ACL 2023
9–14 July, Toronto

Contents

Conference Information

Message from the General Chair

Welcome to ACL 2023, the 61st Annual Meeting of the Association for Computational Linguistics! The conference will be held in Toronto, Canada, July 9-14, 2023.

Following the succession of the recent conferences in our field, ACL 2023 will adopt a hybrid format. While the impact of Covid has considerably diminished in terms of traveling, obtaining visas to Canada entails a very long process. Moreover, the global economic conditions pose challenges for many individuals to travel to conferences. Recognizing these circumstances, we know many participants may not be able to attend the conference in person. Therefore, we are committed to providing a great virtual platform so everyone has the opportunity to interact with other participants and enjoy the conference. Based on the current registered participants, approximately 30% have chosen to attend the conference virtually. Whether you join us in person or virtually, we sincerely hope everyone has a remarkable conference experience.

This General Chair's message is where I express my gratitude to the many individuals who have made enormous contributions to the conference over the past year..

First and foremost, I am grateful for the tremendous efforts by the program chairs: Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. The rapid growth of our field is challenging from the perspective of organizing a conference. Program chairs have admirably handled a huge number of submissions and implemented novel review criteria to improve the quality of reviews and the paper decision process. They responded promptly after ChatGPT was launched and provided guidelines for using it in paper writing. Beyond their responsibilities as program chairs, they have assisted me with various other decisions. Their efforts have truly shaped the conference. Also, thanks to all the senior area chairs, area chairs, reviewers, and the best paper committee, whose commitment and dedication made paper review and selection possible.

Next, I would like to thank the entire organizing committee for their service. It has been an honor for me to collaborate with such a dedicated team. This includes:

- Industry track chairs: Beata Beigman Klebanov, Jason Williams, and Sunayana Sitaram. An addition to this year's ACL is the introduction of a separate industry track. This is motivated by two factors. First, ACL is held in North America this year (and thus no NAACL), and NAACL has an established tradition of hosting an industry track. Second there was an increasing number of industry track submissions at EMNLP last year from previous years. We hope that a separate industry track can foster the dissemination of research on real-world applications in industry settings. Thanks to the industry track chairs for their efforts in coordinating all the logistics associated with this track.

- Demo chairs: Alan Ritter, Danushka Bollegala, and Ruihong Huang, who managed demo submissions and accepted 58 demos that will be presented in the main conference.

- Student research workshop (SRW) chairs: Gisela Vallejo, Vishakh Padmakumar, and Yao Fu, who showed remarkable enthusiasm and dedication in organizing the workshop. They selected 45 papers to be presented in the main conference program. Also thanks to the faculty advisors: Ivan Vulic and Lu Wang, for providing guidance to the SRW chairs and obtaining NSF support for the workshop.

- Workshop chairs: Annie Louis, Eduardo Blanco, and Yang Feng, who collaborated with EACL workshop chairs to select 22 workshops, and served as the vital link between the conference and individual workshop organizers.

- Tutorials chairs: Margot Mieskes, Siva Reddy, and Vivian Chen, who also worked with EACL chairs to select 6 high quality tutorials that cater to the interest and needs of our conference.

- Ethics chairs: Dirk Hovy and Yonatan Bisk, who checked papers flagged with ethics issues. Thanks for their meticulous work to ensure our papers uphold the ethical standards.

- Publication chairs: Ryan Cotterell, Chenghua Lin, Jesse Thomason, Lei Shu, and Lifu Huang, who prepared the conference handbook, ensured proper formatting of papers, and produced the conference proceedings.

- Virtual infrastructure chairs: Jiacheng Xu, Martín Villalba, and Pedro Rodriguez, who worked hard to develop a virtual platform to ensure an engaging conference experience for both in-person and remote participants. They also made various innovations and enhancement on top of the Underline platform, which the conference utilizes.

- Publicity and social media chairs, Devamanyu Hazarika, Eva Vanmassenhove, and Tong Xu, who communicated and publicized the conference through various social media channels, enhancing the visibility and reach of the conference.

- Website chairs: Jinho Choi and Zhongyu Wei, who updated and maintained the conference website to keep participants informed.

- Diversity and inclusion (D&I) chairs: Daniel Beck, Maryam Fazel-Zarandi, and Nedjma Djouhra Ousidhoum, who arranged support to participants facing financial hardships, and organized a diverse array of activities aimed to promoting diversity and inclusion in our community.

- Student volunteer chairs: Ayah Zirikly and Tao Yu, who reviewed applications and selected student volunteers for the conference.

- Sponsorship chairs: Alla Rozovskaya and Lei Li. Thanks to them and Chris Callison-Burch, the ACL sponsorship Director for their efforts in securing sponsorships and managing the relationship between sponsors and the conference. The generous support from our sponsors has played a crucial role in enabling us to maintain a reasonable registration cost for attendees, and the additional sponsorship for D&I initiatives helps our commitment to fostering a diverse and inclusive environment.

- Visa assistance team: Ayana Niwa, Qingwen Liu, Renxiang Zhang, Samridhi Choudhary, and Tao You. Many participants require visas to attend the conference, and we fully understand this lengthy process. This team has been diligently handling visa requests by sending out numerous invitation letters to facilitate visa applications.

- Infrastructure support from Softconf (Richard Gerber) and Underline (Damira Mrsic, Sol Rosenberg). Both platforms kindly accommodated our many, many requests and implemented several new features.

I also want to specially thank Jennifer Rachford, the ACL event director, who handled all the local arrangement for this conference. Though she was relatively new to the role, and often times needed to juggle multiple ACL conferences, she remained well organized, and consistently provided all the necessary information to all members of the organizer committee. Her contributions ensure the success of this conference.

Thanks to previous ACL/EMNLP conference chairs for sharing their knowledge, tips, and best practice on organizing this conference, and ACL Exec for the support they provided throughout the entire planning and execution of this conference.

Lastly, I extend my appreciation to every participant. Regardless of your role, whether as authors or presenters, workshop organizers, tutorial speakers, student volunteers, session chairs, or simply attendees, your involvement is essential in creating a memorable conference.

Welcome everyone to the conference!

ACL 2023 General Chair
Yang Liu
Alexa, Amazon

Message from the Program Chairs

Welcome to the 61st Annual Meeting of the Association for Computational Linguistics! We are excited to welcome everyone in Toronto.

Most of the work of a program chair is behind the scenes: herding reviewers and chairs, wrangling data from various sources, and answering lots and lots of email. This is a volunteer position, so the only reward we get for this is our chance to make the process of submitting and reviewing papers to our conference better. This letter will outline some of those experiments.

First, we asked reviewers for two scores: soundness and excitement. Our goal was that any sound paper would be accepted to some ACL affiliated venue, but that the "main conference" distinction (limited by space) would be focused on the most exciting papers. Our hope was that soundness would be less noisy than a single "overall recommendation" score, which would help reduce the randomness of decisions. Judging by the exit surveys, this change was well received: over 80% of the chairs, reviewers and authors either expressed support or did not object to this change.

Next, we developed a new process for matching papers to reviewers based on keywords for not only the subject matter of the paper, but also its type of contribution and target language(s). This allowed more fine-grained control over the paper-reviewer matches, and we were also able to provide the chairs with context for the paper-reviewer matches.

To improve review quality, we also updated the reviewer guidelines, and developed a system for the authors to flag specific types of issues with reviews. Finally, we have also proposed a new initiative for recognizing outstanding reviewers and chairs (73 awards at ACL'23).

Finally, we have tried to give more options for presentations. Findings papers now have an in-person presentation spotlight slot and virtual posters in addition to recording videos. Virtual posters have portals to link in-person attendees to virtual posters. We have also brought back Miniconf and RocketChat to allow for better virtual communication between papers (regardless of where the authors are).

This conference is a result of the joint efforts of over ten thousand people. We deeply thank them all, and apologize for the many nagging emails we had to send out. In particular:

- the general chair Yang Liu, who led the whole process;

- the incredible team of 70 SACs, 438 ACs, and 4490 reviewers, who were able to handle our record number of submissions;

- the 13,658 authors for their phenomenal scientific contributions, which we were honored to shepherd through the reviewing process;

- the ACL Executive (esp. Iryna Gurevych, Tim Baldwin, David Yarowsky, Yusuke Miyao, Emily M. Bender) for their support of many of our crazy ideas;

- 21 ethics committee reviewers, chaired by Dirk Hovy and Yonatan Bisk, for their hard work to uphold the ACL code of ethics;

- Our Best Paper Award committee (Jonathan Berant, Jose Camacho-Collados, Danqi Chen, Benjamin Van Durme, David Jurgens, Desmond Elliott, Sasha Luccioni, Jonathan May, Tom McCoy, Yusuke Miyao, Ekaterina Shutova, Emma Strubell, Jun Suzuki, Xiaojun Wan, Luke Zettlemoyer), who reviewed a record number of nominated papers under tight schedule;

- Our assistant Youmi Ma, for reducing our email and Softconf workload significantly and suggesting ideas to make the job run smoothly;

- Past ACL PCs, including Smaranda Muresan, Preslav Nakov and Aline Villavicencio (ACL 2022), Yoav Goldberg, Zornitsa Kozareva, Yue Zhang (EMNLP 2022), Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur (NAACL 2021), for their advice and suggestions;

- Publication chairs Ryan Cotterell, Chenghua Lin, Jesse Thomason, Lei Shu, and Lifu Huang, who ensured the proper formatting of camera-ready papers;

- Emma Strubell, Ian Magnusson, and Jesse Dodge for their help in preparing publishable versions of Responsible NLP checklist;

- ACL Anthology director Matt Post;

- TACL editors-in-chief (Asli Celikyilmaz, Roi Reichart, Ani Nenkova) and CL Editor-in-Chief Hwee Tou Ng for coordinating TACL and CL presentations with us;

- Workshop chairs Annie Louis, Eduardo Blanco, and Yang Feng, for helping us to connect the Findings papers to possible presentation slots at workshops;

- Rich Gerber at Softconf, who answered countless emails and implemented several new features on our request;

- Kyle Lo and Semantic Scholar team, who kindly assisted us with data for paper-reviewer matching;

- Our virtual infrastructure chairs (Pedro Rodriguez, Jiacheng Xu, Martín Villalba) and Underline team (Damira Mrsic, Sol Rosenberg) for enabling a new kind of hybrid experience, combining miniconf and Underline;

- the ACL event director Jennifer Rachford and our visa support team (Ayana Niwa, Qingwen Liu, Renxiang Zhang, Samridhi Choudhary, and Tao You), who did everything possible to facilitate the Canada visa situation for ACL attendees.

Submission and Acceptance

We had two routes to submit papers to ACL 2023: directly to the conference or through ACL Rolling Review (ARR). We received a record number of direct submissions (3601 long papers and 958 short papers) in January 2023. In addition, we received 305 commitments from ARR (271 long papers and 34 short papers) in March 2023. In total, we considered 4864 (3872 long and 992 short) papers with 70 senior area chairs, 438 area chairs, 4024 reviewers, 445 secondary reviewers, and 21 ethics reviewers in 27 tracks. We accepted 910 (23.50%) long and 164 (16.53%) short papers for the main conference, and 712 (41.89% including the long papers for the main conference) long and 189 (35.58% including the short papers for the main conference) short papers for Findings. To sum long and short papers, ACL 2023 accepted 1074 (22.08%) papers for the conference and 901 (40.60% including the papers for the main conference) papers for Findings. The ACL 2023 program also features 46 papers from the Transactions of the Association for Computational Linguistics (TACL) journal, and 7 from the Computational Linguistics (CL) journal.

Limitations Section and Responsible NLP Checklist

Following EMNLP 2022 and EACL 2023, we required that each submitted paper must include an explicitly named Limitations section, discussing the limitations of the work. This was to counterbalance the practice of over-hyping the take-away messages of papers, and to encourage more rigorous and honest scientific practice. This discussion did not count towards the page limit, and we asked reviewers to not use the mentioned limitations as reasons to reject the paper, unless there was a really good reason to.

In addition to the mandatory discussion of limitations, a new element at ACL 2023 is that the Responsible NLP Checklist for the accepted papers is not only considered by the reviewers, but also published together with the accepted papers as a special appendix, in an effort to improve transparency and accountability in the field.

Areas

To ensure a smooth process, the submissions to ACL 2023 were divided into 26 areas. The areas mostly followed these of previous ACL, and more broadly *ACL conferences, reflecting the typical divisions in the field. Following EMNLP 2022, we split the "Large Language Models" track away from "Machine learning in NLP", reflecting the growth of submissions in the area. We also offered two new tracks ("Linguistic diversity" and "Multilingualism and Cross-Lingual NLP"). For the papers authored by SACs, the final recommendation decisions were made by a separate SAC team. The most popular areas (with over 250 submissions) were "Dialogue and Interactive Systems", "Information Extraction", "Large Language Models", "Machine Learning for NLP", and "NLP Applications".

Best Paper Awards

ACL'23 implemented the new ACL award policy, aiming to expand the pool of work that is recognized as outstanding. In total, 73 papers were nominated by the reviewers or area chairs for consideration for awards. These papers were assessed by the Best Paper Award Committee, and with their help we selected 4 best papers, 3 special awards (social impact, resource, reproduction), and several dozen outstanding papers. The best and outstanding papers will be announced in a dedicated plenary session for Best Paper Awards on July 10 2023.

Presentation Mode

In ACL 2023, there is no meaningful distinction between oral and poster presentations in terms of paper quality. The composition of the oral sessions were proposed by the SACs of their respective tracks, so as to compose a thematically coherent set of papers on a shared topic or method, which would allow for an engaging discussion. The decisions were not based on the authors' virtual or on-site attendance.

We hope you enjoy the program and the new elements we introduced (but let us know either way). We are looking forward to a great ACL 2023!

Anna Rogers (IT University of Copenhagen, Denmark)
Jordan Boyd-Graber (University of Maryland, USA)
Naoaki Okazaki (Tokyo Institute of Technology, Japan)
ACL 2023 Programme Committee Co-Chairs

Organizing Committee

General Chair

    Yang Liu, Amazon

Program Chairs

    Anna Rogers, IT University of Copenhagen
    Jordan Boyd-Graber, University of Maryland
    Naoaki Okazaki, Tokyo Institute of Technology

Workshop Chairs

    Annie Louis, Google
    Eduardo Blanco, Arizona University
    Yang Feng, Chinese Academy of Science

Tutorials Chairs

    Margot Mieskes, University of Applied Sciences, Darmstadt
    Siva Reddy, McGill University; Mila
    Vivian Chen, National Taiwan University

Demonstrations Chairs

    Alan Ritter, Georgia Institute of Technology
    Danushka Bollegala, University of Liverpool
    Ruihong Huang, Texas A&M University

Industry Track Chairs

    Beata Beigman Klebanov, ETS
    Jason Williams, Apple
    Sunayana Sitaram, Microsoft Research India

Student Research Workshop Chairs

    Gisela Vallejo, University of Melbourne
    Vishakh Padmakumar, New York University
    Yao Fu, University of Edinburgh

Faculty Advisors to SRW

    Ivan Vulič, University of Cambridge
    Lu Wang, University of Michigan

Ethics Chairs

    Dirk Hovy, Bocconi University
    Yonatan Bisk, Carnegie Mellon University

Publication Chairs

Ryan Cotterell, ETH Zürich
Chenghua Lin, University of Sheffield
Jesse Thomason, University of Southern California
Lei Shu, Google

Publicity and Social Media Chairs

Devamanyu Hazarika, Amazon
Eva Vanmassenhove, Tilburg University
Tong Xu, University of Science and Technology of China

Website Chairs

Jinho Choi, Emory University
Zhongyu Wei, Fudan University

Student Volunteer Chairs

Ayah Zirikly, Johns Hopkins University
Tao Yu, University of Hong Kong

Virtual Infrastructure Chairs

Jiacheng Xu, Salesforce
Martín Villalba, Saarland University
Pedro Rodriguez, Meta

Diversity and Inclusion Chairs

Daniel Beck, The University of Melbourne
Maryam Fazel-Zarandi, Meta
Nedjma Djouhra Ousidhoum, University of Cambridge

Sponsorship Chairs

Alla Rozovskaya, The City University of New York
Lei Li, University of California at Santa Barbara

Program Chair Assistant

Youmi Ma, Tokyo Institute of Technology

Visa Assistance Team

Ayana Niwa, Tokyo Institute of Technology
Qingwen Liu, Fudan University
Renxiang Zhang, Amazon
Samridhi Choudhary, Amazon
Tao You, Fudan University

ACL Event Director

    Jennifer Rachford, Association for Computational Linguistics

Program Committee

Computational Social Science and Cultural Analytics

Walid Magdy, Daniel Preotiuc-Pietro, Md. Shad Akhtar, Nikolaos Aletras, Kalina Bontcheva, Kareem Darwish, Mai Elsherief, Kiran Garimella, Marco Guerini, Kokil Jaidka, Barbara Mcgillivray, Yelena Mejova, Usman Naseem, Bjorn Ross, James Thorne, Marco Viviani, Soroush Vosoughi, Ingmar Weber

Dialogue and Interactive Systems

Y-Lan Boureau, Mary Ellen Foster, Minlie Huang, João Sedoc, Luciana Benotti, Paul Crook, Maryam Fazel-Zarandi, Michel Galley, Kallirroi Georgila, Alborz Geramifard, Devamanyu Hazarika, Baotian Hu, Wenqiang Lei, Gina-Anne Levow, Piji Li, Andrea Madotto, Fei Mi, Seungwhan Moon, Lili Mou, Natalie Parde, Baolin Peng, Oleg Rokhlenko, Samira Shaikh, Lei Shu, Kurt Shuster, Ruihua Song, Yiping Song, Shabnam Tafreshi, Ryuichi Takanobu, David Traum, Stefan Ultes, Charles Welch, Min Yang, Zhou Yu, Wei-Nan Zhang, Hao Zhou

Discourse and Pragmatics

Christian Hardmeier, Jey Han Lau, Jacob Andreas, Chloé Braud, Luis Fernando D'haro, Junyi Jessy Li, Sharid Loaiciga, Nafise Sadat Moosavi, Anna Nedoluzhko, Juntao Yu, Amir Zeldes

Ethics and NLP

Vinodkumar Prabhakaran, Diyi Yang, Kai-Wei Chang, Sunipa Dev, Karen Fort, Jack Hessel, Debora Nozza, Zeerak Talat, Yulia Tsvetkov

Generation

Sebastian Gehrmann, Mohit Iyyer, Nina Dethlefs, Nan Duan, Greg Durrett, Angela Fan, Claire Gardent, Albert Gatt, Yeyun Gong, Srinivasan Iyer, Meng Jiang, Sujian Li, Ankur Parikh, Nanyun Peng, Lianhui Qin, Sudha Rao, Hannah Rashkin, Jinsong Su, Hiroya Takamura, John Wieting, Rui Yan, Jiajun Zhang

Information Extraction

Lifu Huang, Chin-Yew Lin, Aaron White, Yixin Cao, Shiyu Chang, Muhao Chen, Brian Davis, Antoine Doucet, Xinya Du, Radu Florian, Xianpei Han, Filip Ilievski, Diana Inkpen, Reno Kriz, Lane Lawley, Manling Li, Kang Liu, Zhiyuan Liu, Bonan Min, Thien Nguyen, Qiang Ning, Alan Ritter, Benjamin Roth, Lei Sha, Jingbo Shang, Ge Shi, Xianzhi Wang, Wenpeng Yin, Mo Yu, Dongyan Zhao, Jun Zhao, Christos Christodoulopoulos

Information Retrieval and Text Mining

Benjamin Piwowarski, Qifan Wang, Yi Fang, Fuli Feng, Yiqun Liu, Jian-Yun Nie, Xiaojun Quan, Yi Tay, Hongning Wang, Jingang Wang, Zenglin Xu, Grace Hui Yang

Interpretability and Analysis of Models for NLP

Carolin Lawrence, Ana Marasovic, Chenhao Tan, Jasmijn Bastings, Dallas Card, Samuel Carton, Oana Cocarascu, Nadir Durrani, Jacob Eisenstein, Mor Geva, Ivan Habernal, Peter Hase, Alon Jacovi, Yangfeng Ji, Divyansh Kaushik, Piyawat Lertvittayakumjorn, Zaiqiao Meng, Pasquale Minervini, Isar Nejadgholi, Danish Pruthi, Abhilasha Ravichander, Roi Reichart, Swabha Swayamdipta, Martin Tutek, Elena Voita, Sarah Wiegreffe, Tongshuang Wu

Language Grounding to Vision, Robotics, and Beyond

Zhongyu Wei, Mark Yatskar, Yoav Artzi, Yi Cai, Jingjing Chen, Zhihao Fan, Daniel Fried, Jiasen Lu, Lin Ma, Aishwarya Padmakumar, Zhaochun Ren, Freda Shi, Carina Silberer, Alessandro Suglia, Alane Suhr, Chen Sun, Hao Tan, Meng Wang, Tong Xu

Large Language Models

Dipanjan Das, Bhuwan Dhingra, Mike Lewis, Xuezhe Ma, Miguel Ballesteros, Kenneth Church, Kumar Dubey, Orhan Firat, Marjan Ghazvininejad, Hila Gonen, Junxian He, Harsh Jhamtani, Mandar Joshi, Xiang Kong, Ni Lao, Moontae Lee, Bing Liu, Peter Liu, Eric Malmi, Huan Sun, Lijun Wu, Chunting Zhou

Linguistic Diversity

Constantine Lignos, Emily Prud'hommeaux, Rebecca Knowles, Zoey Liu, Teresa Lynn, Lane Schwartz, Francis Tyers, Marcos Zampieri

Linguistic Theories, Cognitive Modeling, and Psycholinguistics

Afra Alishahi, Najoung Kim, Lisa Beinborn, Abdellah Fourtassi, Nan-Jiang Jiang, R. Thomas McCoy, Aida Nematzadeh, Grusha Prasad

Machine Learning for NLP

Marie-Francine Moens, Anna Rumshisky, Kevin Small, Heike Adel, Mikhail Burtsev, Giuseppe Castellucci, Trevor Cohn, Danilo Croce, Julian Eisenschlos, Francis Ferraro, Matthias Galle, Dan Goldwasser, Hannaneh Hajishirzi, Ricardo Henao, Estevam Hruschka, Pei Ke, Parisa Kordjamshidi, Omer Levy, Zemin Liu, André Martins, Ashutosh Modi, Ndapa Nakashole, Thanh Tam Nguyen, Giannis Nikolentzos, Barbara Plank, Steven Schockaert, Freda Shi, Vivek Srikumar, Jun Suzuki, Hao Tang, Lu Wang, Taro Watanabe, Ningyu Zhang

Machine Translation

Markus Freitag, Tom Kocmi, Lei Li, Boxing Chen, Colin Cherry, George Foster, Roman Grundkiewicz, Francisco Guzman, Shujian Huang, Philipp Koehn, Qun Liu, Chi-Kiu Lo, Haitao Mi, Jan Niehues, Stephan Peitz, Maja Popović, Ricardo Rei, Felix Stahlberg, Zhaopeng Tu, David Vilar, Mingxuan Wang, Joern Wuebker, Tong Xiao, Jingjing Xu, François Yvon, Yue Zhang, Hao Zhou

Multilingualism and Cross-Lingual NLP

A. Seza Doğruöz, Sunayana Sitaram, Muhammad Abdul-Mageed, David Ifeoluwa Adelani, Alham Fikri Aji, Antonios Anastasopoulos, Mikel Artetxe, Yoshinari Fujinuma, Dan Garrette, Shruti Rijhwani, Sebastian Ruder, Xinyi Wang

NLP Applications

Sophia Ananiadou, Mark Dras, Jing Jiang, Makoto Miwa, Vincent Ng, Hadi Amiri, Riza Batista-Navarro, Jose Camacho-Collados, Fenia Christopoulou, Giovanni Da San Martino, Dina Demner-Fushman, Luigi Di Caro, Haibo Ding, Mariano Felice, Wei Gao, Sanda Harabagiu, Seung-Won Hwang, Naoya Inoue, Shafiq Joty, Ekaterina Kochmar, Mamoru Komachi, Wei Lu, Shervin Malmasi, David Mimno, Preslav Nakov, Maria Leonor Pacheco, Marek Rei, Kirk Roberts, Sara Rosenthal, Alla Rozovskaya, Tulika Saha, Hiroki Sakaji, Matthew Shardlow, Shuohang Wang, Jason Wei, Qianqian Xie, Jianfei Yu, Chrysoula Zerva, Aston Zhang, Arkaitz Zubiaga

Phonology, Morphology, and Word Segmentation

Miikka Silfverberg, Ekaterina Vylomova, Ryan Cotterell, Xuanjing Huang, David R. Mortensen

Question Answering

Eunsol Choi, Mrinmaya Sachan, Rishiraj Saha Roy, Priyanka Agrawal, Chitta Baral, Gianni Barlacchi, Hao Cheng, Danish Contractor, Pradeep Dasigi, Tushar Khot, Rik Koncel-Kedziorski, Bill Yuchen Lin, Bang Liu, Ismini Lourentzou, Sewon Min, Liangming Pan, Panupong Pasupat, Peng Qi, Ashish Sabharwal, Xiaoyu Shen, Veselin Stoyanov, Yu Su, Kai Sun, Mihai Surdeanu, Di Wang, Ziyu Yao, Yuhao Zhang

Resources and Evaluation

Sarvnaz Karimi, Nathan Schneider, Karin Verspoor, Rachel Bawden, Asma Ben Abacha, Doina Caragea, Jennifer D'souza, Rotem Dror, Ondrej Dusek, Steffen Eger, Jorge Gracia, Udo Hahn, Lifeng Han, Radu Tudor Ionescu, David Janiszek, Sudipta Kar, Jin-Dong Kim, Jonathan Kummerfeld, John P. Lalor, Fabrice Lefèvre, Jochen Leidner, Roser Morante, Gabriella Pasi, Maja Popović, German Rigau, Yves Scherrer, Manish Shrivastava, Sowmya Vajjala, Lucy Lu Wang

Semantics: Lexical

Marianna Apidianaki, Gabriella Lapesa, Chris Biemann, Guy Emerson, Allyson Ettinger, Goran Glavaš, Dieuwke Hupkes, Nancy Ide, Andrey Kutuzov, Alessandro Lenci, Mohammad Taher Pilehvar, Yuval Pinter, Edoardo Maria Ponti, Vered Shwartz, Lonneke Van Der Plas, Ivan Vulić

Semantics: Sentence-level Semantics, Textual Inference, and Other Areas

Yuki Arase, Roberto Navigli, Roy Schwartz, Tommaso Caselli, Simone Conia, Lei Cui, Li Dong, Lea Frermann, Atsushi Fujita, Christophe Gravier, Luheng He, Germán Kruszewski, Tommaso Pasini, Adam Poliak, Jakob Prange, Michael Roth, Keisuke Sakaguchi, Abulhair Saparov, Ji-Rong Wen, Wei Xu, Sho Yokoi, Chen Zhao

Sentiment Analysis, Stylistic Analysis, and Argument Mining

Lun-Wei Ku, Henning Wachsmuth, Khalid Al Khatib, Elena Cabrio, Hao Fei, Anette Frank, Lin Gui, Yufang Hou, Ting-Hao Huang, Kentaro Inui, Anne Lauscher, John Lawrence, Saif Mohammad, Joonsuk Park, Shabnam Tafreshi, Orith Toledo-Ronen, Serena Villata, Shuai Wang

Speech and Multimodality

Grzegorz Chrupała, Frank Rudzicz, Laurent Besacier, Manaal Faruqui, Sharon Goldwater, Florian Metze, Okko Rasanen, Andrew Rosenberg, Hao Tang, Wenwu Wang, Xin Wang, Shinji Watanabe

Summarization

Chenghua Lin, Shashi Narayan, Reinald Kim Amplayo, Avi Caciularu, Chung-Chi Chen, Gong Cheng, Markus Dreyer, Xiaocheng Feng, Kathleen Mckeown, Stuart Middleton, Richard Yuanzhe Pang, Xiaojun Wan, Xingxing Zhang, Yao Zhao

Syntax: Tagging, Chunking, and Parsing

Wanxiang Che, Djamé Seddah, Xinchi Chen, Leyang Cui, Lifeng Jin, Zhenghua Li, Joakim Nivre, Kenji Sagae, Meishan Zhang

Theme: Reality Check

Ehud Reiter, Xiang Ren, Malihe Alikhani, Jan Buys, Jesse Dodge, Antske Fokkens, Robin Jia, Daniel Khashabi, Emiel Krahmer, Saad Mahamood, Margaret Mitchell, Richard Sproat, Byron Wallace, Adina Williams

COI

Shay B. Cohen, Daisuke Kawahara

Ethics

Yonatan Bisk, Dirk Hovy, Jin-Dong Kim, Zeerak Talat

Best Paper Selection Committee

Jonathan Berant, Jose Camacho-Collados, Danqi Chen, Benjamin Van Durme, David Jurgens, Desmond Elliott, Sasha Luccioni, Jonathan May, Tom McCoy, Yusuke Miyao, Ekaterina Shutova, Emma Strubell

Primary Reviewers

Amirhossein Abaskohi, Harika Abburi, Asad Abdi, Sadaf Abdul Rauf, Muhammad Abdul-Mageed, Kaori Abe, Omri Abend, Gavin Abercrombie, Sallam Abualhaija, Abdalghani Abujabal, Alafate Abulimiti, Lars Ackermann, Griffin Adams, Ife Adebara, David Ifeoluwa Adelani, Benedikt Adelmann, Tosin Adewumi, Jiban Adhikary, Suman Adhya, Yossi Adi, Somak Aditya, Vaibhav Adlakha, Noëmi Aepli, Stergos Afantenos, Haithem Afli, Ankur Agarwal, Sanchit Agarwal, Shivam Agarwal, Rodrigo Agerri, Arshiya Aggarwal, Karan Aggarwal, Piush Aggarwal, Manex Agirrezabal, Guy Aglionby, Aishwarya Agrawal, Ameeta Agrawal, Sweta Agrawal, Roee Aharoni, Wasi Uddin Ahmad, Sina Ahmadi, Natalie Ahn, Aman Ahuja, Chaitanya Ahuja, Kabir Ahuja, Lin Ai, Xi Ai, Ankit Aich, Annalena Aicher, Laura Aina, Salah Aït-Mokhtar, Akiko Aizawa, Alham Fikri Aji, Aswathy Ajith, Reina Akama, Pritom Saha Akash, Alan Akbik, Adewale Akinfaderin, Nader Akoury, Burak Aksar, Ibrahim Taha Aksu, Mousumi Akter, Arjun Akula, Ekin Akyurek, Hend Al-Khalifa, Hadeel Al-Negheimish, Hussein Al-Olimat, Rami Al-Rfou, Nora Al-Twairesh, Firoj Alam, Mehwish Alam, Alon Albalak, Abdullah Albanyan, Chris Alberti, Hanan Aldarmaki, Vasiliy Alekseev, Jan Alexandersson, Georgios Alexandridis, Mark Alfano, David Alfter, Robin Algayres, Raquel G. Alhama, Abdulaziz Alhamadani, Tariq Alhindi, Hamed Alhoori, Hassan Alhuzali, Badr Alkhamissi, Maxime Allard, Emily Allaway, Liesbeth Allein, Tiago Almeida, Khalid Alnajjar, Omar Alonso, Abdullah Alrajeh, Milad Alshomary, Maha Jarallah Althobaiti, Duygu Altinok, Fernando Alva-Manchego, Rami Aly, Chiara Alzetta, Bharat Ram Ambati, Maxime Amblard, Iqra Ameer, Saadullah Amin, Afra Amini, Silvio Amir, Maaz Amjad, Haozhe An, Jie An, Jisun An, Ashish Anand, Sophia Ananiadou, Raviteja Anantha, Rafael Anchiêta, Mark Anderson, Nicholas Andrews, Raghuram Annasamy, Diego Antognini, Jean-Yves Antoine, Maria Antoniak, Wissam Antoun, Rishita Anubhai, Xiang Ao, Emilia Apostolova, Mario Aragon, Erik Arakelyan, Jun Araki, Rahul Aralikatte, Ayme Arango Monnar, Oscar Araque, Matheus Araujo, John Arevalo, Arturo Argueta, Mozhdeh Ariannezhad, Hiba Arnaout, Akhil Arora, Piyush Arora, Siddhant Arora, Leila Arras, Ekaterina Artemova, Philip Arthur, Ron Artstein, Anjana Arunkumar, Saurav Aryal, Akari Asai, Ehsaneddin Asgari, Elliott Ash, Nicholas Asher, Md.sadek Hossain Asif, Arian Askari, Matthias Assenmacher, Zhenisbek Assylbekov, Berk Atil, Giuseppe Attanasio, Mohammed Attia, Aitziber Atutxa Salazar, Lauriane Aufrant, Tal August, Hayastan Avetisyan, Eleftherios Avramidis, Vera Axelrod, Hammad Ayyubi, Hosein Azarbonyad, Gorka Azkune, Aslan B. Wong, Bogdan Babych, Luca Bacco, Nguyen Bach, Sarkhan Badirli, Ebrahim Bagheri, Petra Bago, Parnia Bahar, Ashutosh Baheti, Vikas Bahirwani, Bing Bai, Fan Bai, He Bai, Jiaxin Bai, Long Bai, Xuefeng Bai, Yinhao Bai, Yu Bai, Yushi Bai, Jinyeong Bak, Amir Bakarov, Collin Baker, Vidhisha Balachandran, Mithun Balakrishna, Oana Balalau, Vevake Balaraman, Ananth Balashankar, Ramya Balasubramaniam, Gunjan Balde, Ioana Baldini, Timothy Baldwin, Simone Balloccu, Mohammadreza Banaei, Dibyanayan Bandyopadhyay, Debayan Banerjee, Pratyay Banerjee, Seojin Bang, Yejin Bang, Vinayshekhar Bannihatti Kumar, Hritik Bansal, Forrest Sheng Bao, Guangsheng Bao, Junwei Bao, Yu Bao, Yuwei Bao, Ankur Bapna, Kfir Bar, Roy Bar-Haim, Claire Barale, Mohamad Hardyman Barawi, Edoardo Barba, Adrien Barbaresi, Verginica Barbu Mititelu, M Saiful Bari, Loic Barrault, Alberto Barrón-Cedeño, Sabine Bartsch, Sabyasachee Baruah, Marco Basaldella, Pierpaolo Basile, Valerio Basile, Ali Basirat, Elisa Bassignana, Mohaddeseh Bastan, Kinjal Basu, Somnath Basu Roy Chowdhury, Tatiana Batura, Daniel Bauer, Timo Baumann, Ian Beaver, Björn Bebensee, Daniel Beck, Lee Becker, Maria Becker, Barend Beekhuizen, Dorothee Beermann, Gasper Begus, Melika Behjati, Shabnam Behzad, Andrei Stefan Bejgu, Nazar Beknazarov, Nuria Bel, Yonatan Belinkov, Eric Bell, Meriem Beloucif, Luca Benedetto, Martin Benjamin, Lauren Benson, Gábor Berend, Benjamin Bergen, Leon Bergen, Maria Berger, Nathaniel Berger, Rafael Berlanga, Gabriel Bernier-Colborne, Dario Bertero, Laurent Besacier, Chandra Bhagavatula, Rasika Bhalerao, Rohan Bhambhoria, Avanti Bhandarkar, Rishabh Bhardwaj, Aditya Bhargava, Pushpak Bhattacharyya, Satwik Bhattamishra, Bimal Bhattarai, Shohini Bhattasali, Anahita Bhiwandiwalla, Plaban Bhowmick, Rajarshi Bhowmik, Mukul Bhutani, Nikita Bhutani, Bin Bi, Guanqun Bi, Wei Bi, Giovanni Biancofiore, Adrien Bibal, Ann Bies, Laura Biester, Geetanjali Bihani, Yi Bin, Arne Binder, Jennifer Bishop, Debmalya

Biswas, Yonatan Bitton, Johannes Bjerva, Henrik Björklund, Johanna Bjorklund, Philippe Blache, Nate Blaylock, Avi Bleiweiss, Terra Blevins, Rexhina Blloshmi, Su Lin Blodgett, Jelke Bloem, Michael Bloodgood, Carlos Bobed Lisbona, Victoria Bobicev, Ben Bogin, Bernd Bohnet, Ondřej Bojar, Huang Bojun, Valeriia Bolotova-Baranova, Necva Bölücü, Rishi Bommasani, Daniele Bonadiman, Alessandro Bondielli, Francesca Bonin, Logan Born, Mihaela Bornea, Emanuela Boros, Johan Bos, Digbalay Bose, Robert Bossy, Kaj Bostrom, Florian Boudin, Mohand Boughanem, Gerlof Bouma, Gosse Bouma, Zied Bouraoui, Andrey Bout, Johan Boye, Faeze Brahman, António Branco, Stephanie Brandl, Kiante Brantley, Pavel Braslavski, Adrian Brasoveanu, Daniel Braun, Jacob Bremerman, Jonathan Brennan, Chris Brew, Shaked Brody, Thomas Brovelli (meyer), Hannah Brown, Caroline Brun, Dominique Brunato, Yi Bu, Emanuele Bugliarello, Trung Bui, Paul Buitelaar, Razvan Bunescu, Laurie Burchell, Susanne Burger, Jill Burstein, Victor Bursztyn, Davide Buscaldi, Hendrik Buschmeier, Miriam Butt, Joan Byamugisha, Bill Byrne, Donna Byron, José G. C. De Souza, Michele Cafagna, Aoife Cahill, Samuel Cahyawijaya, Deng Cai, Han Cai, Hengyi Cai, Hongjie Cai, Pengshan Cai, Xiangrui Cai, Ruken Cakici, Iacer Calixto, Zoraida Callejas, Jesus Calvillo, Giovanni Campagna, Leonardo Campillos-Llanos, Niccolò Campolungo, Daniel Campos, Jon Ander Campos, Ricardo Campos, Burcu Can, M Abdullah Canbaz, Nicola Cancedda, Marie Candito, Ed Cannon, Erion Çano, Boxi Cao, Hailong Cao, Hejing Cao, Jiangxia Cao, Jie Cao, Kris Cao, Mengyun Cao, Pengfei Cao, Qingqing Cao, Qingxing Cao, Ruisheng Cao, Shuyang Cao, Yixuan Cao, Yu Cao, Yu Cao, Yuan Cao, Yuwei Cao, Ziqiang Cao, Cristian Cardellino, Rémi Cardon, Boaz Carmeli, Xavier Carreras, Paula Carvalho, Francisco Casacuberta, Fabio Casati, Helena Caseli, Pierluigi Cassotti, Sheila Castilho, Arie Cattan, Andrew Cattle, Paulo Cavalin, Roberto Centeno, Dumitru-Clementin Cercel, Christophe Cerisara, Mauro Cettolo, Sky Ch-Wang, Haixia Chai, Heyan Chai, Joyce Chai, Junyi Chai, Yekun Chai, Tuhin Chakrabarty, Megha Chakraborty, Tanmoy Chakraborty, Bharathi Raja Chakravarthi, Yllias Chali, Ilias Chalkidis, Nathanael Chambers, Hou Pong Chan, Zhangming Chan, Anshuma Chandak, . Chandrahas, Raman Chandrasekar, Baobao Chang, Buru Chang, Ernie Chang, Haw-Shiuan Chang, Heng Chang, Kent Chang, Serina Chang, Shuaichen Chang, Tyler Chang, Yapei Chang, Yung-Chun Chang, Tai Chang-You, Guan-Lin Chao, Rajen Chatterjee, Akshay Chaturvedi, Iti Chaturvedi, Aditi Chaudhary, Vishrav Chaudhary, Subhajit Chaudhury, Geeticka Chauhan, Kushal Chawla, Chao Che, Ciprian Chelba, Emmanuel Chemla, Beiduo Chen, Berlin Chen, Bo Chen, Boli Chen, Canyu Chen, Catherine Chen, Chacha Chen, Chen Chen, Deli Chen, Derek Chen, Dongsheng Chen, Francine Chen, Fuxiang Chen, Guanhua Chen, Guanliang Chen, Guanyi Chen, Hanjie Chen, Howard Chen, Huiyuan Chen, Hung-Ting Chen, Jia Chen, Jiaao Chen, Jiangjie Chen, Jiaze Chen, Jifan Chen, Jingye Chen, John Chen, Junfan Chen, Junyang Chen, Kehai Chen, Kezhen Chen, Lei Chen, Lichang Chen, Lihu Chen, Lin Chen, Linqing Chen, Long Chen, Lu Chen, Luoxin Chen, Maximillian Chen, Mei-Hua Chen, Meiqi Chen, Meng Chen, Mingda Chen, Nuo Chen, Pei Chen, Qian Chen, Qiang Chen, Qianglong Chen, Qin Chen, Qipin Chen, Qiyuan Chen, Ruey-Cheng Chen, Sanxing Chen, Shijie Chen, Shizhe Chen, Sihao Chen, Tao Chen, Tongfei Chen, Xiaojun Chen, Xiaoli Chen, Xiaoyin Chen, Xilun Chen, Xingran Chen, Xinhong Chen, Xiuyi Chen, Xiuying Chen, Yang Chen, Yangbin Chen, Yangyi Chen, Yanping Chen, Yen-Chun Chen, Yiming Chen, Ying Chen, Yongjun Chen, Yu Chen, Yubo Chen, Yubo Chen, Yue Chen, Yue Chen, Yulong Chen, Yun Chen, Yunmo Chen, Zeming Chen, Zhibin Chen, Zhihong Chen, Zhijun Chen, Zhiyu Chen, Zhiyu Chen, Zhuang Chen, Fei Cheng, Liying Cheng, Lu Cheng, Myra Cheng, Pengxiang Cheng, Qinyuan Cheng, Shanbo Cheng, Sijie Cheng, Weiwei Cheng, Yong Cheng, Yu Cheng, Zhi-Qi Cheng, Vijil Chenthamarakshan, Joe Cheri, Artem Chernodub, Emmanuele Chersoni, Jackie Chi Kit Cheung, Jianfeng Chi, Zewen Chi, Cheng-Han Chiang, David Chiang, Patricia Chiril, Nadezhda Chirkova, Luis Chiruzzo, Billy Chiu, Javier Chiyah-Garcia, Hyunchang Cho, Hyundong Cho, Hyunsoo Cho, Sangwoo Cho, Seunghyuk Cho, Sungjun Cho, Sungzoon Cho, Won Ik Cho, Young Min Cho, Daejin Choi, Jihun Choi, Jinho D. Choi, Seungtaek Choi, Yejin Choi, Yunseok Choi, Shamil Chollampatt, Jaegul Choo, Shubham Chopra, Leshem Choshen, Prafulla Kumar Choubey, Monojit Choudhury, Md Faisal Mahbub Chowdhury, Shammur Absar Chowdhury, Lukas Christ, Chenhui Chu, Yun-Wei Chu, Zewei Chu, Zhendong Chu, Yung-Sung Chuang, Jayeol Chun, Jin-Woo Chung, Abu Nowshed Chy, Alessandra Teresa Cignarella, Philipp

Cimiano, Manuel Ciosici, Jorge Civera Saiz, Christopher Clark, Elizabeth Clark, Vincent Claveau, Ann Clifton, Maximin Coavoux, Anne Cocos, Daniel Cohen, Raphael Cohen, Mariona Coll Ardanuy, Davide Colla, Marcus Collins, Pedro Colon-Hernandez, Andrei Coman, Mathieu Constant, Paul Cook, Asa Cooper Stickland, Anna Corazza, Francesco Corcoglioniti, João Cordeiro, Nathan Cornille, Gonçalo Correia, Erin Crabb, Benoit Crabbé, Mathias Creutz, Liam Cripwell, Fabien Cromieres, Maxwell Crouse, Heriberto Cuayahuitl, Ganqu Cui, Haotian Cui, Peng Cui, Shaobo Cui, Shiyao Cui, Wanyun Cui, Xia Cui, Yiming Cui, Rossana Cunha, Washington Cunha, Jeff Da, Iria Da Cunha, Raj Dabre, Gautier Dagan, Deborah Dahl, Leonard Dahlmann, Daniel Dahlmeier, Damai Dai, Hongliang Dai, Qin Dai, Wenliang Dai, Xiang Dai, Yi Dai, Yinpei Dai, Yong Dai, Daniel Dakota, Fahim Dalvi, Marco Damonte, Sandipan Dandapat, Rumen Dangovski, Verna Dankers, Aswarth Abhilash Dara, Amitava Das, Anubrata Das, Ayan Das, Debopam Das, Dipankar Das, Mithun Das, Sarkar Snigdha Sarathi Das, Souvik Das, Mithun Das Gupta, Sarthak Dash, Debajyoti Datta, Vidas Daudaravicius, Sam Davidson, Forrest Davis, Joe Davison, Luna De Bruyne, Gael De Chalendar, Orphee De Clercq, Adria De Gispert, Michiel De Jong, Kordula De Kuthy, Éric De La Clergerie, Cyprien De Lichy, Ernesto William De Luca, Renato De Mori, Andrea De Varda, Alok Debnath, Mathieu Dehouck, Maksym Del, Luciano Del Corro, Jean-Benoit Delbrouck, Marc Delcroix, Sebastien Delecraz, Louise Deleger, Felice Dell'orletta, Pieter Delobelle, Vera Demberg, Daryna Dementieva, David Demeter, Seniz Demir, Dorottya Demszky, Steve Deneefe, Haolin Deng, Mingkai Deng, Shumin Deng, Xiang Deng, Xun Deng, Yang Deng, Yuntian Deng, Zhi-Hong Deng, Pascal Denis, Michael Denkowski, Leon Derczynski, Jan Deriu, Daniel Deutsch, Premkumar Devanbu, Murthy Devarakonda, Chris Develder, Hannah Devinney, Suvodip Dey, Jay Deyoung, Prajit Dhar, Zonglin Di, Barbara Di Eugenio, Mattia Di Gangi, Luca Di Liello, Giorgio Maria Di Nunzio, Shizhe Diao, Gaël Dias, Alberto Diaz, Dimitar Dimitrov, Emily Dinan, Bosheng Ding, Chenchen Ding, Jie Ding, Kaize Ding, Keyang Ding, Liang Ding, Ning Ding, Shuoyang Ding, Wenjian Ding, Wentao Ding, Yangruibo Ding, Yuning Ding, Zeyuan Ding, Zixiang Ding, Anca Dinu, Liviu P. Dinu, Peter Dirix, Ajay Divakaran, Kalpit Dixit, Tanay Dixit, Nemanja Djuric, Dmitriy Dligach, Sumanth Doddapaneni, Pavel Dolin, Miguel Domingo, Chenhe Dong, Haoyu Dong, Meixing Dong, Mengxing Dong, Ming Dong, Minghui Dong, Qianqian Dong, Qingxiu Dong, Xiangjue Dong, Xin Dong, Christine Doran, Bonaventure F. P. Dossou, Longxu Dou, Zhicheng Dou, Zi-Yi Dou, Jad Doughman, Eduard Dragut, Aleksandr Drozd, Jinhua Du, Li Du, Li Du, Mengnan Du, Pan Du, Tianyu Du, Wanyu Du, Yangkai Du, Yulun Du, Yupei Du, Dheeru Dua, Hanyu Duan, Jiali Duan, Jiaxin Duan, Junwen Duan, Pengfei Duan, Sufeng Duan, Xiangyu Duan, Pablo Duboue, Philipp Dufter, Liam Dugan, Nicolas Dugue, Kevin Duh, Jonathan Dunn, Tejas Duseja, Brian Dusell, Sourav Dutta, Tomasz Dwojak, William Dyer, Haihong E, Kurt Eberle, Sebastian Ebert, Hiroshi Echizen'ya, Lukas Edman, Daniel Edmiston, Aleksandra Edwards, Santiago Egea Gómez, Markus Egg, Koji Eguchi, Yo Ehara, Maud Ehrmann, Roald Eiselen, Jason Eisner, Asif Ekbal, Ismail El Maarouf, Samhaa R. El-Beltagy, Aparna Elangovan, Yanai Elazar, Maha Elbayad, Heba Elfardy, Mohamed Elgaar, Michael Elhadad, Basil Ell, Desmond Elliott, Fatma Elsafoury, Micha Elsner, Chris Chinenye Emezue, Saman Enayati, Joseph Enguehard, Sugyeong Eo, Liana Ermakova, Ori Ernst, Patrick Ernst, Engin Erzin, Carlos Escolano, Arash Eshghi, Cristina España-Bonet, Luis Espinosa Anke, Dominique Estival, Kawin Ethayarajh, Kilian Evang, Kenneth Ezukwoke, Saad Ezzini, Alex Fabbri, Marzieh Fadaee, Michael Faerber, Guglielmo Faggioli, Fahim Faisal, Agnieszka Falenska, Neele Falk, Tobias Falke, James Fan, Jungwei Fan, Yao-Chung Fan, Yimin Fan, Yue Fan, Zhihao Fan, Hui Fang, Qingkai Fang, Tianqing Fang, Yihao Fang, Yimai Fang, Yuwei Fang, Hossein Fani, Ana C Farinha, Nawshad Farruque, Amany Fashwan, Mehwish Fatima, Adam Faulkner, Benoit Favre, Amir Feder, Marc Feger, Zichu Fei, Guy Feigenblat, Nils Feldhus, Sergey Feldman, Virginia Felkner, Jianzhou Feng, Jiazhan Feng, Shangbin Feng, Shi Feng, Shutong Feng, Steven Y. Feng, Weixi Feng, Xiachong Feng, Yang Feng, Yansong Feng, Yu Feng, Yunhe Feng, Zhangyin Feng, Paulo Fernandes, Nigel Fernandez, Ramon Fernandez Astudillo, Javier Fernandez-Cruz, Daniel Fernández-González, Elisa Ferracane, Javier Ferrando, Rafael Ferreira, Besnik Fetahu, Alejandro Figueroa, Matthew Finlayson, Mauajama Firdaus, Mark Fishel, Margaret Fleck, Michael Flor, Jose Fonollosa, Marco Aurelio Fonseca, Tommaso Fornaciari, Karen Fort, Jennifer Foster, Abdellah Fourtassi, Robert Frank, Kathleen C.

Fraser, Flavius Frasincar, Diego Frassinelli, Dayne Freitag, André Freitas, Simona Frenda, Victor Fresno, Dan Friedman, Annemarie Friedrich, Jason Fries, Francesca Frontini, Guohong Fu, Jie Fu, Lisheng Fu, Liye Fu, Peng Fu, Xiyan Fu, Yao Fu, Nancy Fulda, Kotaro Funakoshi, Pascale Fung, Yi Fung, Martin Funkquist, Hagen Fürstenau, Richard Futrell, Matteo Gabburo, Kata Gábor, Marco Gaido, Amit Gajbhiye, Dimitris Galanis, Olivier Galibert, Lukas Galke, Ramiro H. Gálvez, Mihaela Gaman, Leilei Gan, Yujian Gan, Sudeep Gandhe, Ashwinkumar Ganesan, Balaji Ganesan, Ananya Ganesh, Varun Gangal, Debasis Ganguly, William Gantt, Chang Gao, Chongyang Gao, Cuiyun Gao, Ge Gao, Hongyang Gao, Jiahui Gao, Jinhua Gao, Jun Gao, Lingyu Gao, Pengzhi Gao, Qiaozi Gao, Shen Gao, Tianyu Gao, Wei Gao, Xin Gao, Yifan Gao, Yingbo Gao, Cristina Garbacea, Marcos Garcia, Aitor García Pablos, Leibny Paola Garcia Perera, Iker García-Ferrero, Diego Garcia-Olano, Krishna Garg, Muskan Garg, Sarthak Garg, Siddhant Garg, Aina Garí Soler, Ekaterina Garmash, Łukasz Garncarek, Nicolas Garneau, Federico Gaspari, Judith Gaspers, Itai Gat, Susan Gauch, Eric Gaussier, Tanja Gaustad, Dipesh Gautam, Mengshi Ge, Suyu Ge, Xiou Ge, Yixiao Ge, Zhaocheng Ge, Michaela Geierhos, Christian Geishauser, Ruiying Geng, Ariel Gera, Felix Gervits, Luke Gessler, Hamidreza Ghader, Sahar Ghannay, Sarik Ghazarian, Mozhdeh Gheini, Deepanway Ghosal, Amur Ghose, Sayan Ghosh, Sayontan Ghosh, Soumitra Ghosh, Sourav Ghosh, Sreyan Ghosh, Sucheta Ghosh, Filip Ginter, John Giorgi, Salvatore Giorgi, Voula Giouli, Mario Giulianelli, Ameya Godbole, Nathan Godey, Pranav Goel, Rahul Goel, Vaibhava Goel, Anne Göhring, Koldo Gojenola, Tejas Gokhale, Yoav Goldberg, Seraphina Goldfarb-Tarrant, Sujatha Das Gollapalli, Olga Golovneva, Luís Gomes, Jose Manuel Gomez-Perez, Carlos Gómez-Rodríguez, Hugo Goncalo Oliveira, Marcos Goncalves, Teresa Goncalves, Lovedeep Gondara, Hongyu Gong, Jiaying Gong, Linyuan Gong, Shansan Gong, Zhuocheng Gong, Jeff Good, Michael Goodman, Senthilkumar Gopal, Karthik Gopalakrishnan, Jonathan Gordon, Philip John Gorinski, Isao Goto, Yanjie Gou, Antoine Gourru, Cyril Goutte, Venkata Subrahmanyan Govindarajan, Edward Gow-Smith, Thamme Gowda, Kartik Goyal, Navita Goyal, Palash Goyal, Prasoon Goyal, Natalia Grabar, Mario Graff, Damien Graux, David Griol, Milan Gritta, Loïc Grobol, Stig-Arne Grönroos, David Gros, Adam Grycner, Jia-Chen Gu, Jiasheng Gu, Shuhao Gu, Yue Gu, Yuxian Gu, Saiping Guan, Yong Guan, Nuno M. Guerreiro, Liangke Gui, Vincent Guigue, Bruno Guillaume, Adrien Guille, Kalpa Gunaratna, James Gung, Tunga Gungor, Sharath Chandra Guntuku, Biyang Guo, Fengyu Guo, Han Guo, Jiang Guo, Jiaqi Guo, Jinyang Guo, Junliang Guo, Lin Guo, Meiqi Guo, Qipeng Guo, Quan Guo, Ruocheng Guo, Shaoru Guo, Shu Guo, Wangzhen Guo, Xin Guo, Xinnan Guo, Yanzhu Guo, Yinpeng Guo, Zhijiang Guo, Abhirut Gupta, Akshat Gupta, Amulya Gupta, Anchit Gupta, Ankit Gupta, Ankita Gupta, Ashim Gupta, Jai Gupta, Nitish Gupta, Prakhar Gupta, Raghav Gupta, Rishabh Gupta, Sonu Gupta, Sparsh Gupta, Umang Gupta, Vivek Gupta, Ximena Gutierrez-Vasques, Jeremy Gwinnup, Loitongbam Gyanendro Singh, Le An Ha, Nizar Habash, Kais Haddar, Katharina Haemmerl, Christopher Hahn, Joonghyuk Hahn, Michael Hahn, Zhen Hai, Jan Hajič, Eva Hajicova, Hossein Hajipour, Sherzod Hakimov, Kishaloy Halder, Anaïs Halftermeyer, Harald Hammarström, Michael Hammond, Thierry Hamon, Chengcheng Han, Chi Han, Hojae Han, Kelvin Han, Ridong Han, Rujun Han, Ting Han, Xiaochuang Han, Xiaohui Han, Xu Han, Xudong Han, Yo-Sub Han, Yu Han, Zhen Han, Zhongyuan Han, Chung-Wei Hang, Viktor Hangya, Jie Hao, Junheng Hao, Tianyong Hao, Rejwanul Haque, Syed Haque, Tatsuya Harada, David Harbecke, Momchil Hardalov, Daniel Hardt, Hardy Hardy, Keith Harrigian, William Hartmann, John Harvill, Sadid A. Hasan, Maram Hasanain, Taku Hasegawa, Chikara Hashimoto, Sabit Hassan, Bradley Hauer, Claudia Hauff, Shreya Havaldar, William Havard, Adi Haviv, Hiroaki Hayashi, Yoshihiko Hayashi, Amir Hazem, Ben He, Guoxiu He, Jacqueline He, Jianfeng He, Jiangen He, Jiayuan He, Jinzheng He, Kai He, Keqing He, Ru He, Shizhu He, Tianxing He, Wanwei He, Wei He, Xiaodong He, Xingwei He, Xuanli He, Yifan He, Yunjie He, Zexue He, Zhongjun He, Michael Heck, Behnam Hedayatnia, Michael Hedderich, Stefan Heindorf, Johannes Heinecke, Jindřich Helcl, William Held, Oliver Hellwig, Chadi Helwe, Christian Hempelmann, Lisa Anne Hendricks, Iris Hendrickx, Cui Hengbin, Leonhard Hennig, Yu-Jung Heo, David Herel, Delia Irazu Hernandez Farias, Christian Herold, Daniel Hershcovich, Jonathan Herzig, Christian Heumann, John Hewitt, Gerhard Heyer, Christopher Hidey, Derrick Higgins, Stefan Hillmann, Tsutomu Hirao, Tatsuya

Hiraoka, Namgyu Ho, Cong Duy Vu Hoang, Cuong Hoang, Julia Hockenmaier, Chris Hokamp, Samuel Hollands, Nora Hollenstein, Pavan Holur, Christopher Homan, Takeshi Homma, Ukyo Honda, Giwon Hong, Pengyu Hong, Zhi Hong, Mark Hopkins, Ales Horak, Andrea Horbach, Sho Hoshino, Tom Hosking, Md Mosharaf Hossain, Mohammad Javad Hosseini, Pedram Hosseini, Rasa Hosseinzadeh, Lei Hou, Yifan Hou, Phillip Howard, David M. Howcroft, Cheng-Yu Hsieh, Chao-Chun Hsu, Chun-Nan Hsu, I-Hung Hsu, Yi-Li Hsu, Phu Mon Htut, Chi Hu, Dou Hu, Guangneng Hu, Guimin Hu, Hai Hu, Hailin Hu, Han Hu, Hexiang Hu, Jinyi Hu, Linmei Hu, Mengting Hu, Minda Hu, Songbo Hu, Xiang Hu, Xiaodan Hu, Xuming Hu, Yibo Hu, Yuchen Hu, Yushi Hu, Zhe Hu, Zhiwei Hu, Zhiyuan Hu, Zikun Hu, Ziniu Hu, Hang Hua, Wenyue Hua, Xinyu Hua, Chao-Wei Huang, Chen Huang, Chieh-Yang Huang, Fei Huang, Hen-Hsen Huang, Hui Huang, Jen-Tse Huang, Jiaxin Huang, Jie Huang, Jimin Huang, Jimmy Huang, Jin-Xia Huang, Junjie Huang, Kuan-Hao Huang, Kung-Hsiang Huang, Luyang Huang, Qingbao Huang, Quzhe Huang, Rongjie Huang, Shaohan Huang, Tenghao Huang, Xinting Huang, Yinya Huang, Yongjie Huang, Youcheng Huang, Zhiqi Huang, Zhongqiang Huang, Luwen (vivian) Huangfu, Patrick Huber, John Hudzina, Pere-Lluís Huguet Cabot, Mans Hulden, Chia-Chien Hung, Fantine Huot, Ali Hürriyetoğlu, Tin Huynh, Rebecca Hwa, Dae Yon Hwang, Jena D. Hwang, Dongmin Hyun, Ignacio Iacobacci, Muhammad Okky Ibrohim, Adrian Iftene, Ryu Iida, Gabriel Ilharco, Nikolai Ilinykh, Kenji Imamura, Ayyoob Imanigooghari, Joseph Marvin Imperial, Hirofumi Inaguma, Mert Inan, Svanhvít Lilja Ingólfsdóttir, Koji Inoue, Takashi Inui, Hitoshi Isahara, Tatsuya Ishigaki, Etsuko Ishii, Aminul Islam, Tunazzina Islam, Masaru Isonuma, Takumi Ito, Abe Ittycheriah, Hamish Ivison, Tomoya Iwakura, Ran Iwamoto, Kenichi Iwatsuki, Vivek Iyer, Peter Izsak, Bassam Jabaian, Aashi Jain, Alankar Jain, Parag Jain, Rishabh Jain, Milos Jakubicek, Masoud Jalili Sabet, Shoaib Jameel, Richard James, Abhik Jana, Eugene Jang, Hyeju Jang, Myeongjun Jang, Youngsoo Jang, Sepehr Janghorbani, Peter Jansen, Maarten Janssen, Sujay Kumar Jauhar, Tommi Jauhiainen, Inigo Jauregi Unanue, Ganesh Jawahar, Sébastien Jean, Fran Jelenić, Sungho Jeon, Minwoo Jeong, Myeongho Jeong, Young-Seob Jeong, Kevin Jesse, Elisabetta Jezek, Akshita Jha, Prince Jha, Sneha Jha, Bin Ji, Haozhe Ji, Seunghyun Ji, Shaoxiong Ji, Ziwei Ji, Chen Jia, Qi Jia, Zixia Jia, Yiren Jian, Aiqi Jiang, Chao Jiang, Feng Jiang, Hang Jiang, Hao Jiang, Jie Jiang, Jiyue Jiang, Junfeng Jiang, Jyun-Yu Jiang, Lan Jiang, Lavender Jiang, Ming Jiang, Ridong Jiang, Tianwen Jiang, Tianyu Jiang, Wenbin Jiang, Xiaotong Jiang, Xuhui Jiang, Yichen Jiang, Yong Jiang, Yuxin Jiang, Zhengbao Jiang, Zhiwei Jiang, Zhiying Jiang, Zhuoren Jiang, Zhuoxuan Jiang, Cathy Jiao, Wenxiang Jiao, Yizhu Jiao, Zhanming Jie, Bernal Jimenez Gutierrez, Di Jin, Li Jin, Lisa Jin, Mali Jin, Qiao Jin, Shuning Jin, Woojeong Jin, Xiaomeng Jin, Yiping Jin, Zhi Jin, Zhijing Jin, Zijian Jin, Hwiyeol Jo, Richard Johansson, Kristen Johnson, Michael Johnston, Erik Jones, Kenneth Joseph, Abhinav Joshi, Aditya Joshi, Brihi Joshi, Nitish Joshi, Rishabh Joshi, Xincheng Ju, Yiming Ju, Zeqian Ju, Jaap Jumelet, Kyomin Jung, Myong Chol Jung, Taehee Jung, Juraj Juraska, David Jurgens, Raquel Justo, Prathyusha Jwalapuram, Preethi Jyothi, Vimal Kumar K, Kishan K C, Besim Kabashi, Srikanth Doss Kadarundalagi Raghuram Doss, Kazuma Kadowaki, Andrea Kahn, Magdalena Kaiser, Ivana Kajic, Tomoyuki Kajiwara, Mihir Kale, Oren Kalinsky, Laura Kallmeyer, Aikaterini-Lida Kalouli, Katikapalli Subramanyam Kalyan, Abu Raihan Kamal, Ehsan Kamalloo, Nishant Kambhatla, Hidetaka Kamigaito, Jaap Kamps, Hiroshi Kanayama, Kamil Kanclerz, Masahiro Kaneko, Gi-Cheon Kang, Jaewook Kang, Minki Kang, Yoshinobu Kano, Diptesh Kanojia, Pinar Karagoz, Giannis Karamanolakis, Siddharth Karamcheti, Mladen Karan, Akbar Karimi, Younes Karimi, Payam Karisani, Börje Karlsson, Shubhra Kanti Karmaker Santu, Sanjeev Kumar Karn, Constantinos Karouzos, Marzena Karpinska, Omid Kashefi, Zdeněk Kasner, Aly Kassem, Anisia Katinskaia, Yoav Katz, David Kauchak, Pride Kavumba, Noriaki Kawamae, Hideto Kazawa, Ashkan Kazemi, Ghazaleh Kazeminejad, Amirhossein Kazemnejad, Zixuan Ke, Akhil Kedia, Sedrick Scott Keh, Katherine Keith, Amr Keleg, Frank Keller, Casey Kennington, Tom Kenter, Roman Kern, Santosh Kesiraju, Lee Kezar, Shahram Khadivi, Muhammad Khalifa, Salam Khalifa, Anant Khandelwal, Dinesh Khandelwal, Shima Khanehzar, Simran Khanuja, Kyung Seo Ki, Mert Kilickaya, Halil Kilicoglu, Bugeun Kim, Gangwoo Kim, Gene Kim, Geonmin Kim, Gunhee Kim, Gyuhak Kim, Harksoo Kim, Hong Kook Kim, Hyounghun Kim, Hyunjae Kim, Hyunwoo Kim, Jaeyoung Kim, Jihyuk Kim, Jongwon Kim, Joo-Kyung Kim, Joshua Y. Kim,

Jung-Jae Kim, Kangil Kim, Kyungho Kim, Minsoo Kim, Sungdong Kim, Taeuk Kim, Yekyung Kim, Young Jin Kim, Youngwoo Kim, Yu Jin Kim, Yasutomo Kimura, Milton King, Tracy Holloway King, Svetlana Kiritchenko, Christo Kirov, Denis Kiselev, Hirokazu Kiyomaru, Shun Kiyono, Christopher Klamm, Ayal Klein, Tassilo Klein, Jan-Christoph Klie, Roman Klinger, Julien Kloetzer, Miyoung Ko, Goro Kobayashi, Hayato Kobayashi, Thomas Kober, Elena Kochkina, Jan Kocon, Prashant Kodali, Jordan Kodner, Arne Koehn, Rob Koeling, Svetla Koeva, Jing Yu Koh, Mare Koit, Noriyuki Kojima, Stanley Kok, Daan Kolkman, Anton Kolonin, Kazunori Komatani, Kanako Komiya, Grzegorz Kondrak, Cunliang Kong, Lingkai Kong, Miloslav Konopík, Ioannis Konstas, Selcuk Kopru, Michalis Korakakis, Katerina Korre, Ana Kotarcic, Suraj Kothawade, Fajri Koto, Neema Kotonya, Alexander Kotov, Manolis Koubarakis, Anna Koufakou, Vasiliki Kougia, Punit Singh Koura, Venelin Kovatchev, Ivan Koychev, Michael Kranzlein, Matthias Kraus, Simon Krek, Brigitte Krenn, Amrith Krishna, Kalpesh Krishna, Kundan Krishna, Adit Krishnan, Nikhil Krishnaswamy, Canasai Kruengkrai, Udo Kruschwitz, Anna Kruspe, Da Kuang, Andrei Kucharavy, Ilia Kulikov, Aditya Prakash Kulkarni, Ashish Kulkarni, Atharva Kulkarni, Vivek Kulkarni, Ashutosh Kumar, Puneet Kumar, Ritesh Kumar, Sachin Kumar, Sawan Kumar, Shankar Kumar, Shanu Kumar, Sumeet Kumar, Varun Kumar, Sadhana Kumaravel, Anoop Kunchukuttan, Adhiguna Kuncoro, Tsung-Ting Kuo, Yuri Kuratov, Murathan Kurfalı, Tatsuki Kuribayashi, Mikko Kurimo, Shuhei Kurita, Ugur Kursuncu, Guy Kushilevitz, Mucahid Kutlu, Ilia Kuznetsov, Haewoon Kwak, Sunjun Kweon, Yeonsu Kwon, Moreno La Quatra, Philippe Laban, Sofie Labat, Matthieu Labeau, Yanis Labrak, Faisal Ladhak, Katrien Laenen, Allison Lahnala, Huiyuan Lai, Kenneth Lai, Viet Lai, Yi-An Lai, Yuxuan Lai, Veronika Laippala, Surafel M. Lakew, Kushal Lakhotia, Yash Kumar Lal, Tsz Kin Lam, Wai Lam, Hemank Lamba, Vasileios Lampos, Gerasimos Lampouras, Nur Lan, Yunshi Lan, Lukas Lange, Maurice Langner, Mateusz Lango, Mirella Lapata, Issam Laradji, Samuel Larkin, Mikel Larrañaga, Stefan Larson, Samuel Läubli, Frances Adriana Laureano De Leon, Alberto Lavelli, Alexandra Lavrentovich, Dawn Lawrie, Phong Le, Joseph Le Roux, Kevin Leach, Gianluca Lebani, Lynda Lechani, Andrew Lee, Bruce W. Lee, Deokjae Lee, Dong-Ho Lee, Donghun Lee, Dongkyu Lee, Dongyub Lee, Fei-Tzin Lee, Gibbeum Lee, Grandee Lee, Hung-Yi Lee, Hwaran Lee, I-Ta Lee, Jackson Lee, Jae Hee Lee, Jae Sung Lee, Jay Yoon Lee, Jeong Min Lee, Ji-Ung Lee, Jihwan Lee, Jinhyuk Lee, John Lee, Jongwuk Lee, Jun-Min Lee, Koanho Lee, Kyumin Lee, Lung-Hao Lee, Mina Lee, Minho Lee, Minwoo Lee, Mong Li Lee, Nayeon Lee, Roy Ka-Wei Lee, Sang-Woo Lee, Seolhwa Lee, Wonkee Lee, Yongjae Lee, Yoonjoo Lee, Young-Suk Lee, Younghun Lee, Els Lefever, Joël Legrand, Jens Lemmens, Yves Lepage, Leo Leppänen, Pietro Lesci, Chun Wa Leung, Gregor Leusch, Ran Levy, Sharon Levy, Alexander Hanbo Li, Baoli Li, Bei Li, Belinda Z. Li, Bin Li, Bo Li, Bobo Li, Boyang Li, Changmao Li, Cheng Li, Cheng-Te Li, Chengming Li, Chenliang Li, Chong Li, Dianqi Li, Fangtao Li, Fei Li, Guanlin Li, Haizhou Li, Haochen Li, Haonan Li, Haoqi Li, Haoran Li, Haoran Li, Irene Li, Jiacheng Li, Jialu Li, Jiangnan Li, Jiangtong Li, Jiaqi Li, Jiaxuan Li, Jieyu Li, Jing Li, Jinpeng Li, Jiyi Li, Juanhui Li, Juncheng Li, Junyi Li, Junyi Li, Keyi Li, Lei Li, Li Erran Li, Liangyou Li, Linjie Li, Linyang Li, Liunian Harold Li, Maoxi Li, Margaret Li, Miao Li, Miaoran Li, Mingda Li, Mingjie Li, Mukai Li, Peifeng Li, Peiguang Li, Peng Li, Qian Li, Qintong Li, Ru Li, Rui Li, Ruifan Li, Ruizhe Li, Sha Li, Shaobo Li, Sheng Li, Shengjie Li, Shimin Li, Shiyang Li, Shuangyin Li, Shujun Li, Shuyang Li, Si Li, Siyan Li, Tao Li, Tianjian Li, Wei Li, Wei Li, Wenyan Li, Xia Li, Xiang Li, Xiang Lisa Li, Xiao Li, Xiaonan Li, Ximing Li, Xin Li, Xintong Li, Xinxin Li, Xue Li, Yafu Li, Yanran Li, Yanyang Li, Yanzeng Li, Yanzhou Li, Yaoyiran Li, Yinghui Li, Yingjie Li, Yingya Li, Yitong Li, Yiyuan Li, Yu Li, Yuan-Fang Li, Yucheng Li, Yuliang Li, Yuncong Li, Yunji Li, Zekun Li, Zhenhao Li, Zhi Li, Zhongli Li, Zongxi Li, Zuchao Li, Vladislav Lialin, Yixin Lian, Bin Liang, Chao-Chun Liang, Davis Liang, Di Liang, Hongru Liang, Junjie Liang, Miya Liang, Paul Pu Liang, Ping Liang, Sheng Liang, Yaobo Liang, Zhengzhong Liang, Zhenwen Liang, Zhicheng Liang, Baohao Liao, Lizi Liao, Peiyuan Liao, Siyu Liao, Jindřich Libovický, Veronica Liesaputra, Daniil Likhobaba, Gilbert Lim, Heuiseok Lim, Jungwoo Lim, Kwan Hui Lim, Tomasz Limisiewicz, Nut Limsopatham, Bingqian Lin, Bo Lin, Chuan-Jie Lin, Hongyu Lin, Huan Lin, Kevin Lin, King Ip Lin, Li Lin, Lucy Lin, Nankai Lin, Peiqin Lin, Qika Lin, Sheng-Chieh Lin, Ting-En Lin, Victoria Lin, Wei Lin, Weizhe Lin, Xiang Lin, Xinshi Lin, Yankai Lin,

Ying-Jia Lin, Yu-Hsiang Lin, Zeqi Lin, Zhaojiang Lin, Zhenxi Lin, Zhouhan Lin, Zi Lin, Matthias Lindemann, Jeffrey Ling, Zhenhua Ling, Tal Linzen, Marco Lippi, Pierre Lison, Diane Litman, Robert Litschko, Marina Litvak, Alisa Liu, Ao Liu, Bing Liu, Boyang Liu, Chen Liu, Chi-Liang Liu, Dayiheng Liu, Dexi Liu, Emmy Liu, Fangyu Liu, Fenglin Liu, Guangliang Liu, Guisheng Liu, Han Liu, Haokun Liu, Hui Liu, Hui Liu, Jiacheng Liu, Jiangming Liu, Jiawei Liu, Jiduan Liu, Jie-Jyun Liu, Jinglin Liu, Jingzhou Liu, Junhao Liu, Lei Liu, Linlin Liu, Linqing Liu, Luyang Liu, Ming Liu, Minqian Liu, Nayu Liu, Nelson F. Liu, Peng Liu, Qian Liu, Qian Liu, Qianying Liu, Shuaiqi Liu, Siyang Liu, Song Liu, Tianyuan Liu, Wenqiang Liu, Xianggen Liu, Xiangyang Liu, Xiao Liu, Xiao Liu, Xiaoyuan Liu, Xingxian Liu, Xuebo Liu, Xuye Liu, Yang Liu, Yang Janet Liu, Ye Liu, Ye Liu, Yijia Liu, Yiren Liu, Yixin Liu, Yizhu Liu, Yong Liu, Yongbin Liu, Yongfei Liu, Yonghao Liu, Yongkang Liu, Yuanxin Liu, Zechun Liu, Zeming Liu, Zequn Liu, Zeyu Liu, Zhe Liu, Zhenghao Liu, Zhengyuan Liu, Zhengzhong Liu, Zhijian Liu, Zihan Liu, Zitao Liu, Zuozhu Liu, Nikola Ljubešić, Kuan-Chieh Lo, Kyle Lo, Robert L Logan Iv, Lajanugen Logeswaran, Abhay Lokesh Kashyap, Damien Lolive, Guodong Long, Shangbang Long, Yunfei Long, Lucelene Lopes, Marcos Lopes, Henrique Lopes Cardoso, Oier Lopez De Lacalle, Adrian Pastor Lopez Monroy, Isabelle Lorge, Chao Lou, Jian-Guang Lou, Renze Lou, Natalia Loukachevitch, Anastassia Loukina, Daniel Loureiro, Nicholas Lourie, Pablo Loyola, Di Lu, Hongyuan Lu, Jianqiao Lu, Jinghui Lu, Jinliang Lu, Junru Lu, Pan Lu, Peng Lu, Weiming Lu, Wenpeng Lu, Xiaolei Lu, Yao Lu, Yaojie Lu, Yu Lu, Yu Lu, Yujie Lu, Nurul Lubis, Li Lucy, Stephanie M. Lukin, Gunnar Lund, Cheng Luo, Haoran Luo, Haozheng Luo, Hongyin Luo, Jiaming Luo, Jiebo Luo, Junyu Luo, Ling Luo, Man Luo, Renqian Luo, Ruipu Luo, Wencan Luo, Zhunchen Luo, Ziyang Luo, Kelvin Luu, Qi Lv, Chenyang Lyu, Weimin Lyu, Yajuan Lyu, Yougang Lyu, Meryem M'hamdi, Chenkai Ma, Chunpeng Ma, Congbo Ma, Danni Ma, Huifang Ma, Kaixin Ma, Longxuan Ma, Mingyu Derek Ma, Qianli Ma, Ruotian Ma, Tengfei Ma, Wei-Yun Ma, Weizhi Ma, Xinyin Ma, Yubo Ma, Yukun Ma, Zhanyu Ma, Ziqiao Ma, Dominik Macháček, Wolfgang Macherey, Jakub Macina, Aman Madaan, Avinash Madasu, Mounica Maddela, Brielen Madureira, Manuel Mager, Bernardo Magnini, Rahmad Mahendra, Ayush Maheshwari, Kyle Mahowald, Wolfgang Maier, Jean Maillard, Bodhisattwa Prasad Majumder, Navonil Majumder, Márton Makrai, Prodromos Malakasiotis, Ankur Mali, Itzik Malkiel, Anton Malko, Valentin Malykh, Jonathan Mamou, Arpan Mandal, Pranav Maneriker, Emma Manning, Irene Manotas, Elman Mansimov, Saab Mansour, Ramesh Manuvinakurike, Emaad Manzoor, Jiaxin Mao, Kelong Mao, Rui Mao, Wenji Mao, Yuning Mao, Zhendong Mao, Zhiming Mao, Zhuoyuan Mao, Piotr Mardziel, Katerina Margatina, Alex Marin, Mirko Marras, Edison Marrese-Taylor, Santiago Marro, Federico Martelli, Eugenio Martínez Cámara, Eva Martínez Garcia, Abelardo Carlos Martínez Lorenzo, Fernando Martínez-Plumed, Juan Martinez-Romo, Bruno Martins, Pedro Henrique Martins, David Martins De Matos, Luisa März, Laura Mascarell, Lambert Mathias, Sandeep Mathias, Sergio Matos, Yuichiroh Matsubayashi, Yuji Matsumoto, Takuya Matsuzaki, Evgeny Matusov, Borislav Mavrin, Jonathan May, Tobias Mayer, Joshua Maynez, Amir Mazaheri, Sahisnu Mazumder, Alessandro Mazzei, R. Thomas McCoy, Nick Mckenna, Paul Mcnamee, Quentin Meeus, Alexander Mehler, Ninareh Mehrabi, Nikhil Mehta, Sanket Vaibhav Mehta, Clara Meister, Dheeraj Mekala, Julia Mendelsohn, Erick Mendez Guzman, Arul Menezes, Telmo Menezes, Chuan Meng, Rui Meng, Yu Meng, Yuanliang Meng, Zhao Meng, Samuel Mensah, William Merrill, Mohsen Mesgar, Kourosh Meshgi, Eleni Metheniti, Lars Meyer, Adam Meyers, Ivan Vladimir Meza Ruiz, Yisong Miao, Alessio Miaschi, Antonio Valerio Miceli Barone, Timothee Mickus, Lesly Miculicich, Margot Mieskes, Todor Mihaylov, Nandana Mihindukulasooriya, Simon Mille, Timothy Miller, Hye-Jin Min, Koji Mineshima, Gosse Minnema, Andrei Mircea, Seyedabolghasem Mirroshandel, Paramita Mirza, Maryam Sadat Mirzaei, Abhijit Mishra, Pushkar Mishra, Shubhanshu Mishra, Siddhartha Mishra, Kanishka Misra, Masato Mita, Mitch Mithun, Ashish Mittal, Sarthak Mittal, Vibhu Mittal, Yasuhide Miura, Tong Mo, Yijun Mo, Daichi Mochihashi, Daniela Moctezuma, Ali Modarressi, Sandip Modha, Hans Moen, Aditya Mogadala, Nikita Moghe, Hosein Mohebbi, Behrang Mohit, Mrinal Mohit, Afroz Mohiuddin, Tasnim Mohiuddin, Michael Mohler, Luis Mojica De La Vega, Negar Mokhberian, Diego Molla, Nicholas Monath, Sneha Mondal, Helena Moniz, Ali Montazeralghaem, Manuel Montes, Johanna Monti, Hyeonseok Moon, Jihyung Moon,

Lori Moon, Raymond Mooney, Jared Moore, Richard Moot, Mehrad Moradshahi, Goncalo Mordido, Erwan Moreau, Antonio Moreno-Ortiz, Antonio Moreno-Sandoval, Mathieu Morey, Yusuke Mori, Véronique Moriceau, Emmanuel Morin, Gaku Morio, Makoto Morishita, John Morris, Marius Mosbach, Larry Moss, Xiangyang Mou, Maximilian Mozes, Frank Mtumbuka, Jesse Mu, Aaron Mueller, David Mueller, Aldrian Obaja Muis, Shashank Mujumdar, Animesh Mukherjee, Rajdeep Mukherjee, Matthew Mulholland, Benjamin Muller, Mathias Müller, Philippe Muller, Max Müller-Eberstein, Emir Munoz, Rafael Muñoz Guillena, Saliha Muradoglu, Koji Murakami, Deepak Muralidharan, Yugo Murawaki, Kenton Murray, Rudra Murthy, Shikhar Murty, Karthik Murugadoss, Skatje Myers, Agnieszka Mykowiecka, Sheshera Mysore, Anandhavelu N, Seung-Hoon Na, Nona Naderi, Seema Nagar, Masaaki Nagata, Aakanksha Naik, Saeed Najafi, Tetsuji Nakagawa, Yukiko Nakano, Yuta Nakashima, Hideki Nakayama, Christoforos Nalmpantis, Sungjin Nam, Marcin Namysl, Subhrangshu Nandi, Abhilash Nandy, Tarek Naous, Diane Napolitano, Jason Naradowsky, Sharan Narasimhan, Tahira Naseem, Sudip Naskar, Alexis Nasr, Vivi Nastase, Borja Navarro-Colorado, Tapas Nayak, Mojtaba Nayyeri, Claire Nedellec, Carina Negreanu, Preksha Nema, Joshua Nemecek, Graham Neubig, Guenter Neumann, Aurélie Névéol, Mariana Neves, Hwee Tou Ng, Axel-Cyrille Ngonga Ngomo, Cam Tu Nguyen, Dang Tuan Nguyen, Dat Quoc Nguyen, Dong Nguyen, Duc-Vu Nguyen, Huy Nguyen, Huyen Nguyen, Kiet Nguyen, Nhung Nguyen, Thanh Nguyen, Thanh-Tung Nguyen, Trang Nguyen, Truc-Vien T. Nguyen, Trung Hieu Nguyen, Vincent Nguyen, Hoang-Quoc Nguyen-Son, Ansong Ni, Jianmo Ni, Jingwei Ni, Minheng Ni, Zhaoheng Ni, Eric Nichols, Garrett Nicolai, Massimo Nicosia, Feng Nie, Ping Nie, Shaoliang Nie, Zhijie Nie, Sofia Nikiforova, Dmitry Nikolaev, Nikola I. Nikolov, Vassilina Nikoulina, Iftitahu Nimah, Lasguido Nio, Noriki Nishida, Masaaki Nishino, Sergiu Nisioi, Malvina Nissim, Tong Niu, Xing Niu, Yulei Niu, Zheng-Yu Niu, Bill Noble, Mariana Noguti, Tadashi Nomoto, Armineh Nourbakhsh, Jekaterina Novikova, Pierre Nugues, Diarmuid Ó Séaghdha, Alexander O'connor, Brendan O'connor, Tim O'gorman, Stephen Obadinma, Jose Ochoa-Luna, Kemal Oflazer, Maciej Ogrodniczuk, Kelechi Ogueji, Tolulope Ogunremi, Alice Oh, Shin Ah Oh, Mayumi Ohta, Kiyonori Ohtake, Atul Kr. Ojha, Oleg Okun, Eda Okur, Amy Olex, Anais Ollagnier, Ali Omrani, Byung-Won On, Donovan Ong, Ethel Ong, Yasumasa Onoe, Juri Opitz, Abigail Oppong, Matan Orbach, Hadas Orgad, Riccardo Orlando, John E. Ortega, Pedro Ortiz Suarez, Yohei Oseki, Naoki Otani, Zhijian Ou, Hiroki Ouchi, Nedjma Ousidhoum, Nedjma Ousidhoum, Jessica Ouyang, Iris Oved, Lilja Øvrelid, Kehinde Owoeye, Deepak P, Trilok Padhi, Ankur Padia, Vishakh Padmakumar, Gustavo Paetzold, Artidoro Pagnoni, Vardaan Pahuja, Santanu Pal, Vaishali Pal, Shriphani Palakodety, Chester Palen-Michel, Alexis Palmer, Alessio Palmero Aprosio, Shramay Palta, Junshu Pan, Xiang Pan, Xiaoman Pan, Yi-Cheng Pan, Youcheng Pan, Yu Pan, Yudai Pan, Artemis Panagopoulou, Alexander Panchenko, Mugdha Pandya, Liang Pang, Sheena Panthaplackel, Alessandro Panunzi, Isabel Papadimitriou, Pinelopi Papalampidi, Alexandros Papangelis, Nikos Papasarantopoulos, Paolo Papotti, Nikolaos Pappas, Emerson Paraiso, Bhargavi Paranjape, Letitia Parcalabescu, Antonio Pareja-Lora, Tanmay Parekh, Shantipriya Parida, Pierre-Henri Paris, Chaehun Park, Chan Young Park, Jun-Hyung Park, Jungsoo Park, Kunwoo Park, Seong-Bae Park, Seongmin Park, Seongsik Park, Shinwoo Park, Sunghyun Park, Sungjoon Park, Yannick Parmentier, Patrick Paroubek, Ankita Pasad, Lucia Passaro, Rebecca Passonneau, Ramakanth Pasunuru, Arkil Patel, Raj Patel, Roma Patel, Sapan Patel, Braja Gopal Patra, Jasabanta Patro, Parth Patwa, Manasi Patwardhan, Siddharth Patwardhan, Debjit Paul, Indraneil Paul, Shounak Paul, Adam Pauls, Nikita Pavlichenko, Ellie Pavlick, John Pavlopoulos, Siddhesh Pawar, Justin Payan, Pavel Pecina, Jiahuan Pei, Jiaxin Pei, Weiping Pei, Hao Peng, Hao Peng, Qianqian Peng, Qiwei Peng, Siyao Peng, Tao Peng, Wei Peng, Wei Peng, Wenjun Peng, Xutan Peng, Yifan Peng, Gerald Penn, Oren Pereg, Ethan Perez, Juan Antonio Perez-Ortiz, Gabriele Pergola, Charith Peris, Stanislav Peshterliev, Denis Peskoff, Ben Peters, Slav Petrov, Miriam R. L. Petruck, Pavel Petrushkov, Maxime Peyrard, Sandro Pezzelle, Jonas Pfeiffer, Quang Nhat Minh Pham, Thang Pham, Jason Phang, Maciej Piasecki, Massimo Piccardi, Matúš Pikuliak, Nisha Pillai, Tiago Pimentel, Juan Pino, Leticia Pinto-Alva, Irina Piontkovskaya, Telmo Pires, Flammie Pirinen, Jakub Piskorski, Lidia Pivovarova, Daniel Platt, Laura Plaza, Flor Miriam Plaza-Del-Arco, Lahari Poddar, Massimo Poesio, Thierry Poibeau, Lucie Polakova, Marco Polignano, Senja Pollak, Maria

Pontiki, Simone Paolo Ponzetto, Andrei Popescu-Belis, Maja Popović, Beatrice Portelli, Rafał Poświata, Martin Potthast, Christopher Potts, Amir Pouran Ben Veyseh, Rohit Prabhavalkar, Shrimai Prabhumoye, Aniket Pramanick, Soumajit Pramanik, Animesh Prasad, Radityo Eko Prasojo, Adithya Pratapa, Pavel Přibáň, Prokopis Prokopidis, Piotr Przybyła, Michal Ptaszynski, Dongqi Pu, Ratish Surendran Puduppully, Rajkumar Pujari, Stephen Pulman, Hemant Purohit, Alberto Purpura, Matthew Purver, James Pustejovsky, Valentina Pyatkin, Ehsan Qasemi, Fanchao Qi, Ji Qi, Jianzhong Qi, Jingyuan Qi, Shuhan Qi, Siya Qi, Wang Qi, Weizhen Qi, Chen Qian, Hongjin Qian, Jing Qian, Yujie Qian, Zhong Qian, Yaqiong Qiao, Bosheng Qin, Bowen Qin, Chuan Qin, Jinghui Qin, Kechen Qin, Libo Qin, Yujia Qin, Jielin Qiu, Liang Qiu, Long Qiu, Xinying Qiu, Zhaopeng Qiu, Zimeng Qiu, Chen Qu, Tingyu Qu, Rakesh R. Menon, Ella Rabinovich, Alexandre Rademaker, Daniele Radicioni, Alessandro Raganato, Preethi Raghavan, Dinesh Raghu, Afshin Rahimi, Sunny Rai, Vyas Raina, Nishant Raj, Navid Rajabi, Hossein Rajaby Faghihi, Dheeraj Rajagopal, Kanagasabai Rajaraman, Taraka Rama, Heri Ramampiaro, Naveen Raman, Giulia Rambelli, Owen Rambow, Abhinav Ramesh Kashyap, Sahana Ramnath, Rita Ramos, Alan Ramponi, Tharindu Ranasinghe, Surangika Ranathunga, Priya Rani, Yanghui Rao, Okko Rasanen, Mohammad Sadegh Rasooli, Fedor Ratnikov, Vikas Raunak, Andrea Amelio Ravelli, Shauli Ravfogel, Manikandan Ravikiran, Srinivas Ravishankar, Bhanu Pratap Singh Rawat, Vipula Rawte, Soumya Ray, Jishnu Ray Chowdhury, Manny Rayner, Anastasiia Razdaibiedina, Yasaman Razeghi, Evgeniia Razumovskaia, Livy Real, Traian Rebedea, Gabor Recski, Hanumant Redkar, Michael Regan, Ines Rehbein, Georg Rehm, Machel Reid, Markus Reiter-Haas, Navid Rekabsaz, Da Ren, Feiliang Ren, Haopeng Ren, Liliang Ren, Pengjie Ren, Ruiyang Ren, Shuhuai Ren, Steven Rennie, Christian Retoré, Kiamehr Rezaee, Mehdi Rezagholizadeh, Ryokan Ri, Eugénio Ribeiro, Leonardo F. R. Ribeiro, Giuseppe Riccardi, Kyle Richardson, Caitlin Richter, Martin Riedl, Stefan Riezler, Davide Rigoni, Mattia Rigotti, Shruti Rijhwani, Matīss Rikters, Fabio Rinaldi, Ruty Rinott, Annette Rios, Anthony Rios, Elijah Rippeth, Andrey Risukhin, Yara Rizk, Brian Roark, Alvaro Rodrigo, Melissa Roemmele, Morteza Rohanian, Mukesh Kumar Rohil, Mahdin Rohmatillah, Paul Roit, Lina M. Rojas Barahona, Roland Roller, Julia Romberg, Salvatore Romeo, Julien Romero, Srikanth Ronanki, Md Rashad Al Hasan Rony, Tanya Roosta, Rudolf Rosa, Domenic Rosati, Guy Rosin, Alexis Ross, Robert Ross, Sophie Rosset, Paolo Rosso, Guy Rotman, Hossein Rouhizadeh, Dmitri Roussinov, Rachel Edita Roxas, Aurko Roy, Shamik Roy, Soumyadeep Roy, Sumegh Roychowdhury, Jos Rozen, Antoine Rozenknop, Yu-Ping Ruan, Susanna Rücker, Koustav Rudra, Amina Rufai, Federico Ruggeri, Ramon Ruiz-Dolz, Mukund Rungta, Josef Ruppenhofer, Benjamin Ruppik, Thomas Ruprecht, Alexander Rush, Irene Russo, Piotr Rybak, Maciej Rybinski, Maria Ryskina, Hadeel Saadany, Arkadiy Saakyan, Caroline Sabty, Devendra Sachan, Fatiha Sadat, Farig Sadeque, Arka Sadhu, Philipp Sadler, Sahar Sadrizadeh, Mehrnoosh Sadrzadeh, Niloofar Safi Samghabadi, Sylvie Saget, Alsu Sagirova, Amrita Saha, Punyajoy Saha, Sougata Saha, Swarnadeep Saha, Tanay Kumar Saha, Tulika Saha, Saurav Sahay, Gözde Şahin, Mihir Sahoo, Sovan Kumar Sahoo, Sunil Kumar Sahu, Surya Kant Sahu, Ananya Sai B, Oscar Sainz, Tarek Sakakini, Sakriani Sakti, Ander Salaberria, Julian Salazar, Elizabeth Salesky, Jonne Saleva, Avneesh Saluja, Tanja Samardžić, Rajhans Samdani, Younes Samih, Iñaki San Vicente, Abhilasha Sancheti, Vicente Ivan Sanchez Carmona, Danae Sánchez Villegas, Víctor M. Sánchez-Cartagena, German Sanchis-Trilles, Mario Sänger, Ananth Sankar, Chinnadhurai Sankar, Scott Sanner, Sashank Santhanam, Andrea Santilli, Diana Santos, Rodrigo Santos, Bishal Santra, Sebastin Santy, Soumya Sanyal, Maarten Sap, Naomi Saphra, Ruhi Sarikaya, Efsun Sarioglu Kayi, Anoop Sarkar, Kamal Sarkar, Ritesh Sarkhel, Prathusha K Sarma, Prof. Shikhar Kumar Sarma, Gabriele Sarti, Kengatharaiyer Sarveswaran, Sheikh Sarwar, Felix Sasaki, Minoru Sasaki, Shota Sasaki, Ryohei Sasano, Giorgio Satta, Danielle Saunders, Ketki Savle, Guergana Savova, Apoorv Saxena, Michael Saxon, Asad Sayeed, Shigehiko Schamoni, Wout Schellaert, Frank Schilder, David Schlangen, Viktor Schlegel, Michael Sejr Schlichtkrull, Jörg Schlötterer, Helmut Schmid, Robin Schmidt, Patricia Schmidtova, Martin Schmitt, Tyler Schnoebelen, Stephanie Schoch, Annika Marie Schoene, Mirco Schoenfeld, Lenhart Schubert, Hendrik Schuff, William Schuler, Sabine Schulte Im Walde, Claudia Schulz, Hannes Schulz, Elliot Schumacher, Raphael Schumann, Sebastian Schuster, Ineke Schuurman, Jackson Scott, Kyle Seelman, Ethan Selfridge, Thibault Sellam, David Semedo,

Nasredine Semmar, Cansu Sen, Srinivasan Sengamedu Hanumantha Rao, Ayan Sengupta, Shubhashis Sengupta, Rico Sennrich, Jaehyung Seo, Ronald Seoh, Yeon Seonwoo, Royal Sequiera, Sofia Serrano, Mahsa Shafaei, Stephen Shaffran, Simra Shahid, Omar Shaikh, Igor Shalyminov, Chao Shang, Mingyue Shang, Chenze Shao, Wei Shao, Yijia Shao, Yutong Shao, Ori Shapira, Aditya Sharma, Ashish Sharma, Piyush Sharma, Serge Sharoff, Tatiana Shavrina, Shuaijie She, Artem Shelmanov, Aili Shen, Hua Shen, Jiaming Shen, Jianhao Shen, Sheng Shen, Shiqi Shen, Siqi Shen, Tianhao Shen, Xudong Shen, Yatian Shen, Ying Shen, Yongliang Shen, Yuming Shen, Zejiang Shen, Zhengyuan Shen, Emily Sheng, Qiang Sheng, Ashish Shenoy, Tom Sherborne, Botian Shi, Bowen Shi, Chen Shi, Jihao Shi, Kaize Shi, Ning Shi, Peng Shi, Tian Shi, Tianze Shi, Weijia Shi, Xiao Shi, Yangyang Shi, Zhan Shi, Zhouxing Shi, Tomohide Shibata, Hidetoshi Shimodaira, Jamin Shin, Seungjae Shin, Kazutoshi Shinoda, Takahiro Shinozaki, Keiji Shinzato, Prashant Shiralkar, Yow-Ting Shiue, Harry Shomer, Ziyi Shou, Mohit Shridhar, Ritvik Shrivastava, Dimitar Shterionov, Kai Shu, Raphael Shu, Kai Shuang, Zeren Shui, Alexander Shvets, Chenglei Si, Suzanna Sia, Anthony Sicilia, A.b. Siddique, Melanie Siegel, Ingo Siegert, Alejandro Sierra-Múnera, Ankur Sikarwar, Sandipan Sikdar, Andrew Silva, João Ricardo Silva, Danilo Silva De Carvalho, Fabrizio Silvestri, Stefano Silvestri, Robert Sim, Michel Simard, Patrick Simianer, Dharani Simma, Dan Simonson, Edwin Simpson, Jyotika Singh, Mayank Singh, Pranaydeep Singh, Thoudam Doren Singh, Sneha Singhania, Priyanka Sinha, Olivier Siohan, Amy Siu, Inguna Skadiņa, Gabriel Skantze, Victor Skobov, Aviv Slobodkin, Alisa Smirnova, David Smith, Noah A. Smith, Vésteinn Snæbjarnarson, Felipe Soares, Marco Antonio Sobrevilla Cabezudo, Artem Sokolov, Luca Soldaini, Amir Soleimani, Ilia Sominsky, Pia Sommerauer, Junyoung Son, Seonil (simon) Son, Youngseo Son, Haiyue Song, Haoyu Song, Hyeonho Song, Hyun-Je Song, Kai Song, Kaiqiang Song, Kaitao Song, Linfeng Song, Ran Song, Wei Song, Xiaohui Song, Yan Song, Yangqiu Song, Yifan Song, Zhenqiao Song, Sarvesh Soni, Shashank Sonkar, Taylor Sorensen, Ionut-Teodor Sorodoc, Alexey Sorokin, Daniil Sorokin, Anna Sotnikova, Xabier Soto, Sajad Sotudeh, Gerasimos Spanakis, Manuela Speranza, Andreas Spitz, Richard Sproat, Rachele Sprugnoli, Makesh Narsimhan Sreedhar, Mukund Srinath, Kavya Srinet, Balaji Vasan Srinivasan, Tejas Srinivasan, Vijay Srinivasan, Ankit Srivastava, Saurabh Srivastava, Efstathios Stamatatos, Dominik Stammbach, Karolina Stanczak, Marija Stanojevic, Gabriel Stanovsky, Katherine Stasaski, Manfred Stede, Julius Steen, Michal Štefánik, Shane Steinert-Threlkeld, Georg Stemmer, Evgeny Stepanov, Zachary Stine, Regina Stodden, Niklas Stoehr, Alessandro Stolfo, Matthew Stone, Shane Storks, Kevin Stowe, Marco Antonio Stranisci, Karl Stratos, Kristina Striegnitz, Phillip Ströbel, David Strohmaier, Jannik Strötgen, Tomek Strzalkowski, Sara Stymne, Dan Su, Hsuan Su, Qi Su, Qinliang Su, Ruolin Su, Xin Su, Ying Su, Yixuan Su, Yusheng Su, Nishant Subramani, Katsuhito Sudoh, Saku Sugawara, Hiroaki Sugiyama, Kazunari Sugiyama, Yoshi Suhara, Zhifang Sui, Octavia Şulea, Elior Sulem, Md Arafat Sultan, Aixin Sun, Changzhi Sun, Chengjie Sun, Chenkai Sun, Guangzhi Sun, Haipeng Sun, Hao Sun, Haohai Sun, Jian Sun, Jiao Sun, Ming Sun, Mingwei Sun, Qingfeng Sun, Renliang Sun, Shichao Sun, Simeng Sun, Tianxiang Sun, Weiwei Sun, Zewei Sun, Zhaoyue Sun, Zhiqing Sun, Dhanasekar Sundararaman, Mujeen Sung, Yi-Lin Sung, Yoo Yeon Sung, Hanna Suominen, Marek Suppa, Benjamin Suter, Mirac Suzgun, Sandesh Swamy, Stan Szpakowicz, Piotr Szymański, Anaïs Tack, Oyvind Tafjord, Shabnam Tafreshi, Dima Taji, Sho Takase, Ece Takmaz, George Tambouratzis, Aleš Tamchyna, Aniruddha Tammewar, Akihiro Tamura, Chao-Hong Tan, Haochen Tan, Hongye Tan, Samson Tan, Wei Tan, Xiao Tan, Ryota Tanaka, Karan Taneja, Buzhou Tang, Chengguang Tang, Gongbo Tang, Hao Tang, Jialong Tang, Raphael Tang, Shuai Tang, Tianyi Tang, Wei Tang, Xiangyun Tang, Xuemei Tang, Xunzhu Tang, Yun Tang, Yun Tang, Zheng Tang, Simon Tannert, Chaofan Tao, Wei Tao, Allahsera Auguste Tapo, Shiva Taslimipoor, Sandeep Tata, Michiaki Tatsubori, Marta Tatu, Simone Tedeschi, Selma Tekir, Serra Sinem Tekiroğlu, Zhiyang Teng, Ian Tenney, Alberto Testoni, Joel Tetreault, Martin Teuffenbach, Kapil Thadani, Katherine Thai, Urmish Thakker, Surendrabikram Thapa, Avijit Thawani, Anton Thielmann, Krishnaprasad Thirunarayan, Brian Thompson, Jana Thompson, Craig Thomson, Sam Thomson, David Thulke, Chang Tian, Yuanhe Tian, Zhiliang Tian, Zuoyu Tian, Jörg Tiedemann, Christoph Tillmann, Tiago Timponi Torrent, Prayag Tiwari, Amalia Todirascu, Nadi Tomeh, Nicholas Tomlin, Antonio Toral, Cagri Toraman, Manabu Torii,

Kentaro Torisawa, Juan-Manuel Torres-Moreno, Lucas Torroba Hennigen, Shubham Toshniwal, Samia Touileb, Yannick Toussaint, Benjamin Towle, Amine Trabelsi, Khanh Tran, Trang Tran, Marcos Treviso, Jan Trienes, Bayu Distiawan Trisedya, Harsh Trivedi, Enrica Troiano, Chen-Tse Tsai, Adam Tsakalidis, Bo-Hsiang Tseng, Ioannis Tsiamas, Masaaki Tsuchida, Oren Tsur, Satoshi Tsutsui, Jingxuan Tu, Kewei Tu, Lifu Tu, Yunbin Tu, Yi-Lin Tuan, Marco Turchi, Ferhan Ture, Elena Tutubalina, Rutuja Ubale, Ana Sabina Uban, Adrian Ulges, Eddie Ungless, Bhargav Upadhyay, Kartikeya Upasani, Olga Uryupina, Asahi Ushio, Dmitry Ustalov, Ahmet Üstün, Masao Utiyama, Venktesh V, Saujas Vaduguru, Ashwini Vaidya, Marco Valentino, Gisela Vallejo, Jannis Vamvas, Tim Van De Cruys, Antal Van Den Bosch, Rob Van Der Goot, Daan Van Esch, Josef Van Genabith, Emiel Van Miltenburg, Rik Van Noord, Vincent Vandeghinste, Keith Vanderlinden, David Vandyke, Natalia Vanetik, Eva Vanmassenhove, Daniel Varab, Francielle Vargas, Siddharth Varia, Neeraj Varshney, Rossella Varvara, Siddharth Vashishtha, Jake Vasilakes, Eva Maria Vecchi, Nikhita Vedula, Aswathy Velutharambath, Giulia Venturi, Gaurav Verma, Rakesh Verma, Yannick Versley, Anvesh Rao Vijjini, David Vilares, Jesús Vilares, Manuel Vilares Ferro, Martina Vilas, Veronika Vincze, Lucas Vinh Tran, Sami Virpioja, Juraj Vladika, Nikolai Vogler, Rob Voigt, Pius Von Däniken, Spencer Von Der Ohe, Nikos Voskarides, Ali Vosoughi, Pavlos Vougiouklis, Thuy Vu, Thuy-Trang Vu, Yogarshi Vyas, Akifumi Wachi, Takashi Wada, Joachim Wagner, Jan Philip Wahle, Hiromi Wakaki, David Wan, Stephen Wan, Xingchen Wan, Yao Wan, Yu Wan, Ante Wang, Bailin Wang, Bang Wang, Baoxin Wang, Baoxun Wang, Beilun Wang, Benyou Wang, Bin Wang, Bin Wang, Bingqing Wang, Bingyu Wang, Bo Wang, Bo Wang, Boxin Wang, Chao Wang, Chengyi Wang, Chengyu Wang, Chuan-Ju Wang, Chunliu Wang, Cunxiang Wang, Dingquan Wang, Fei Wang, Guangrun Wang, Guoyin Wang, Hai Wang, Han Wang, Han Wang, Han Wang, Hanrui Wang, Hao Wang, Haobo Wang, Haoyu Wang, Haoyu Wang, Haoyu Wang, Heyuan Wang, Hong Wang, Hongfei Wang, Hsin-Min Wang, Huimin Wang, Jiaan Wang, Jian Wang, Jianing Wang, Jianyu Wang, Jianzong Wang, Jiayi Wang, Jie Wang, Jin Wang, Jin Wang, Jinpeng Wang, Jue Wang, Jun Wang, Lei Wang, Liang Wang, Lidan Wang, Lingzhi Wang, Longshaokan Wang, Longyue Wang, Meiqi Wang, Peifeng Wang, Pidong Wang, Ping Wang, Qiang Wang, Qingyun Wang, Qiqi Wang, Rui Wang, Rui Wang, Runze Wang, Ryan Wang, Shufan Wang, Shuhe Wang, Shuo Wang, Sijia Wang, Sirui Wang, Tao Wang, Tianduo Wang, Tianlu Wang, Wei Wang, Wen Wang, Wenping Wang, Wenxuan Wang, Wenya Wang, Xiangdong Wang, Xiao Wang, Xiaojie Wang, Xiaolin Wang, Xiaozhi Wang, Xin Wang, Xindi Wang, Xing Wang, Xingjin Wang, Xintong Wang, Xinyi Wang, Xinyu Wang, Xuewei Wang, Xun Wang, Yan Wang, Yanlin Wang, Yanshan Wang, Ye Wang, Ye Wang, Yibo Wang, Yifan Wang, Yigong Wang, Yihan Wang, Yiwei Wang, Yizhong Wang, Yue Wang, Yun Cheng Wang, Yuxuan Wang, Zekun Wang, Zhaowei Wang, Zhen Wang, Zheng Wang, Zheng Wang, Zhenhailong Wang, Zhenyi Wang, Zhichun Wang, Zhiguang Wang, Zhilin Wang, Zhiqiang Wang, Zhiwei Wang, Zhuoer Wang, Zhuoyi Wang, Zifeng Wang, Zihan Wang, Zihao Wang, Zijian Wang, Zijie Wang, Zilong Wang, Zirui Wang, Prashan Wanigasekara, Leo Wanner, Alex Warstadt, Cedric Waterschoot, Julia Watson, Bonnie Webber, Leon Weber, Albert Webson, Kellie Webster, Tharindu Cyril Weerasooriya, Chengkun Wei, Jerry Wei, Lingwei Wei, Penghui Wei, Tianxin Wei, Xiangpeng Wei, Xiaochi Wei, Shira Wein, Nathaniel Weir, Henry Weld, Orion Weller, Marion Weller-Di Marco, Simon Wells, Bingyang Wen, Haoyang Wen, Jiaxin Wen, Liang Wen, Lijie Wen, Rongxiang Weng, Lukas Wertz, Peter West, Matthijs Westera, Jennifer C. White, Richard Wicentowski, Michael Wiegand, Ethan Wilcox, Rodrigo Wilkens, Bram Willemsen, Ronald Wilson, Shomir Wilson, Steven Wilson, Grégoire Winterstein, Shuly Wintner, Sam Wiseman, Guillaume Wisniewski, Emilia Wisnios, Tomer Wolfson, Marcin Woliński, Diedrich Wolter, Derek F. Wong, Ka Ho Wong, Raymond Wong, Tak-Lam Wong, Alina Wróblewska, Anna Wroblewska, Anne Wu, Bowen Wu, Changxing Wu, Chen Wu, Chen Henry Wu, Chien-Sheng Wu, Chuhan Wu, Di Wu, Di Wu, Fangzhao Wu, Hua Wu, Hui Wu, Junda Wu, Ledell Wu, Lianwei Wu, Linzhi Wu, Shengqiong Wu, Shih-Hung Wu, Sixing Wu, Stephen Wu, Te-Lin Wu, Tianxing Wu, Ting-Wei Wu, Weibin Wu, Wenhao Wu, Winston Wu, Xian Wu, Xianchao Wu, Xin Wu, Xixin Wu, Yang Wu, Yangjun Wu, Yaoyao Wu, Yike Wu, Yimeng Wu, Youzheng Wu, Yu Wu, Yuanbin Wu, Yuexin Wu, Yunfang Wu, Yuting Wu, Yuxiang Wu, Zeqiu Wu, Zhaofeng Wu, Zhen Wu, Zhijing Wu, Zhiyong Wu,

Zhiyong Wu, Zhizheng Wu, Zhuofeng Wu, Zihao Wu, Zixiu Wu, Jian Xi, Zhaohan Xi, Fei Xia, Menglin Xia, Mengzhou Xia, Patrick Xia, Qingrong Xia, Yingce Xia, Anhao Xiang, Jiannan Xiang, Suncheng Xiang, Changrong Xiao, Chaojun Xiao, Chunyang Xiao, Jinfeng Xiao, Jing Xiao, Jinghui Xiao, Min Xiao, Yanghua Xiao, Zhaomin Xiao, Jun Xie, Kaige Xie, Ning Xie, Ruobing Xie, Shangyu Xie, Yiqing Xie, Yuqing Xie, Yuxi Xie, Zhiwen Xie, Zhouhang Xie, Ji Xin, Chen Xing, Linzi Xing, Zhenchang Xing, Bo Xiong, Chao Xiong, Jing Xiong, Kai Xiong, Wenhan Xiong, Binfeng Xu, Boyan Xu, Canwen Xu, Chen Xu, Chenchen Xu, Chunpu Xu, Dongfang Xu, Fan Xu, Fangyuan Xu, Frank F. Xu, Guandong Xu, Guangyue Xu, Hanzi Xu, Hongfei Xu, Hongzhi Xu, Jiacheng Xu, Jiashu Xu, Jin Xu, Jinan Xu, Jitao Xu, Jun Xu, Kang Xu, Keyang Xu, Kun Xu, Lei Xu, Lu Xu, Mingbin Xu, Mingzhou Xu, Nan Xu, Peng Xu, Peng Xu, Qiongkai Xu, Ruifeng Xu, Ruochen Xu, Shicheng Xu, Wang Xu, Weiran Xu, Weiwen Xu, Wenda Xu, Wenduan Xu, Xiao Xu, Xinnuo Xu, Yan Xu, Yang Xu, Yang Xu, Yi Xu, Yige Xu, Yiheng Xu, Yumo Xu, Zhen Xu, Zhenhui Xu, Zhichao Xu, Zhiyang Xu, Fuzhao Xue, Nianwen Xue, Shan Xue, Deshraj Yadav, Prateek Yadav, Yadollah Yaghoobzadeh, Bryce Yahn, Ikuya Yamada, Ivan Yamshchikov, An Yan, Hang Yan, Hanqi Yan, Jianhao Yan, Jun Yan, Lingyong Yan, Xifeng Yan, Xu Yan, Zhao Yan, Hitomi Yanaka, An Yang, Cheng Yang, Dejie Yang, Eugene Yang, Fan Yang, Guanqun Yang, Haoran Yang, Jian Yang, Jian Yang, Jianing Yang, Jie Yang, Jingfeng Yang, Jun Yang, Kexin Yang, Li Yang, Liner Yang, Linyi Yang, Liu Yang, Longfei Yang, Nan Yang, Sen Yang, Songlin Yang, Tsung-Yen Yang, Wei Yang, Wenmian Yang, Xianjun Yang, Xiaocong Yang, Yaqin Yang, Yazheng Yang, Yiben Yang, Yinfei Yang, Yuanhang Yang, Yue Yang, Zhao Yang, Zixiaofan Yang, Zonglin Yang, Ken Yano, Tae Yano, Barry Yao, Bingsheng Yao, Liang Yao, Peiran Yao, Zijun Yao, Mahsa Yarmohammadi, Bingyang Ye, Fanghua Ye, Hai Ye, Jiacheng Ye, Jiasheng Ye, Junjie Ye, Muchao Ye, Qinyuan Ye, Rong Ye, Seonghyeon Ye, Wei Ye, Wenting Ye, Xi Ye, An-Zi Yen, Jinyoung Yeo, Yu Ting Yeung, Jingwei Yi, Xiaoyuan Yi, Wen-Wai Yim, Seid Muhie Yimam, Chuantao Yin, Congchi Yin, Fan Yin, Kayo Yin, Qingyu Yin, Wenjie Yin, Xuwang Yin, Yu Yin, Yuwei Yin, Jiahao Ying, Anssi Yli-Jyra, Michael Yoder, Hikaru Yokono, Zheng Xin Yong, Kiyoon Yoo, Soyeop Yoo, Seunghyun Yoon, Sunjae Yoon, Susik Yoon, Wonjin Yoon, Naoki Yoshinaga, Koichiro Yoshino, Chenyu You, Haoxuan You, Weiqiu You, Steve Young, Tom Young, Abdou Youssef, Bei Yu, Changlong Yu, Cheng Yu, Dian Yu, Dong Yu, Heng Yu, Jianfei Yu, Jifan Yu, Liang-Chih Yu, Nan Yu, Ning Yu, Pengfei Yu, Philip Yu, Shoubin Yu, Tao Yu, Tiezheng Yu, Tong Yu, Wenhao Yu, Xiaodong Yu, Xinchen Yu, Yue Yu, Zac Yu, Zhiwei Yu, Bo Yuan, Caixia Yuan, Chenhan Yuan, Fei Yuan, Jianhua Yuan, Lifan Yuan, Nicholas Jing Yuan, Xiaojie Yuan, Ye Yuan, Zheng Yuan, Chuan Yue, Tianwei Yue, Mert Yuksekgonul, Hyeongu Yun, Frances Yung, Polina Zablotskaia, Ofir Zafrir, Wajdi Zaghouani, Hamada Zahera, Nasser Zalmout, Olga Zamaraeva, Roberto Zamparelli, Fabio Massimo Zanzotto, Alessandra Zarcone, Sina Zarrieß, Vicky Zayats, Albin Zehe, Eric Zelikman, Yury Zemlyanskiy, Jiali Zeng, Jiandian Zeng, Kaisheng Zeng, Qi Zeng, Qingkai Zeng, Weixin Zeng, Xingshan Zeng, Yan Zeng, Yawen Zeng, Ziqian Zeng, Deniz Zeyrek, Zenan Zhai, Haolan Zhan, Jingtao Zhan, Pengwei Zhan, Runzhe Zhan, Aston Zhang, Biao Zhang, Boliang Zhang, Bowen Zhang, Bowen Zhang, Chao Zhang, Chen Zhang, Chen Zhang, Chenwei Zhang, Chiyu Zhang, Dan Zhang, Duzhen Zhang, Fan Zhang, Ge Zhang, Hainan Zhang, Hao Zhang, Haopeng Zhang, Hongkuan Zhang, Jieyu Zhang, Jinchao Zhang, Jingqing Zhang, Junchi Zhang, Junwen Zhang, Kai Zhang, Ke Zhang, Kechi Zhang, Kun Zhang, Lei Zhang, Li Zhang, Licheng Zhang, Lingyu Zhang, Lining Zhang, Longhui Zhang, Meng Zhang, Michael Zhang, Mike Zhang, Min Zhang, Qi Zhang, Qiao Zhang, Ran Zhang, Richong Zhang, Rongsheng Zhang, Ruisi Zhang, Ruiyi Zhang, Ruohong Zhang, Ruoyu Zhang, Shaokun Zhang, Shaolei Zhang, Sheng Zhang, Shiyue Zhang, Shuai Zhang, Shujian Zhang, Shuo Zhang, Songming Zhang, Tianlin Zhang, Tianyi Zhang, Tongtao Zhang, Wei Zhang, Wei Emma Zhang, Wen Zhang, Wenqiang Zhang, Wenxuan Zhang, Xiangliang Zhang, Xiaojun Zhang, Xiaoqiang Zhang, Xin Zhang, Xinbo Zhang, Xinliang Frederick Zhang, Xinsong Zhang, Xuanwei Zhang, Yan Zhang, Yan Zhang, Yanzhe Zhang, Yichi Zhang, Yifan Zhang, Yiming Zhang, Ying Zhang, Yong Zhang, Yu Zhang, Yuan Zhang, Yuan Zhang, Yuanzhe Zhang, Yue Zhang, Yuhao Zhang, Yuhui Zhang, Yun Zhang, Yunqi Zhang, Yunxiang Zhang, Yunyi Zhang, Yuqi Zhang, Yusen Zhang, Yuxiang Zhang, Yuxin Zhang, Zequn Zhang, Zhengkun Zhang, Zhengyan Zhang, Zhexin Zhang, Zhi-

han Zhang, Zhirui Zhang, Zhisong Zhang, Zhiyuan Zhang, Zhuosheng Zhang, Bing Zhao, Chao Zhao, Fei Zhao, Guangxiang Zhao, Jeffrey Zhao, Jiahao Zhao, Jianyu Zhao, Jieyu Zhao, Jinming Zhao, Kai Zhao, Kaiqi Zhao, Mengjie Zhao, Qinghua Zhao, Ruihui Zhao, Sanqiang Zhao, Shuai Zhao, Shuai Zhao, Tiancheng Zhao, Tianyu Zhao, Wenting Zhao, Xiaoyan Zhao, Xinran Zhao, Xueliang Zhao, Yang Zhao, Yangyang Zhao, Yiyun Zhao, Yu Zhao, Yunlong Zhao, Zhixue Zhao, Zhuanzhe Zhao, Boyuan Zheng, Changmeng Zheng, Chujie Zheng, Jing Zheng, Junhao Zheng, Kai Zheng, Rui Zheng, Xianrui Zheng, Xiaosen Zheng, Xin Zheng, Xinyi Zheng, Yinhe Zheng, Zaixiang Zheng, Jialun Zhong, Ming Zhong, Ruiqi Zhong, Victor Zhong, Wanjun Zhong, Yang Zhong, Zexuan Zhong, Baohang Zhou, Ben Zhou, Daniel Xiaodan Zhou, Deyu Zhou, Dong Zhou, Giulio Zhou, Guangyou Zhou, Han Zhou, Jiawei Zhou, Jiawei Zhou, Jie Zhou, Jinfeng Zhou, Jingbo Zhou, Jingyan Zhou, Junpei Zhou, Junsheng Zhou, Kankan Zhou, Kun Zhou, Lexin Zhou, Pei Zhou, Peilin Zhou, Peng Zhou, Qiang Zhou, Qingyu Zhou, Shuchang Zhou, Shuyan Zhou, Tong Zhou, Wangchunshu Zhou, Wenjie Zhou, Wenxuan Zhou, Xiang Zhou, Xiangyang Zhou, Xixi Zhou, Yangqiaoyu Zhou, Yi Zhou, Yi Zhou, Yichao Zhou, Yichu Zhou, Yucheng Zhou, Yufan Zhou, Zhengyu Zhou, Zhihan Zhou, Zhong Zhou, Dawei Zhu, Fangwei Zhu, Haichao Zhu, Kenny Zhu, Linchao Zhu, Luyao Zhu, Muhua Zhu, Pengcheng Zhu, Qi Zhu, Qiannan Zhu, Qingfu Zhu, Su Zhu, Suyang Zhu, Tong Zhu, Wang Zhu, Wanzheng Zhu, Wei Zhu, Wenhao Zhu, Xiaodan Zhu, Xiaofeng Zhu, Xuan Zhu, Yilun Zhu, Yutao Zhu, Zining Zhu, Fuzhen Zhuang, Honglei Zhuang, Yimeng Zhuang, Yuan Zhuang, Yuchen Zhuang, Caleb Ziems, Leonardo Zilio, Heike Zinsmeister, Ayah Zirikly, Yftah Ziser, Imed Zitouni, Shi Zong, Bowei Zou, Wei Zou, Yicheng Zou, Amal Zouaq, Vilém Zouhar, Andrej Zukov Gregoric, Simiao Zuo, Xinyu Zuo

Secondary Reviewers

Sharon Adar, Sneha Agarwal, Utkarsh Agarwal, Akiko Aizawa, Christopher Akiki, Ilseyar Alimova, Falah Amro, Miriam Anschütz, William Armstrong, Yuya Asano, Md Rabiul Awal, Ansar Aynetdinov, Andrea Bacciu, Yinhao Bai, Oliver Baumann, Alessandro De Bellis, Guillaume Le Berre, Marie Bexte, Hanoz Bhathena, Abari Bhattacharya, Mukul Bhutani, Verena Blaschke, Moritz Blum, Marc Brinner, Reynier Ortega Bueno, Kishan K C, Mingchen Cai, Yucheng Cai, Paul Caillon, Eduardo Calò, Marco Casavantes, Giulia Cassara, Roman Castagné, Brittany Cates, Amanda Chan, Ayon Chattopadhyay, Huiyao Chen, Liang Chen, Pei Chen, Tianyu Chen, Tongfei Chen, Weidong Chen, Xi Chen, Xingyu Chen, Yuan Chen, Yue Chen, Zhenghan Chen, Zhi Chen, Zhijia Chen, Zhikai Chen, Zifeng Cheng, Jae Sook Cheong, Lin Lee Cheong, Yan Kin Chi, Hanjun Cho, Eunsenog Choi, Sahil Chopra, Rennan Cordeiro, Matthias Cosler, Adrian Cosma, Liam Cripwell, Yudivián Almeida Cruz, Israel Cuevas, Shih-Chieh Dai, Yinpei Dai, Parag Dakle, Niklas Deckers, Zhongfen Deng, Sourabh Deoghare, Simma Dharani, Harshita Diddee, Qiuyu Ding, Yuning Ding, Zixiang Ding, Mingwen Dong, Kefei Duan, Fanny Ducel, Tobias Eder, Pavel Efimov, Suilan Estevez-Velarde, Saad Ezzini, Maurice Falk, Meng Fan, Ziwei Fan, Qingkai Fang, Mohsen Fayyaz, James Finch, Sarah Finch, Sheema Firdous, Martina Forster, Cady Gansen, Alberto Gasparin, Qiming Ge, Shiping Ge, Kinga Gémes, Lei Geng, Yaroslav Getman, Sadaf Ghaffari, Sarvjeet Singh Ghotra, Lukas Gienapp, Jonas Golde, Mahsa Goodarzi, Shuhao Gu, Gael Guibon, Mika Hämäläinen, Kelvin Han, Shiyi Han, Yu Han, Sami Ul Haq, Bradley Hauer, Hui He, Junyi He, Yunjie He, Zhiwei He, Julien Heitmann, Alexander Henlein, Ondřej Herman, Xanh Ho, Julian Hoellig, Chun-Cheng Hsieh, Echo Hu, Langlin Huang, Shih-Cheng Huang, Shuyan Huang, Yerin Hwang, Radu Cristian Alexandru Iacob, Etsuko Ishii, Itay Itzhak, Adam Ivankay, Nazanin Jafari, Anubhav Jangra, Seongjun Jeong, Tianbo Ji, Qi Jia, Yiren Jian, Chengyue Jiang, Junfeng Jiang, Yiwei Jiang, Hailong Jin, Omisa Jinsi, Richard Jonker, Minjoon Jung, Danial Kamali, Jeongwoo Kang, Beatrice Kanyi, Abhinav Ramesh Kashyap, Prachuryya Kaushik, Joschka Kersting, Shamir Khandaker, Aditi Khandelwal, Niama El Khbir, Sopan Khosla, Mohammad Khosravani, Hajung Kim, Hyunjong Kim, Jeonghwan Kim, Jiwoo Kim, Seungone Kim, Yongil Kim, Youngbin Kim, Chaitanya Kirti, Xenia Klinge, Erik Körner, Ádám Kovács, Vojtěch Kovář, Shachi H Kumar, Vivek Kumar, Gitanjali Kumari, Maddalen López De Lacalle, Jack Lanchantin, Loic De Langhe, Anna Laskina, Chaeeun Lee, Dongryeol Lee, Kang-Il Lee, Sunkyung Lee, Yongjae Lee, Els Lefever, Zhihong Lei, Elisa Leonardelli, Hang Li, Jiazhao Li, Junlong Li, Mengyu Li, Ming-

han Li, Senyu Li, Shiyang Li, Shuqin Li, Wenyan Li, Xinhang Li, Yan Li, Yichen Li, Yichuan Li, Yunshui Li, Zekun Li, Zhaoqun Li, Zhuoqun Li, Zitong Li, Zhenwen Liang, Ruotong Liao, Boda Lin, Jiuheng Lin, Hali Lindsay, Alisa Liu, Andy T. Liu, Hong Liu, Hongyi Liu, Huijun Liu, Mengying Liu, Zhexiong Liu, Alessandro Locaputo, Roberto López, Sebastian Lopez-Cot, Yuze Lou, Xuantao Lu, Kamile Lukosiute, Gunnar Lund, Chu Fei Luo, Haoran Luo, Xin Lv, Congbo Ma, Da Ma, Andrew Mackey, Hiren Madhu, Daniele Malitesta, Oscar Mañas, Fabienne Marco, Salima Mdhaffar, Marek Medveď, Nikhil Mehta, Di Mei, Althis Mendes, Augusto Mendes, Stefano Menini, Elena Merdjanovska, Hossein Mohammadi, Samraj Moorjani, Yusuke Mori, Durgesh Nandini, Gaurav Negi, Hoang Nguyen, Vincent Nguyen, Feng Nie, Anna Nikiforovskaya, Jingcheng Niu, Gibson Nkhata, Rik Van Noord, Michael Ogezi, Olubusayo Olabisi, Katrina Olsen, Talgat Omarov, Andreas Opedal, Junshu Pan, Suehyun Park, Daraksha Parveen, Maya Pavlova, Diogo Pernes, Jan Pfister, Alejandro Piad-Morffis, Max Ploner, Alexander Podolskiy, Dejan Porjazovski, Pradyot Prakash, Adrien Pupier, Maarten De Raedt, Pétur Orri Ragnarsson, Sai Krishna Rallabandi, Leonardo Ranaldi, Abhinav Rao, Anton Razzhigaev, Sebastian Reimann, Raphael Reinauer, François Remy, Jiaqian Ren, Siyu Ren, Akseli Reunamo, Valentin Richard, Ruty Rinott, Elsa Rizk, Giulia Rizzi, Sean Robertson, Cristian Rodriguez, Sudipta Singha Roy, Susanna Rücker, Elena Sofia Ruzzetti, Tasnim Kabir Sadik, Joy Sain, Jose Ignacio Abreu Salas, Hossein Salemi, Mufan Sang, Twisampati Sarkar, Simone Scaboro, Felix Schmidt, Frederik Schmitt, Christopher Schröder, Simeon Schüz, Nina Seemann, Vincent Segonne, Yasas Senarath, Ashish Seth, Silvio Severino, Lele Sha, Stephen Shaffran, Anastassia Shaitarova, Hee Ming Shan, Kai Shen, Xingyu Shen, Shuqian Sheng, Kaize Shi, Ke Shi, Yuanjun Shi, Yuxuan Shu, Lucas Dos Santos Silva, Harmanpreet Singh, Pranaydeep Singh, Salam Michael Singh, Iustin Sirbu, Sonish Sivarajkumar, Mohamed Soliman, Chenyang Song, Kunzhe Song, William Soto, Florian Steuber, Manuel Stoeckel, Vit Suchomel, Bin Sun, Changzhi Sun, Cong Sun, Jingdong Sun, Qiujie Sun, Xiaohui Sun, Xueyao Sun, Shahbaz Syed, Zhaoxuan Tan, Shaowen Tang, Ziming Tang, Kumar Tanmay, Jingxuan Tu, Sichang Tu, Xiao Chi Tu, Mehmet Deniz Turkmen, Sagar Uprety, Hannah Vanderhoeven, Julien Velcin, Elad Venezian, Radhakrishnan Venkatakrishnan, Ivo Vigan, Fedor Vitiugin, Nikolas Vitsakis, Xiangpeng Wan, An Wang, Bingyu Wang, Cong Wang, Haoran Wang, Hu Wang, Junlin Wang, Junting Wang, Ke Wang, Lei Wang, Lingzhi Wang, Qianli Wang, Ruofan Wang, Shih-Heng Wang, Teng Wang, Weizhi Wang, Xinyou Wang, Yigong Wang, Yiming Wang, Yueguan Wang, Zihao Wang, Haitian Wei, Martyna Wiącek, Ronald Wilson, Moritz Wolf, Haibin Wu, Jay Zhangjie Wu, Jian Wu, Yexin Wu, Yuan-Kuei Wu, Siyuan Xiang, Yang Xiao, Yao Xiao, Zhouhang Xie, Benfeng Xu, Chenwei Xu, Kaishuai Xu, Yuzhuang Xu, Zhichao Xu, Zhiyang Xu, Bo Xue, Siyuan Xue, Xiaojun Xue, Baosong Yang, Kaiqi Yang, Shiping Yang, Yanjie Yang, Yinguan Yang, Jiarui Yao, Bingyang Ye, Yongjing Yin, Yuwei Yin, Tarik Yousef, Guoxin Yu, Nan Yu, Tiezheng Yu, Zhengqing Yuan, Klim Zaporojets, Urchade Zaratiana, Omnia Zayed, Weihao Zeng, Ge Zhang, Hanlei Zhang, Jingyu Zhang, Le Zhang, Mian Zhang, Qi Zhang, Ruike Zhang, Songyang Zhang, Tao Zhang, Weijia Zhang, Yidan Zhang, Yunan Zhang, Zhiling Zhang, Ziheng Zhang, Ziqiing Zhang, Ziqing Zhang, Honghong Zhao, Jiahao Zhao, Jinman Zhao, Siyang Zhao, Wei Zhao, Xingmeng Zhao, Yingxiu Zhao, Yu Zhao, Gui Zhen, Kangjie Zhen, Kai Zheng, Kaiwen Zhou, Terry Zhou, Zhengping Zhou, Zhijie Zhou, Ming Zhu, Zhihong Zhu, Haojie Zhuang, Anni Zou

2

Anti-Harassment Policy

ACL 2023 adheres to the ACL Anti-Harassment Policy. Any participant who experiences harassment or hostile behaviour may contact any current member of the ACL Professional Conduct Committee or Priscilla Rasmussen, who is usually available at the registration desk of the conference. Please be assured that if you approach us, your concerns will be kept in strict confidence, and we will consult with you on any actions taken.

The open exchange of ideas, the freedom of thought and expression, and respectful scientific debate are central to the aims and goals of a ACL conference. These require a community and an environment that recognizes the inherent worth of every person and group, that fosters dignity, understanding, and mutual respect, and that embraces diversity. For these reasons, ACL is dedicated to providing a harassment-free experience for participants at our events and in our programs.

Harassment and hostile behavior are unwelcome at any ACL conference. This includes: speech or behavior (including in public presentations and on-line discourse) that intimidates, creates discomfort, or interferes with a person's participation or opportunity for participation in the conference. We aim for ACL conferences to be an environment where harassment in any form does not happen, including but not limited to: harassment based on race, gender, religion, age, color, national origin, ancestry, disability, sexual orientation, or gender identity. Harassment includes degrading verbal comments, deliberate intimidation, stalking, harassing photography or recording, inappropriate physical contact, and unwelcome sexual attention.

The ACL board members are listed at:
`https://www.aclweb.org/portal/about`
The full policy and its implementation is defined at:
`https://2023.aclweb.org/participants/ethical_policies/`

3

Ethics Policy

ACL 2023 adheres to the ACM's code of ethics, the preamble of which is given below.
Computing professionals' actions change the world. To act responsibly, they should reflect upon the wider impacts of their work, consistently supporting the public good. The ACM Code of Ethics and Professional Conduct ("the Code") expresses the conscience of the profession.

The Code is designed to inspire and guide the ethical conduct of all computing professionals, including current and aspiring practitioners, instructors, students, influencers, and anyone who uses computing technology in an impactful way. Additionally, the Code serves as a basis for remediation when violations occur. The Code includes principles formulated as statements of responsibility, based on the understanding that the public good is always the primary consideration. Each principle is supplemented by guidelines, which provide explanations to assist computing professionals in understanding and applying the principle.
Section 1 outlines fundamental ethical principles that form the basis for the remainder of the Code. Section 2 addresses additional, more specific considerations of professional responsibility. Section 3 guides individuals who have a leadership role, whether in the workplace or in a volunteer professional capacity. Commitment to ethical conduct is required of every ACM member, ACM SIG member, ACM award recipient, and ACM SIG award recipient. Principles involving compliance with the Code are given in Section 4. The Code as a whole is concerned with how fundamental ethical principles apply to a computing professional's conduct. The Code is not an algorithm for solving ethical problems; rather it serves as a basis for ethical decision-making. When thinking through a particular issue, a computing professional may find that multiple principles should be taken into account, and that different principles will have different relevance to the issue. Questions related to these kinds of issues can best be answered by thoughtful consideration of the fundamental ethical principles, understanding that the public good is the paramount consideration. The entire computing profession benefits when the ethical decision-making process is accountable to and transparent to all stakeholders. Open discussions about ethical issues promote this accountability and transparency.

The full policy is available at `https://www.acm.org/code-of-ethics`.

4

Meal Info

*Breakfast: A light breakfast will be served for the Main Conference Days. The breakfast buffet will be served in the Regatta Ballroom located in the Hotel Lobby. Hours will be from 07:30 - 08:30. There will be seating in the Lobby Bar and additional seating in the Marina Ballroom. Please check the hotel maps for locations.

*Break: Coffee, tea , and light snacks will be provided late morning (approximately 10:30) and midafternoon (approximately 15:30)

Lunch: Lunch is not provided, but there are plenty of Hotels, cafes, restaurants and shops within walking distance. You can pick up a list of options at registration.

Dinner: Dinner is not provided, but there are plenty of Hotels, cafes, restaurants and shops within walking distance. You can pick up a list of options at registration.

Welcome Reception: Light Canapés will be provided on Sunday Evening July 09, 2023, at the Welcome Reception which will be held in the Metropolitan Ballroom Room. Welcome Reception tickets are included as part of the Full Conference Registration and can be added on for Guests, Tutorial, Workshop and Exhibitors to attend at the Registration Solutions Desk or through your Yes Events registration login. **No admission without an entry ticket.

** Social Event: Social Event: Dinner will only be provided on Tuesday Evening July 11, 2023, at the Social Event which will be held at the Steam Whistle Brewing Company. Social Event tickets are included as part of Entire Conference Registration and can be added on for Guests, Tutorial, Workshop and Exhibitors to attend at the Registration Solutions Desk or through your Yes Events registration login. **No admission without an entry ticket.**

Please note the following:
*Denotes Workshop/Tutorial/Main Conference Days
**Denotes Full Conference Days

5

Social Events

Welcome Reception—Sunday, July 9, 2023

Venue: **Westin Harbour Castle Hotel**
Time: **19:00–21:30**

One entry ticket will be included with each full conference registration. To get admission into the event you will need to have your name badge on your person as the QR code that is located on your badge is how the ACL Staff member(s) Scan and account for admission(s). No name badge, no entrance. Social Event tickets can be added on for Guests, Tutorial, Workshop and Exhibitors to attend at the Registration Solutions Desk or through your Yes Events registration login.
Location: Harbour Ballroom.

Directions: From the hotel lobby head up the escalator and follow the signs to the Metropolitan Ballroom.

Schedule:

- 19:00–21:30 - Pass around Canapés & Cash Bar: Each Full Conference attendee will receive one complimentary drink ticket upon admission into the Welcome Reception

Social Event—Tuesday, July 11, 2023

Venue: **Steam Whistle Brewing Company**
Time: **18:30–22:00**

We have planned the following social events for ACL 2023. Please follow ACL'S Code of Conduct while attending these events.

One entry ticket will be included with each full conference registration. To get admission into the event you will need to have your name badge on your person as the QR code that is located on your badge is how the ACL Staff member(s) Scan and account for admission(s). No name badge, no entrance. Social Event tickets can be added on for Guests, Tutorial, Workshop and Exhibitors to attend at the Registration Solutions Desk or through your Yes Events registration login.
Location: Steam Whistle Brewing Company located at The Roundhouse, 255 Bremner Blvd, Toronto ON M5V 3M9.

Directions: Walking distance from the main conference or Westin Harbour Castle Hotel is an 11-minute walk. There will be an accessibility walking map provided at check in.

Schedule:

- 18:30–20:30 - Buffet Dinner & Cash Bar: Each attendee will receive one complimentary drink ticket upon admission into the Social Event.

- 20:30–22:00 - Dessert & Entertainment: Local Canadian DJ will provide international hits for all walks of life participating in the Social Event.

- Last Call at 22:00

6

Keynotes, Panels and Discussions

Two Paths to Intelligence

Keynote: Geoffrey Hinton
Cohere

Monday, July 10 - Time: **09:30–10:30 EDT**

Abstract: I will briefly describe the forty year history of neural net language models with particular attention to whether they understand what they are saying. I will then discuss some of the main differences between digital and biological intelligences and speculate on how the brain could implement something like transformers. I will conclude by addressing the contentious issue of whether current multimodal LLMs have subjective experience.

Bio: Geoffrey Hinton received his PhD in Artificial Intelligence from Edinburgh in 1978. After five years as a faculty member at Carnegie-Mellon he became a fellow of the Canadian Institute for Advanced Research and moved to the University of Toronto where he is now an emeritus professor. He is also the Chief Scientific Adviser at the Vector Institute.

He was one of the researchers who introduced the backpropagation algorithm and the first to use back-propagation for learning word embeddings. His other contributions to neural network research include Boltzmann machines, distributed representations, time-delay neural nets, mixtures of experts, variational learning and deep learning. His research group in Toronto made major breakthroughs in deep learning that revolutionized speech recognition and object classification.

He is a fellow of the UK Royal Society and a foreign member of the US National Academy of Engineering, the US National Academy of Sciences and the American Academy of Arts and Sciences. His awards include the David E. Rumelhart prize, the IJCAI award for research excellence, the Killam prize for Engineering, the Royal Society Royal Medal, the NSERC Herzberg Gold Medal, the IEEE James Clerk Maxwell Gold medal, the NEC C&C award, the BBVA award, the Honda Prize and the Turing Award.

Large Language Models as Cultural Technologies: Imitation and Innovation in Children and Models

Alison Gopnik
University of California, Berkeley

Wednesday, July 12 - Time: **14:00–15:00 EDT**

Abstract: Its natural to ask whether large language models like LaMDA or GPT-3 are intelligent agents. But I argue that this is the wrong question. Intelligence and agency are the wrong categories for understanding them. Instead, these Al systems are what we might call cultural technologies, like writing, print, libraries, internet search engines or even language itself. They are new techniques for passing on information from one group of people to another. Cultural technologies arent like intelligent humans, but they are essential for human intelligence. Many animals can transmit some information from one individual or one generation to another, but no animal does it as much as we do or accumulates as much information over time, . New technologies that make cultural transmission easier and more effective have been among the greatest engines of human progress, but they have also led to negative as well as positive social consequences. Moreover, while cultural technologies allow transmission of existing information cultural evolution, which is central to human success, also depends on innovation, exploration and causal learning. Comparing LLM's responses in prompts based on developmental psychology experiments to the responses of children may provide insight into which capacities can be learned through language and cultural transmission, and which require innovation and exploration in the physical world. I will present results from several studies making such comparisons.

Bio: Alison Gopnik is a professor of psychology and affiliate professor of philosophy at the University of California at Berkeley, and a member of the Berkeley AI Research Group. She received her BA from McGill University and her PhD. from Oxford University. She is a leader in the study of cognitive science and of children's learning and development and was one of the founders of the field of "theory of mind", an originator of the "theory of cognitive development", and the first to apply Bayesian probabilistic models to children's learning. She has received both the APS Lifetime Achievement Cattell and William James Awards, the Bradford Washburn Award for Science Communication, and the SRCD Lifetime Achievement Award for Basic Science in Child Development. She is an elected member of the Society of Experimental Psychologists and the American Academy of Arts and Sciences and a Cognitive Science Society, American Association for the Advancement of Science, and Guggenheim Fellow. She was 2022-23 President of the Association for Psychological Science.
She is the author or coauthor of over 140 journal articles and several books including "Words, thoughts and theories" MIT Press, 1997, and the bestselling and critically acclaimed popular books "The Scientist

in the Crib" William Morrow, 1999, "The Philosophical Baby; What children's minds tell us about love, truth and the meaning of life" 2009, and "The Gardener and the Carpenter" 2016, Farrar, Strauss and Giroux, the latter two won the Cognitive Development Society Best Book Prize in 2009 and 2016. Since 2013 she has written the Mind and Matter column for the Wall Street Journal and she has also written widely about cognitive science and psychology for The New York Times, The Economist, The Atlantic, The New Yorker, Scientific American, The Times Literary Supplement, The New York Review of Books, New Scientist and Slate, among others. Her TED talk on her work has been viewed more than 5.2 million times. She has frequently appeared on TV, radio and podcasts including "The Charlie Rose Show", "The Colbert Report", "Radio Lab" and "The Ezra Klein Show". She lives in Berkeley with her husband Alvy Ray Smith and has three children and five grandchildren.

Panel: The Future of Computational Linguistics in the LLM Age

Chair: Iryna Gurevych
Technische Universität Darmstadt

Tuesday, July 11 - Time: **14:45-15:45**

This is a panel discussion with:

- Dan Klein (UC Berkeley)

- Meg Mitchell (Hugging Face)

- Roy Schwartz (the Hebrew University of Jerusalem)

They will present short statements (5 to 7 min.) related to the main topic of the panel

- New opportunities (e.g., artificial general intelligence, responsible NLP);

- Technical challenges (e.g., multimodality, instruction-tuning, etc.)

- Real life problems & societal implications (e.g., hallucinations, biases, future job market);

- LLMs and the future of NLP; and

- Open-science vs. commercial LLMs

Followed by discussion with the panel and audience.

Memorial

Tuesday, July 11, 2023 - Room: **Metropolitan** - Time: **13:00–13:30**

Dragomir Radev, the A. Bartlett Giamatti Professor of Computer Science at Yale University, passed away this year on Wed, March 29th. Drago contributed in substantial ways to research in NLP, to the organization of the ACL and to mentoring the next generation of computational linguists. Drago's role in our ACL community spans four decades. He was recognized for his work over this period through his selection as an ACL Fellow in 2018 for his significant contributions to text summarization and question answering, and through his receipt of the Distinguished ACL Service Award in 2022. In this session, speakers from different time periods of his life will discuss his contributions to the field and the impact his life had on so many of us.

ACL Rolling Review Update and Discussion

Mausam, Professor, IIT Delhi (ARR EIC), Jonathan K. Kummerfeld, Assistant Professor, University of Sydney (ARR CTO)
Tuesday, July 11, 2023 - Room: **Metropolitan** - Time: **14:15–14:45**

This session will contain a presentation on progress in ARR over the past year and provide an opportunity for community questions and discussion.
We will briefly present:

> Personnel Updates New aspects: Tracks, Senior Action Editors Improvements, e.g. changes to the review - paper matching process Statistics on timeliness and paper outcomes Next steps

With that context we will open the floor to questions.

Ethics Panel

Karën Fort, Min-Yen Kan and Yulia Tsvetkov (ACL Ethics Committee co-chairs) Committee Members: Luciana Benotti, Mark Dredze, Pascale Fung, Dirk Hovy, Jin-Dong Kim, Malvina Nissim
Tuesday, July 11, 2023 - Room: **Pier 4&5** - Time: **16:15–17:45**

We present our ACL Ethics Committee's progress over the last few years. Of core interest, we will present the results of the ACL stakeholder survey about the role of ethics and ethics training exposure. Results from the survey respondents indicate that ethics is of primary interest to the community and that there is a mandate for the further creation and dissemination of ethics related training for authors, reviewers and event organisers. We will briefly review the survey results and feature a lengthed question and answer session in support of extended dialogue with our community. Our session will culminate through a dialogue with our session's participants in a moderated panel that includes participation from the entire ethics committee.

7

Tutorials: Sunday, July 9, 2023

Overview

09:00 - 12:30	***Morning Tutorials***	
	Tutorial 1 – Goal Awareness for Conversational AI: Proactivity, Non-collaborativity, and Beyond	*Metropolitan East*
	Yang Deng, Wenqiang Lei, Minlie Huang and Tat-Seng Chua	
	Tutorial 2 – Complex Reasoning in Natural Language	*Metropolitan Centre*
	Wenting Zhao, Mor Geva, Bill Yuchen Lin, Michihiro Yasunaga, Aman Madaan and Tao Yu	
	Tutorial 3 – Everything you need to know about Multilingual LLMs: Towards fair, performant and reliable models for languages of the world	*Metropolitan West*
	Sunayana Sitaram, Monojit Choudhury, Barun Patra, Vishrav Chaudhary, Kabir Ahuja and Kalika Bali	
14:00 - 17:30	***Afternoon Tutorials***	
	Tutorial 4 – Generating Text from Language Models	*Metropolitan East*
	Afra Amini, Ryan Cotterell, John Hewitt, Clara Meister and Tiago Pimentel	
	Tutorial 5 – Indirectly Supervised Natural Language Processing	*Metropolitan Centre*
	Wenpeng Yin, Muhao Chen, Ben Zhou, Qiang Ning, Kai-Wei Chang and Dan Roth	
	Tutorial 6 – Retrieval-based Language Models and Applications	*Metropolitan West*
	Akari Asai, Sewon Min, Zexuan Zhong and Danqi Chen	

Message from the Tutorial Chairs

Welcome to the Tutorials Session of ACL 2023.

The ACL tutorials session is organized to give conference attendees a comprehensive introduction by expert researchers to some topics of importance drawn from our rapidly growing and changing research field. This year, as has been the tradition over the past few years, the call, submission, reviewing and selection of tutorials were coordinated jointly for multiple conferences: EACL, ACL, and EMNLP. We formed a review committee including the EACL tutorial chairs (Sameer Pradhan and Fabio Massimo Zanzotto) and ACL tutorial chairs (Yun-Nung Vivian Chen, Margot Mieskes, and Siva Reddy). A reviewing process was organized so that each proposal receives at least 2 reviews. The selection criteria included clarity, preparedness, novelty, timeliness, instructors' experience, likely audience, open access to the teaching materials, diversity (multilingualism, gender, age and geolocation) and the compatibility of preferred venues. A total of 42 tutorial submissions were received, of which 6 were selected for presentation at ACL.

We would like to thank the tutorial authors for their contributions and flexibility while organizing the conference in a hybrid format. Finally, our thanks go to the conference organizers for effective collaboration, and in particular to the general chair Yang Liu.

We hope you enjoy the tutorials.

ACL 2023 Tutorial Co-chairs

Yun-Nung (Vivian) Chen
Margot Mieskes
Siva Reddy

T1: Goal Awareness for Conversational AI: Proactivity, Non-collaborativity, and Beyond

Yang Deng, Wenqiang Lei, Minlie Huang and Tat-Seng Chua

Conversational systems are envisioned to provide social support or functional service to human users via natural language interactions. Conventional conversation researches mainly focus on the responseability of the system, such as dialogue context understanding and response generation, but overlooks the design of an essential property in intelligent conversations, i.e., goal awareness. The awareness of goals means the state of not only being responsive to the users but also aware of the target conversational goal and capable of leading the conversation towards the goal, which is a significant step towards higher-level intelligence and artificial consciousness. It can not only largely improve user engagement and service efficiency in the conversation, but also empower the system to handle more complicated conversation tasks that involve strategical and motivational interactions. In this tutorial, we will introduce the recent advances on the design of agent's awareness of goals in a wide range of conversational systems.

Yang Deng, Ph.D. Candidate, Chinese University of Hong Kong
email: `ydeng@se.cuhk.edu.hk`
website: `https://dengyang17.github.io`
Yang Deng is a final-year Ph.D. candidate in The Chinese University of Hong Kong. His research lies in natural language processing and information retrieval, especially for dialogue and QA systems. He has published over 20 papers at top venues such as ACL, EMNLP, SIGIR,WWW, TKDE, and TOIS.

Wenqiang Lei, Professor, Sichuan University
email: `wenqianglei@gmail.com`
website: `https://sites.google.com/view/wenqianghome/home`
Wenqiang Lei is a Professor in Sichuan University. His research interests focus on conversational AI, including conversational recommendation, dialogue and QA systems. He has published relevant papers at top venues such as ACL, EMNLP, KDD, SIGIR, TOIS, and received the ACM MM 2020 best paper award. He has given tutorials on the topic of conversational recommendation at RecSys 2021, SIGIR 2020, and co-organized special issues about conversational information seeking on ACM Trans. on Web. Specifically, his tutorial on SIGIR 2020 accepts over 1600 audiences, being one of the most popular tutorials in SIGIR 2020.

Minlie Huang, Associate Professor, Tsinghua University
email: `aihuang@tsinghua.edu.cn`
website: `http://coai.cs.tsinghua.edu.cn/hml`
Minlie Huang is an Associate Professor with the Department of Computer Science and Technology, Tsinghua University. He has authored or co-authored more than 100 papers in premier conferences and journals (ACL, EMNLP, TACL, etc). His research interests include natural language processing, particularly in dialog systems, reading comprehension, and sentiment analysis. He is an editor of TACL, CL, TNNLS, the Area Chair or SAC of ACL/EMNLP for more than 10 times. He is the recipient of IJCAI 2018 distinguished paper award, a nominee of ACL 2019 best demo papers, and SIGDIAL 2020 best paper award.

Tat-Seng Chua, KITHCT Chair Professor, National University of Singapore
email: `chuats@comp.nus.edu.sg`

website: https://www.chuatatseng.com

Tat-Seng Chua is the KITHCT Chair Professor with the School of Computing, National University of Singapore. His main research interest include multimedia information retrieval and social media analytics. He is the 2015 winner of the prestigious ACM SIGMM Technical Achievement Award and receives the best papers (or candidates) over 10 times in top conferences (SIGIR, WWW, MM, etc). He serves as the general co-chair of top conferences multiple times (MM 2005, SIGIR 2008, WSDM 2023, etc), and the editors of multiple journals (TOIS, TMM, etc). He has given invited keynote talks at multiple top conferences, including the recent one on the topic of multimodal conversational search and recommendation.

T2: Complex Reasoning in Natural Language

Wenting Zhao, Mor Geva, Bill Yuchen Lin, Michihiro Yasunaga, Aman Madaan and Tao Yu

```
https:
//wenting-zhao.github.io/complex-reasoning-tutorial
```

Teaching machines to reason over texts has been a long-standing goal of natural language processing (NLP). To this end, researchers have designed a diverse set of complex reasoning tasks that involve compositional reasoning, knowledge retrieval, grounding, commonsense reasoning, etc.

A standard choice for building systems that perform a desired type of reasoning is to fine-tune a pretrained language model (LM) on specific downstream tasks. However, recent research has demonstrated that such a straightforward approach is often brittle. For example, Elazar et al. (2021) and Branco et al. (2021) show that, on question-answering (QA) tasks, similar performance can be achieved with questions removed from the inputs. Min et al. (2019), Chen and Durrett (2019), and Tang et al. (2021) show that models trained on multi-hop QA do not generalize to answer single-hop questions. The reasoning capabilities of these models thus remain at a surface level, i.e., exploiting data patterns. Consequently, augmenting LMs with techniques that make them robust and effective becomes an active research area.

We will start the tutorial by providing an overview of complex reasoning tasks where the standard application of pretrained language models fails. This tutorial then reviews recent promising directions for tackling these tasks. Specifically, we focus on the following groups of approaches that explicitly consider problem structures: (1) knowledge-augmented methods, where the knowledge is either incorporated during fine-tuning or pretraining; (2) few-shot prompting methods, which effectively guide the models to follow instructions; (3) neuro-symbolic methods, which produce explicit intermediate representations; and, (4) rationale-based methods, one of the most popular forms of the neuro-symbolic methods, which highlight subsets of input as explanations for individual model predictions.

Wenting Zhao, Ph.D. student, Cornell University
email: wzhao@cs.cornell.edu
Wenting Zhao is a Ph.D. student in Computer Science at Cornell University. Her research focuses on the intersection of reasoning and NLP. She is especially interested in developing explainable methods for complex reasoning problems. He has published over 20 papers at top venues such as ACL, EMNLP, SIGIR,WWW, TKDE, and TOIS.

Mor Geva, Postdoctoral Researcher, Google Research
email: pipek@google.com
Mor Geva is a postdoctoral researcher, now at Google Research and previously at the Allen Institute for AI. Her research focuses on debugging the inner workings of black-box NLP models, to increase their transparency, control their operation, and improve their reasoning abilities. She is organizing the next edition of the Workshop on Commonsense Reasoning and Representation.

Bill Yuchen Lin, Postdoctoral Researcher, Allen Institute for AI
email: yuchenl@allenai.org
Bill Yuchen Lin is a postdoctoral researcher at the Allen Institute for AI. He obtained his Ph.D. at USC advised by Prof. Xiang Ren. His research goal is to teach machines to think, talk, and act with commonsense knowledge and commonsense reasoning ability as humans do. He was a co-author of the tutorial on

Knowledge-Augmented Methods for Natural Language Processing and the Workshop on Commonsense Representation and Reasoning at ACL 2022.

Michihiro Yasunaga, Ph.D. Student, Stanford University
email: `myasu@cs.stanford.edu`
Michihiro Yasunaga is a Ph.D. student in Computer Science at Stanford University. His research interest is in developing generalizable models with knowledge, including commonsense, science, and reasoning abilities. He co-organized the Workshop on Structured and Unstructured Knowledge Integration (SUKI) at NAACL 2022.

Aman Madaan, Ph.D. Student, Carnegie Mellon University
email: `amadaan@cs.cmu.edu`
Aman Madaan is a Ph.D. student at the School of Computer Science, Carnegie Mellon University. He is interested in large language models, feedback-driven generation, and the intersection of code generation and natural language reasoning. He helped organize the 1st and 2nd Workshops on Natural Language Generation, Evaluation, and Metrics (GEM) at ACL 2021 and EMNLP 2022.

Tao Yu, Assistant Professor, University of Hong Kong
email: `tyu@cs.hku.hk`
Tao Yu is an assistant professor of computer science at The University of Hong Kong. He completed his Ph.D. at Yale University and was a postdoctoral fellow at the University of Washington. He works on executable language understanding, such as semantic parsing and code generation, and large LMs. Tao is the recipient of an Amazon Research Award. He co-organized multiple workshops in Semantic Parsing and Structured and Unstructured Knowledge Integration at EMNLP and NAACL.

T3: Everything you need to know about Multilingual LLMs: Towards fair, performant and reliable models for languages of the world

Sunayana Sitaram, Monojit Choudhury, Barun Patra, Vishrav Chaudhary, Kabir Ahuja and Kalika Bali

This tutorial will describe various aspects of scaling up language technologies to many of the world's languages by describing the latest research in Massively Multilingual Language Models (MMLMs). We will cover topics such as data collection, training and fine-tuning of models, Responsible AI issues such as fairness, bias and toxicity, linguistic diversity and evaluation in the context of MMLMs, specifically focusing on issues in non-English and low-resource languages. Further, we will also talk about some of the real-world challenges in deploying these models in language communities in the field. With the performance of MMLMs improving in the zero-shot setting for many languages, it is now becoming feasible to use them for building language technologies in many languages of the world, and this tutorial will provide the computational linguistics community with unique insights from the latest research in multilingual models

Sunayana Sitaram, Senior Researcher, Microsoft Research India
email: `sunayana,sitaram@microsoft.com`
Sunayana Sitaram is a Senior Researcher at Microsoft Research India, where she works on multilingual speech and NLP. Her current research interests include training and evaluation of Massively Multilingual Language Models and Responsible AI for NLP. Prior to coming to MSRI as a Post Doc, Sunayana completed her MS and PhD at the Language Technologies Institute, Carnegie Mellon University in 2015. Sunayana's research has been published in top NLP and Speech conferences including ACL, NAACL, EMNLP, Interspeech, ICASSP. She has organized special sessions and workshops on under-resourced languages, code-switching, multilingual evaluation and speech for social good. She has also led the creation of several benchmarks and datasets in code-switching, ASR, NLI and TTS that have been used by research groups all over the world.

Monojit Choudhury, Principal Applied Scientist, Microsoft Turing
email: `monojitc@microsoft.com`
Monojit Choudhury is a Principal Applied Scientist at Microsoft Turing, prior to which he was a Principal Researcher at Microsoft Research India. He is also a Professor of Practice at Plaksha University, and had held adjunct faculty positions at Ashoka University, IIIT Hyderabad and IIT Kharagpur. Over the past 15 years, Monojit has worked on several impactful projects on processing of code-mixed text, evaluation and linguistic fairness of large language models, and social impact through participatory design of technology for under-resourced languages like Gondi, Mundari, Idu Mishmi and Swahili. Monojit has served as Senior Area Chair and Area chair in leading NLP and AI conferences including EMNLP, ACL, NAACL, IJCNLP and AAAI. He has organized several successful workshops in *ACL conferences (SUMEval 2022, CALCS series, TextGraph series, etc.) and has delivered a tutorial on Code-mixed text processing at EMNLP 2019. He is the general chair of the Panini Linguistics Olympiad and the founding co-chair of Asia Pacific Linguistics Olympiad — programs to introduce bright young students to linguistics and computational linguistics through puzzles. Dr. Choudhury holds PhD and B.Tech degrees in Computer Science and Engineering from IIT Kharagpur.

Vishrav Chaudhary, Principal Researcher, Microsoft Turing
email: `vchaudhary@microsoft.com`
Vishrav Chaudhary is a Principal Researcher at Microsoft Turing where he works on scaling and building efficient Multilingual and Multimodal representation and generation models. Prior to Microsoft, Vishrav was a Lead Researcher at FAIR and focused on several aspects of Machine Translation, Quality Estimation and Cross-lingual understanding. Over the past 10 years, Vishrav's research work has been published in several leading NLP and AI conferences and journals including ACL, EMNLP, NAACL, EACL, AACL, TACL, JMLR and AMTA. He has also organized several workshops successfully including SUMEval 2022, AmericasNLP 2021, WMT 2021 etc. He has also served as an Area Chair for EMNLP 2022. Vishrav has also led creation of benchmarks and datasets targeting 100+ languages which have been used to train state-of-the-art Cross-Lingual Representation and Machine Translation models.

Barun Patra, Applied Scientist, Microsoft Turing
email: `bapatra@microsoft.com`
Barun Patra is an Applied Scientist at Microsoft Turing. His research interest revolves around building better foundational models that can help support numerous NLP tasks across different languages. Barun's research work focuses on improving the quality and efficiency of training these large multilingual foundational models, helping achieve state-of-the-art performance on crosslingual NLP tasks.

Kabir Ahuja, Research Fellow, Microsoft Research India
email: `t-kabirahuja@microsoft.com`
Kabir Ahuja is a Research Fellow at Microsoft Research India, where he works on building linguistically fair multilingual models covering different aspects around their performance, calibration, evaluation, interpretation, and data collection. He is also interested in the analysis and interpretability of the computation mechanisms utilized by neural sequence models for solving different tasks.

Kalika Balia, Principal Researcher, Microsoft Research India
email: `kalikab@microsoft.com`
Kalika Bali is a Principal Researcher at Microsoft Research India working in the areas of Machine Learning, Natural Language Systems and Applications, as well as Technology for Emerging Markets. Her research interests lie broadly in the area of Speech and Language Technology especially in the use of linguistic models for building technology that offers a more natural Human- Computer as well as Computer-Mediated interactions.

T4: Generating Text from Language Models

Afra Amini, Ryan Cotterell, John Hewitt, Clara Meister and Tiago Pimentel

An increasingly large percentage of natural language processing (NLP) tasks center around the generation of text from probabilistic language models. Despite this trend, techniques for improving or specifying preferences in these generated texts rely mostly on intuition-based heuristics. Further, there lacks a unified presentation of their motivations, practical implementation, successes and pitfalls. Practitioners must, therefore, choose somewhat blindly between generation algorithms—like top-p sampling or beam search—which can lead to wildly different results. At the same time, language generation research continues to criticize and improve the standard toolboxes, further adding entropy to the state of the field. In this tutorial, we will provide a centralized and cohesive discussion of critical considerations when choosing how to generate from a language model. We will cover a wide range of empirically-observed problems (like degradation, hallucination, repetition) and their corresponding proposed algorithmic solutions from recent research (like top-p sampling and its successors). We will then discuss a subset of these algorithms under a unified light; most stochastic generation strategies can be framed as locally adapting the probabilities of a model to avoid failure cases. Finally, we will then cover methods in controlled generation, that go beyond just ensuring coherence to ensure text exhibits specific desired properties. We aim for NLP practitioners and researchers to leave our tutorial with a unified framework which they can use to evaluate and contribute to the latest research in language generation.

Afra Amini, Ph.D. Student, ETH Zűrich
email: `afra.amini@inf.ethz.ch`
Afra Amini is a PhD student at ETH Zűrich in the ETH AI Center. Her current foci include language generation and parsing.

Ryan Cotterell, Assistant Professor, ETH Zűrich
email: `ryan.cotterell@inf.ethz.ch`
Ryan Cotterell is an assistant professor at ETH Zűrich in the Institute for Machine Learning.

John Hewitt, Ph.D. Student, Stanford University
email: `johnhew@cs.stanford.edu`
John Hewitt Is a PhD student at Stanford University. His research tackles basic problems in learning models from broad distributions over language, characterizing and understanding those models, and building smaller, simpler models.

Clara Meister, Ph.D. Student, ETH Zűrich
email: `clara.meister@inf.ethz.ch`
Clara Meister is a PhD student at ETH Ziűrich in the Institute for Machine Learning and a Google PhD Fellow. Her current foci include language generation, pyscholinguistics, and the general application of statistical methods to natural language processing.

Tiago Pimentel, Ph.D. Student, University of Cambridge
email: `tp472@cam.ac.uk`
Tiago Pimentel is a PhD student at the University of Cambridge and a Facebook Fellow. His research focuses on information theory, and its applications to the analysis of pre-trained language models and natural languages.

T5: Indirectly Supervised Natural Language Processing

Wenpeng Yin, Muhao Chen, Ben Zhou, Qiang Ning, Kai-Wei Chang and Dan Roth

This tutorial targets researchers and practitioners who are interested in ML technologies for NLP from indirect supervision. In particular, we will present a diverse thread of indirect supervision studies that try to answer the following questions: (i) when and how can we provide supervision for a target task T, if all we have is data that corresponds to a related task T? (ii) humans do not use exhaustive supervision; they rely on occasional feedback, and learn from incidental signals from various sources; how can we effectively incorporate such supervision in machine learning? (iii) how can we leverage multi-modal supervision to help NLP? To the end, we will discuss several lines of research that address those challenges, including (i) indirect supervision from T' that handles T with outputs spanning from a moderate size to an open space, (ii) the use of sparsely occurring and incidental signals, such as partial labels, noisy labels, knowledge-based constraints, and cross-domain or cross-task annotations—all having statistical associations with the task, (iii) principled ways to measure and understand why these incidental signals can contribute to our target tasks, and (iv) indirect supervision from vision-language signals. We will conclude the tutorial by outlining directions for further investigation.

Wenpeng Yin, Assistant Professor, Penn State University
email: `wenpeng@psu.edu`
website: `http://www.wenpengyin.org`
Wenpeng Yin is an Assistant Professor in the Department of Computer Science and Engineering at Penn State University. Prior to joining Penn State, he was a tenure-track faculty member at Temple University (1/2022-12/2022), Senior Research Scientist at Salesforce Research (8/2019-12/2021), a postdoctoral researcher at UPenn (10/2017-7/2019), and got his Ph.D. degree from the Ludwig Maximilian University of Munich, Germany, in 2017. Dr. Yin's research focuses on natural language processing with three sub-areas: (i) learning from task instructions; (ii) information extraction; (iii) learning with limited supervision.

Muhao Chen, Assistant Research Professor, USC
email: `muhaoche@usc.edu`
website: `http://luka-group.github.io/`
Muhao Chen is an Assistant Research Professor of Computer Science at USC, where he directs the Language Understanding and Knowledge Acquisition (LUKA) Group. His research focuses on data-driven machine learning approaches for natural language understanding and knowledge acquisition. His work has been recognized with an NSF CRII Award, a Cisco Faculty Research Award, an ACM SIGBio Best Student Paper Award, and a Best Paper Nomination at CoNLL. Muhao obtained his PhD degree from UCLA Department of Computer Science in 2019, and was a postdoctoral researcher at UPenn prior to joining USC.

Ben Zhou, Ph.D. Student, University of Pennsylvania
email: `xyzhou@seas.upenn.edu`
website: `http://xuanyu.me/`
Ben Zhou is a fourth-year Ph.D. student at the Department of Computer and Information Science, University of Pennsylvania. Ben's research interests are distant supervision extraction and experiential knowledge reasoning, and he has more than 5 recent papers on related topics. He is a recipient of the ENIAC fellowship from the University of Pennsylvania, and a finalist of the CRA outstanding undergraduate researcher

award.

Qiang Ning, Senior Applied Scientist, Amazon AWS AI
email: qning@amazon.com
website: https://www.qiangning.info/
Qiang Ning is currently a senior applied scientist at AWS AI (2022-). Prior to that, Qiang was an applied scientist at Alexa AI (2020-2022) and a research scientist at the Allen Institute for AI (2019-2020). Qiang received his Ph.D. from the University of Illinois at Urbana-Champaign in 2019 in Electrical and Computer Engineering. Qiang's research interests span in information extraction, question answering, and the application of weak supervision methods in these NLP problems in both theoretical and practical aspects.

Kai-Wei Chang, Associate Professor, University of California Los Angeles
email: kwchang@cs.ucla.edu
website: http://kwchang.net/
Kai-Wei Chang is an associate professor in the Department of Computer Science at the University of California Los Angeles. His research interests include designing robust, fair, and accountable machine learning methods for building reliable NLP systems. His awards include the EMNLP Best Long Paper Award (2017), the KDD Best Paper Award (2010), and the Sloan Resaerch Fellowship (2021). Kai-Wei has given tutorials at NAACL 15, AAAI 16, FAccT18, EMNLP 19, AAAI 20, EMNLP 21, MLSS 21 on different research topics.

Dan Roth, Eduardo D. Glandt Distinguished Professor, UPenn
email: danroth@seas.upenn.edu
website: http://www.cis.upenn.edu/ danroth
Dan Roth is the Eduardo D. Glandt Distinguished Professor at the Department of Computer and Information Science, UPenn, the NLP Lead at AWS AI Labs, and a Fellow of the AAAS, ACM, AAAI, and ACL. In 2017 Roth was awarded the John McCarthy Award, the highest award the AI community gives to mid-career AI researchers. Roth was recognized "for major conceptual and theoretical advances in the modeling of natural language understanding, machine learning, and reasoning." Roth has published broadly in machine learning, NLP, KRR, and learning theory, and has given keynote talks and tutorials in all ACL and AAAI major conferences. Roth was the Editor-in-Chief of JAIR until 2017, and was the program chair of AAAI'11, ACL'03 and CoNLL'02; he serves regularly as an area chair and senior program committee member in the major conferences in his research areas.

T6: Retrieval-based Language Models and Applications

Akari Asai, Sewon Min, Zexuan Zhong and Danqi Chen

Retrieval-based language models (LMs) have shown impressive performance on diverse NLP tasks. In this tutorial, we will provide a comprehensive and coherent overview of recent advances in retrieval-based LMs. We will start by providing preliminaries covering the foundation of LMs (e.g., masked LMs, autoregressive LMs) and retrieval systems (e.g., nearest-neighbor search). We will then detail recent progress in retrieval-based models, focusing on their model architectures and learning approaches. Finally, we will show how retrieval-based LMs are adapted to downstream applications, and extended to multilingual and multi-modal settings. Finally, we will use an exercise to showcase the effectiveness of retrieval-based LMs.

Akari Asai, Ph.D. Student, University of Washington
email: `akari@cs.washington.edu`
Akari Asai is a Ph.D. student in the Paul G. Allen School of Computer Science & Engineering at the University of Washington, advised by Prof. Hannaneh Hajishirzi. Her research lies in natural language processing and machine learning. Her recent research focuses on question answering, retrieval-based LMs, multilingual NLP, and entity-aware representations. She received the IBM Fellowship in 2022. She is a lead organizer of the Workshop on Multilingual Information Access (NAACL 2022) and serves as an area chair in question answering at EACL 2023.

Sewon Min, Ph.D. Student, University of Washington
email: `sewon@cs.washington.edu`
Sewon Min is a Ph.D. student in the Paul G. Allen School of Computer Science & Engineering at the University of Washington, and a visiting researcher at Meta AI. Her research spans question answering, representation and retrieval of factoid knowledge, and language modeling. She was a co-instructor and a co-organizer of multiple tutorials and workshops at ACL, NAACL-HLT, EMNLP, NeurIPS and AKBC, including a tutorial on Few-Shot NLP with Pretrained Language Models (ACL 2022), a tutorial on NLP for Long Sequences (NAACL-HLT 2021), and the Workshop on Semiparametric Methods in NLP (ACL 2022).

Zexuan Zhong, Ph.D. Student, University of Washington
email: `zzhong@cs.princeton.edu`
Zexuan Zhong is a Ph.D. student in the Department of Computer Science at Princeton University, advised by Prof. Danqi Chen. His research interests lie in natural language processing and machine learning. His recent research focuses on retrieval-based LMs, generalization of retrieval models, and efficient models in NLP. He received a J.P. Morgan PhD Fellowship in 2022.

Danqi Chen, Assistant Professor, Princeton University
email: `danqic@cs.princeton.edu`
Danqi Chen is an Assistant Professor of Computer Science at Princeton University and co-leads the Princeton NLP Group. Her recent research focuses on training, adapting, and understanding large LMs, and developing scalable and generalizable NLP systems for question answering, information extraction, and conversational agents. Danqi is a recipient of a Sloan Fellowship, a Samsung AI Researcher of the Year award, outstanding paper awards from ACL 2016, EMNLP 2017 and ACL 2022, and multiple industry faculty awards. Danqi served as the program chair for AKBC 2021 and (senior) area chairs for many *ACL conferences. She taught a tutorial on "Open-domain Question Answering" at ACL 2020.

8

Main Conference

Main Conference Program (Overview)

Main Conference Program (Overview): Day 1

7:30-8:45	Breakfast
9:00-9:30	Welcome Ceremony
9:30-10:30	**Keynote: Geoffrey Hinton**
10:30-11:00	Coffee Break

11:00-12:30	**Session 1:**	**NLP Applications** *Metropolitan East*	**Large Language Models** *Metropolitan Centre*
		Question Answering *Metropolitan West*	**In Person Poster Session** *Frontenac Ballroom*
		Ethics and NLP *Pier 2&3*	**Multilingualism and Cross-Lingual NLP** *Pier 4&5*
		Virtual Poster	

| 12:30-14:00 | Lunch Break |

14:00-15:30	**Session 2:**	**Theme: Reality Check** *Metropolitan East*	**Machine Learning for NLP** *Metropolitan Centre*
		Machine Translation *Metropolitan West*	**In Person Poster Session** *Frontenac Ballroom*
		Sentiment Analysis, Stylistic Analysis, and Argument Mining *Pier 2&3*	**Language Grounding to Vision, Robotics, and Beyond** *Pier 4&5*
		Syntax: Tagging, Chunking, and Parsing *Pier 7&8*	

| 15:30-16:00 | Coffee Break |
| 16:00-17:30 | **Best Paper Awards** |

19:00-21:00	**Spotlight Session:**	**Findings Spotlights I** *Metropolitan East*	**Findings Spotlights II** *Metropolitan Centre*
		Findings Spotlights III *Metropolitan West*	

Main Conference Program (Overview): Day 2

7:30-8:45	Breakfast			

9:00-10:30	**Session 3:**	**Interpretability and Analysis of Models for NLP** *Metropolitan East*	**Large Language Models** *Metropolitan Centre*	
		Dialogue and Interactive Systems *Metropolitan West*	**In Person Poster Session** *Frontenac Ballroom*	
		Computational Social Science and Cultural Analytics *Pier 2&3*	**Industry track: Model efficiency, Information Extraction** *Pier 4&5*	
		Linguistic Diversity *Pier 7&8*		

10:30-11:00	Coffee Break

11:00-12:30	**Session 4:**	**Resources and Evaluation** *Metropolitan East*	**Large Language Models** *Metropolitan Centre*
		Summarization *Metropolitan West*	**In Person Poster Session** *Frontenac Ballroom*
		Student Research Workshop *Pier 2&3*	**Language Grounding to Vision, Robotics, and Beyond** *Pier 4&5*
		Virtual Poster *Pier 7&8*	

12:30-13:00	Lunch Break
13:00-13:30	**Dr. Dragomir Radev Memorial**
13:30-14:10	Business Meeting
14:15-14:45	ARR Discussion
14:45-15:45	**Panel: Large Language Models**
15:45-16:15	Coffee Break

16:15-17:45	**Session 5:**	**Interpretability and Analysis of Models for NLP** *Metropolitan East*	**Information Extraction** *Metropolitan Centre*
		Generation *Metropolitan West*	**In Person Poster Session** *Frontenac Ballroom*
		Semantics: Lexical *Pier 2&3*	**Linguistic Theories, Cognitive Modeling, and Psycholinguistics** *Pier 7&8*

18:30-22:00	Social Event (Steam Whistle Brewing / Canada's Premium Beer)

Main Conference Program (Overview): Day 3

7:30-8:45	Breakfast			

9:00-10:30	**Session 6:**	**NLP Applications** *Metropolitan East*	**Machine Learning for NLP** *Metropolitan Centre*
		Machine Translation *Metropolitan West*	**In Person Poster Session** *Frontenac Ballroom*
		Semantics: Sentence-level Semantics, Textual Inference, and Other Areas *Pier 2&3*	**Industry track: Interactive Systems, Speech** *Pier 4&5*
		Phonology, Morphology, and Word Segmentation *Pier 7&8*	

10:30-11:00 Coffee Break

11:00-12:30	**Session 7:**	**Resources and Evaluation** *Metropolitan East*	**Information Extraction / Generation** *Metropolitan Centre*
		Information Retrieval and Text Mining *Metropolitan West*	**In Person Poster Session** *Frontenac Ballroom*
		Discourse and Pragmatics *Pier 2&3*	**Speech and Multimodality** *Pier 4&5*
		Virtual Poster *Pier 2&3*	

12:30-14:00 Lunch Break

14:00-15:00 **Kehynote: Alison Gopnik**

15:00-15:30 Coffee Break

15:30-17:00 **ACL Lifetime / ToT Awards**

17:00-17:30 Closing Session

Main Conference: Monday, July 10, 2023

Session 1 - 11:00-12:30

NLP Applications

11:00-12:30 (Metropolitan East)

CoAD: Automatic Diagnosis through Symptom and Disease Collaborative Generation
Huimin Wang, Wai Chung Kwan, Kam-Fai Wong and Yefeng Zheng   11:00-11:15 (Metropolitan East)
Automatic diagnosis (AD), a critical application of AI in healthcare, employs machine learning techniques to assist doctors in gathering patient symptom information for precise disease diagnosis. The Transformer-based method utilizes an input symptom sequence, predicts itself through auto-regression, and employs the hidden state of the final symptom to determine the disease. Despite its simplicity and superior performance demonstrated, a decline in disease diagnosis accuracy is observed caused by 1) a mismatch between symptoms observed during training and generation, and 2) the effect of different symptom orders on disease prediction. To address the above obstacles, we introduce the CoAD, a novel disease and symptom collaborative generation framework, which incorporates several key innovations to improve AD: 1) aligning sentence-level disease labels with multiple possible symptom inquiry steps to bridge the gap between training and generation; 2) expanding symptom labels for each sub-sequence of symptoms to enhance annotation and eliminate the effect of symptom order; 3) developing a repeated symptom input schema to effectively and efficiently learn the expanded disease and symptom labels. We evaluate the CoAD framework using four datasets, including three public and one private, and demonstrate that it achieves an average 2.3% improvement over previous state-of-the-art results in automatic disease diagnosis. For reproducibility, we release the code and data at https://github.com/KwanWaiChung/coad.

Clinical Note Owns its Hierarchy: Multi-Level Hypergraph Neural Networks for Patient-Level Representation Learning
Nayeon Kim, Yinhua Piao and Sun Kim   11:15-11:30 (Metropolitan East)
Leveraging knowledge from electronic health records (EHRs) to predict a patient's condition is essential to the effective delivery of appropriate care. Clinical notes of patient EHRs contain valuable information from healthcare professionals, but have been underused due to their difficult contents and complex hierarchies. Recently, hypergraph-based methods have been proposed for document classifications. Directly adopting existing hypergraph methods on clinical notes cannot sufficiently utilize the hierarchy information of the patient, which can degrade clinical semantic information by (1) frequent neutral words and (2) hierarchies with imbalanced distribution. Thus, we propose a taxonomy-aware multi-level hypergraph neural network (TM-HGNN), where multi-level hypergraphs assemble useful neutral words with rare keywords via note and taxonomy level hyperedges to retain the clinical semantic information. The constructed patient hypergraphs are fed into hierarchical message passing layers for learning more balanced multi-level knowledge at the note and taxonomy levels. We validate the effectiveness of TM-HGNN by conducting extensive experiments with MIMIC-III dataset on benchmark in-hospital-mortality prediction.

DICE: Data-Efficient Clinical Event Extraction with Generative Models
Mingyu Derek Ma, Alexander K Taylor, Wei Wang and Nanyun Peng   11:30-11:45 (Metropolitan East)
Event extraction for the clinical domain is an under-explored research area. The lack of training data along with the high volume of domain-specific terminologies with vague entity boundaries makes the task especially challenging. In this paper, we introduce DICE, a robust and data-efficient generative model for clinical event extraction. DICE frames event extraction as a conditional generation problem and introduces a contrastive learning objective to accurately decide the boundaries of biomedical mentions. DICE also trains an auxiliary mention identification task jointly with event extraction tasks to better identify entity mention boundaries, and further introduces special markers to incorporate identified entity mentions as trigger and argument candidates for their respective tasks. To benchmark clinical event extraction, we compose MACCROBAT-EE, the first clinical event extraction dataset with argument annotation, based on an existing clinical information extraction dataset MACCROBAT. Our experiments demonstrate state-of-the-art performances of DICE for clinical and news domain event extraction, especially under low data settings.

TemplateGEC: Improving Grammatical Error Correction with Detection Template
Yinghao Li, Xuebo Liu, Shuo Wang, Peiyuan Gong, Derek F. Wong, Yang Gao, Heyan Huang and Min Zhang 11:45-12:00 (Metropolitan East)
Grammatical error correction (GEC) can be divided into sequence-to-edit (Seq2Edit) and sequence-to-sequence (Seq2Seq) frameworks, both of which have their pros and cons. To utilize the strengths and make up for the shortcomings of these frameworks, this paper proposes a novel method, TemplateGEC, which capitalizes on the capabilities of both Seq2Edit and Seq2Seq frameworks in error detection and correction respectively. TemplateGEC utilizes the detection labels from a Seq2Edit model, to construct the template as the input. A Seq2Seq model is employed to enforce consistency between the predictions of different templates by utilizing consistency learning. Experimental results on the Chinese NLPCC18, English BEA19 and CoNLL14 benchmarks show the effectiveness and robustness of TemplateGEC. Further analysis reveals the potential of our method in performing human-in-the-loop GEC. Source code and scripts are available at https://github.com/li-aolong/TemplateGEC.

Towards Domain-Agnostic and Domain-Adaptive Dementia Detection from Spoken Language
Shahla Farzana and Natalie Parde   12:00-12:15 (Metropolitan East)
Health-related speech datasets are often small and varied in focus. This makes it difficult to leverage them to effectively support healthcare goals. Robust transfer of linguistic features across different datasets orbiting the same goal carries potential to address this concern. To test this hypothesis, we experiment with domain adaptation (DA) techniques on heterogeneous spoken language data to evaluate generalizability across diverse datasets for a common task: dementia detection. We find that adapted models exhibit better performance across conversational and task-oriented datasets. The feature-augmented DA method achieves a 22% increase in accuracy adapting from a conversational to task-specific dataset compared to a jointly trained baseline. This suggests promising capacity of these techniques to allow for productive use of disparate data for a complex spoken language healthcare task.

Using Neural Machine Translation for Generating Diverse Challenging Exercises for Language Learner
Frank Palma Gomez, Subhadarshi Panda, Michael M Flor and Alla Rozovskaya   12:15-12:30 (Metropolitan East)
We propose a novel approach to automatically generate distractors for cloze exercises for English language learners, using round-trip neural machine translation. A carrier sentence is translated from English into another (pivot) language and back, and distractors are produced by

aligning the original sentence with its round-trip translation. We make use of 16 linguistically-diverse pivots and generate hundreds of translation hypotheses in each direction. We show that using hundreds of translations allows us to generate a rich set of challenging distractors. Moreover, we find that typologically unrelated language pivots contribute more diverse candidate distractors, compared to language pivots that are closely related. We further evaluate the use of machine translation systems of varying quality and find that better quality MT systems produce more challenging distractors. Finally, we conduct a study with language learners, demonstrating that the automatically generated distractors are of the same difficulty as the gold distractors produced by human experts.

Large Language Models

11:00-12:30 (Metropolitan Centre)

Searching for Needles in a Haystack: On the Role of Incidental Bilingualism in PaLM's Translation Capability

Eleftheria Briakou, Colin Cherry and George Foster 11:00-11:15 (Metropolitan Centre)

Large, multilingual language models exhibit surprisingly good zero- or few-shot machine translation capabilities, despite having never seen the intentionally-included translation examples provided to typical neural translation systems. We investigate the role of incidental bilingualism—the unintentional consumption of bilingual signals, including translation examples—in explaining the translation capabilities of large language models, taking the Pathways Language Model (PaLM) as a case study. We introduce a mixed-method approach to measure and understand incidental bilingualism at scale. We show that PaLM is exposed to over 30 million translation pairs across at least 44 languages. Furthermore, the amount of incidental bilingual content is highly correlated with the amount of monolingual in-language content for non-English languages. We relate incidental bilingual content to zero-shot prompts and show that it can be used to mine new prompts to improve PaLM's out-of-English zero-shot translation quality. Finally, in a series of small-scale ablations, we show that its presence has a substantial impact on translation capabilities, although this impact diminishes with model scale.

KILM: Knowledge Injection into Encoder-Decoder Language Models

Yan Xu, Mahdi Namazifar, Devamanyu Hazarika, Aishwarya Padmakumar, Yang Liu and Dilek Hakkani-Tur 11:15-11:30 (Metropolitan Centre)

Large pre-trained language models (PLMs) have been shown to retain implicit knowledge within their parameters. To enhance this implicit knowledge, we propose Knowledge Injection into Language Models (KILM), a novel approach that injects entity-related knowledge into encoder-decoder PLMs, via a generative knowledge infilling objective through continued pre-training. This is done without architectural modifications to the PLMs or adding additional parameters. Experimental results over a suite of knowledge-intensive tasks spanning numerous datasets show that KILM enables models to retain more knowledge and hallucinate less while preserving their original performance on general NLU and NLG tasks. KILM also demonstrates improved zero-shot performances on tasks such as entity disambiguation, outperforming state-of-the-art models having 30x more parameters.

When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

Alex Troy Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi and Hannaneh Hajishirzi 11:30-11:45 (Metropolitan Centre)

Despite their impressive performance on diverse tasks, large language models (LMs) still struggle with tasks requiring rich world knowledge, implying the difficulty of encoding a wealth of world knowledge in their parameters. This paper aims to understand LMs' strengths and limitations in memorizing factual knowledge, by conducting large-scale knowledge probing experiments on two open-domain-centric QA datasets: PopQA, our new dataset with 14k questions about long-tail entities, and EntityQuestions, a widely used open-domain QA dataset. We find that LMs struggle with less popular factual knowledge, and that retrieval augmentation helps significantly in these cases. Scaling, on the other hand, mainly improves memorization of popular knowledge, and fails to appreciably improve memorization of factual knowledge in the tail. Based on those findings, we devise a new method for retrieval-augmentation that improves performance and reduces inference costs by only retrieving non-parametric memories when necessary.

Unified Demonstration Retriever for In-Context Learning

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, GUOTONG XIE, Xiaoling Wang and Xipeng Qiu 11:45-12:00 (Metropolitan Centre)

In-context learning is a new learning paradigm where a language model conditions on a few input-output pairs (demonstrations) and a test input, and directly outputs the prediction. It has been shown sensitive to the provided demonstrations and thus promotes the research of demonstration retrieval: given a test input, relevant examples are retrieved from the training set to serve as informative demonstrations for in-context learning. While previous works train task-specific retrievers for several tasks separately, these methods are hard to transfer and scale on various tasks, and separately trained retrievers will cause a lot of parameter storage and deployment cost. In this paper, we propose Unified Demonstration Retriever (UDR), a single model to retrieve demonstrations for a wide range of tasks. To train UDR, we cast various tasks' training signals into a unified list-wise ranking formulation by language model's feedback. Then we propose a multi-task list-wise ranking training framework with an iterative mining strategy to find high-quality candidates, which can help UDR fully incorporate various tasks' signals. Experiments on 30+ tasks across 13 task families and multiple data domains show that UDR significantly outperforms baselines. Further analyses show the effectiveness of each proposed component and UDR's strong ability in various scenarios including different LMs (1.3B 175B), unseen datasets, varying demonstration quantities, etc. We will release the code and model checkpoint after review.

Measuring Inductive Biases of In-Context Learning with Underspecified Demonstrations

Chenglei Si, Dan Friedman, Nitish Joshi, Shi Feng, Danqi Chen and He He 12:00-12:15 (Metropolitan Centre)

In-context learning (ICL) is an important paradigm for adapting large language models (LLMs) to new tasks, but the generalization behavior of ICL remains poorly understood. We investigate the inductive biases of ICL from the perspective of feature bias: which feature ICL is more likely to use given a set of underspecified demonstrations in which two features are equally predictive of the labels. First, we characterize the feature biases of GPT-3 models by constructing underspecified demonstrations from a range of NLP datasets and feature combinations. We find that LLMs exhibit clear feature biases—for example, demonstrating a strong bias to predict labels according to sentiment rather than shallow lexical features, like punctuation. Second, we evaluate the effect of different interventions that are designed to impose an inductive bias in favor of a particular feature, such as adding a natural language instruction or using semantically relevant label words. We find that, while many interventions can influence the learner to prefer a particular feature, it can be difficult to overcome strong prior biases. Overall, our results provide a broader picture of the types of features that ICL may be more likely to exploit and how to impose inductive biases that are better aligned with the intended task.

Prompting PaLM for Translation: Assessing Strategies and Performance

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar and George Foster 12:15-12:30 (Metropolitan Centre)

Large language models (LLMs) that have been trained on multilingual but not parallel text exhibit a remarkable ability to translate between

languages. We probe this ability in an in-depth study of the pathways language model (PaLM), which has demonstrated the strongest machine translation (MT) performance among similarly-trained LLMs to date. We investigate various strategies for choosing translation examples for few-shot prompting, concluding that example quality is the most important factor. Using optimized prompts, we revisit previous assessments of PaLM's MT capabilities with more recent test sets, modern MT metrics, and human evaluation, and find that its performance, while impressive, still lags that of state-of-the-art supervised systems. We conclude by providing an analysis of PaLM's MT output which reveals some interesting properties and prospects for future work.

Question Answering

11:00-12:30 (Metropolitan West)

Multi-Source Test-Time Adaptation as Dueling Bandits for Extractive Question Answering
Hai Ye, Qizhe Xie and Hwee Tou Ng   11:00-11:15 (Metropolitan West)
In this work, we study multi-source test-time model adaptation from user feedback, where K distinct models are established for adaptation. To allow efficient adaptation, we cast the problem as a stochastic decision-making process, aiming to determine the best adapted model after adaptation. We discuss two frameworks: multi-armed bandit learning and multi-armed dueling bandits. Compared to multi-armed bandit learning, the dueling framework allows pairwise collaboration among K models, which is solved by a novel method named Co-UCB proposed in this work. Experiments on six datasets of extractive question answering (QA) show that the dueling framework using Co-UCB is more effective than other strong baselines for our studied problem.

Query Refinement Prompts for Closed-Book Long-Form QA
Reinald Kim Amplayo, Kellie Webster, Michael Collins, Dipanjan Das and Shashi Narayan   11:15-11:30 (Metropolitan West)
Large language models (LLMs) have been shown to perform well in answering questions and in producing long-form texts, both in few-shot closed-book settings. While the former can be validated using well-known evaluation metrics, the latter is difficult to evaluate. We resolve the difficulties to evaluate long-form output by doing both tasks at once – to do question answering that requires long-form answers. Such questions tend to be multifaceted, i.e., they may have ambiguities and/or require information from multiple sources. To this end, we define query refinement prompts that encourage LLMs to explicitly express the multifacetedness in questions and generate long-form answers covering multiple facets of the question. Our experiments on two long-form question answering datasets, ASQA and AQuAMuSe, show that using our prompts allows us to outperform fully finetuned models in the closed book setting, as well as achieve results comparable to retrieve-then-generate open-book models.

Won't Get Fooled Again: Answering Questions with False Premises
Shengding Hu, Yifan Luo, Huadong Wang, Xingyi Cheng, Zhiyuan Liu and Maosong Sun   11:30-11:45 (Metropolitan West)
Pre-trained language models (PLMs) have shown unprecedented potential in various fields, especially as the backbones for question-answering (QA) systems. However, they tend to be easily deceived by tricky questions such as "How many eyes does the sun have?". Such frailties of PLMs often allude to the lack of knowledge within them. In this paper, we find that the PLMs already possess the knowledge required to rebut such questions, and the key is how to activate the knowledge. To systematize this observation, we investigate the PLMs' responses to one kind of tricky questions, i.e., the false premises questions (FPQs). We annotate a FalseQA dataset containing 2365 human-written FPQs, with the corresponding explanations for the false premises and the revised true premise questions. Using FalseQA, we discover that PLMs are capable of discriminating FPQs by fine-tuning on moderate numbers (e.g., 256) of examples. PLMs also generate reasonable explanations for the false premise, which serve as rebuttals. Further replaying a few general questions during training allows PLMs to excel on FPQs and general questions simultaneously. Our work suggests that once the rebuttal ability is stimulated, knowledge inside the PLMs can be effectively utilized to handle FPQs, which incentivizes the research on PLM-based QA systems. The FalseQA dataset and code are available at https://github.com/thunlp/FalseQA .

Abductive Commonsense Reasoning Exploiting Mutually Exclusive Explanations
Wenting Zhao, Justin Chiu, Claire Cardie and Alexander Rush   11:45-12:00 (Metropolitan West)
Abductive reasoning aims to find plausible explanations for an event. This style of reasoning is critical for commonsense tasks where there are often multiple plausible explanations. Existing approaches for abductive reasoning in natural language processing (NLP) often rely on manually generated annotations for supervision; however, such annotations can be subjective and biased. Instead of using direct supervision, this work proposes an approach for abductive commonsense reasoning that exploits the fact that only a subset of explanations is correct for a given context. The method uses posterior regularization to enforce a mutual exclusion constraint, encouraging the model to learn the distinction between fluent explanations and plausible ones. We evaluate our approach on a diverse set of abductive reasoning datasets; experimental results show that our approach outperforms or is comparable to directly applying pretrained language models in a zero-shot manner and other knowledge-augmented zero-shot methods.

To Adapt or to Annotate: Challenges and Interventions for Domain Adaptation in Open-Domain Question Answering
Dheeru Dua, Emma Strubell, Sameer Singh and Pat Verga   12:00-12:15 (Metropolitan West)
Recent advances in open-domain question answering (ODQA) have demonstrated impressive accuracy on general-purpose domains like Wikipedia. While some work has been investigating how well ODQA models perform when tested for out-of-domain (OOD) generalization, these studies have been conducted only under conservative shifts in data distribution and typically focus on a single component (i.e., retriever or reader) rather than an end-to-end system. This work proposes a more realistic end-to-end domain shift evaluation setting covering five diverse domains. We not only find that end-to-end models fail to generalize but that high retrieval scores often still yield poor answer prediction accuracy. To address these failures, we investigate several interventions, in the form of data augmentations, for improving model adaption and use our evaluation set to elucidate the relationship between the efficacy of an intervention scheme and the particular type of dataset shifts we consider. We propose a generalizability test that estimates the type of shift in a target dataset without training a model in the target domain and that the type of shift is predictive of which data augmentation schemes will be effective for domain adaption. Overall, we find that these interventions increase end-to-end performance by up to 24 points.

Post-Abstention: Towards Reliably Re-Attempting the Abstained Instances in QA
Neeraj Varshney and Chitta Baral   12:15-12:30 (Metropolitan West)
Despite remarkable progress made in natural language processing, even the state-of-the-art models often make incorrect predictions. Such predictions hamper the reliability of systems and limit their widespread adoption in real-world applications. 'Selective prediction' partly addresses the above concern by enabling models to abstain from answering when their predictions are likely to be incorrect. While selective prediction is advantageous, it leaves us with a pertinent question 'what to do after abstention'. To this end, we present an explorative study on 'Post-Abstention', a task that allows re-attempting the abstained instances with the aim of increasing **coverage** of the system without

significantly sacrificing its **accuracy**. We first provide mathematical formulation of this task and then explore several methods to solve it. Comprehensive experiments on 11 QA datasets show that these methods lead to considerable risk improvements –performance metric of the Post-Abstention task– both in the in-domain and the out-of-domain settings. We also conduct a thorough analysis of these results which further leads to several interesting findings. Finally, we believe that our work will encourage and facilitate further research in this important area of addressing the reliability of NLP systems.

Posters

11:00-12:30 (Frontenac Ballroom and Queen's Quay)

[TACL] Generative Spoken Dialogue Language Modeling

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed and Emmanuel Dupoux 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

We introduce dGSLM, the first "textless" model able to generate audio samples of naturalistic spoken dialogues. It uses recent work on unsupervised spoken unit discovery coupled with a dual-tower transformer architecture with cross-attention trained on 2000 hours of two-channel raw conversational audio (Fisher dataset) without any text or labels. We show that our model is able to generate speech, laughter and other paralinguistic signals in the two channels simultaneously and reproduces more naturalistic and fluid turn taking compared to a text-based cascaded model.

[TACL] INSCIT: Information-Seeking Conversations with Mixed-Initiative Interactions

Zequi Wu, Ryu Parish, Hao Cheng, Sewon Min, Prithviraj Ammanabrolu, Mari Ostendorf and Hannaneh Hajishirzi 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

In an information-seeking conversation, a user may ask questions that are under-specified or unanswerable. An ideal agent would interact by initiating different response types according to the available knowledge sources. However, most current studies either fail to or artificially incorporate such agent-side initiative. This work presents INSCIT, a dataset for Information-Seeking Conversations with mixed-initiative Interactions. It contains 4.7K user-agent turns from 805 human-human conversations where the agent searches over Wikipedia and either directly answers, asks for clarification, or provides relevant information to address user queries. The data supports two subtasks, evidence passage identification and response generation, as well as a human evaluation protocol to assess model performance. We report results of two systems based on state-of-the-art models of conversational knowledge identification and open-domain question answering. Both systems significantly underperform humans, suggesting ample room for improvement in future studies.

[TACL] Explainable Abuse Detection as Intent Classification and Slot Filling

Agostina Calabrese, Björn Ross and Mirella Lapata 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

To proactively offer social media users a safe online experience, there is a need for systems that can detect harmful posts and promptly alert platform moderators. In order to guarantee the enforcement of a consistent policy, moderators are provided with detailed guidelines. In contrast, most state-of-the-art models learn what abuse is from labelled examples and as a result base their predictions on spurious cues, such as the presence of group identifiers, which can be unreliable. In this work we introduce the concept of policy-aware abuse detection, abandoning the unrealistic expectation that systems can reliably learn which phenomena constitute abuse from inspecting the data alone. We propose a machine-friendly representation of the policy that moderators wish to enforce, by breaking it down into a collection of intents and slots. We collect and annotate a dataset of 3,535 English posts with such slots, and show how architectures for intent classification and slot filling can be used for abuse detection, while providing a rationale for model decisions.

[TACL] FeelingBlue: A Corpus for Understanding the Emotional Connotation of Color in Context

Amith Ananthram, Olivia Winn and Smaranda Muresan 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

While the link between color and emotion has been widely studied, how context-based changes in color impact the intensity of perceived emotions is not well understood. In this work, we present a new multimodal dataset for exploring the emotional connotation of color as mediated by line, stroke, texture, shape and language. Our dataset, FeelingBlue, is a collection of 19,788 4-tuples of abstract art ranked by annotators according to their evoked emotions and paired with rationales for those annotations. Using this corpus, we present a baseline for a new task: Justified Affect Transformation. Given an image I, the task is to 1) recolor I to enhance a specified emotion e and 2) provide a textual justification for the change in e. Our model is an ensemble of deep neural networks which takes I, generates an emotionally transformed color palette p conditioned on I, applies p to I, and then justifies the color transformation in text via a visual-linguistic model. Experimental results shed light on the emotional connotation of color in context, demonstrating both the promise of our approach on this challenging task and the considerable potential for future investigations enabled by our corpus. Our dataset, code and models are available at https://github.com/amith-ananthram/feelingblue.

[TACL] Tracking Brand-Associated Polarity-Bearing Topics in User Reviews

Runcong Zhao, Lin Gui, Hanqi Yan and Yulan He 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Monitoring online customer reviews is important for business organisations to measure customer satisfaction and better manage their reputations. In this paper, we propose a novel dynamic Brand-Topic Model (dBTM) which is able to automatically detect and track brand-associated sentiment scores and polarity-bearing topics from product reviews organised in temporally-ordered time intervals. dBTM models the evolution of the latent brand polarity scores and the topic-word distributions over time by Gaussian state space models. It also incorporates a meta learning strategy to control the update of the topic-word distribution in each time interval in order to ensure smooth topic transitions and better brand score predictions. It has been evaluated on a dataset constructed from MakeupAlley reviews and a hotel review dataset. Experimental results show that dBTM outperforms a number of competitive baselines in brand ranking, achieving a good balance of topic coherence and uniqueness, and extracting well-separated polarity-bearing topics across time intervals.

[TACL] Efficient Long-Text Understanding with Short-Text Models

Maor Ivgi, Uri Shaham and Jonathan Berant 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Transformer-based pretrained language models (LMs) are ubiquitous across natural language understanding, but cannot be applied to long sequences such as stories, scientific articles and long documents, due to their quadratic complexity. While a myriad of efficient transformer variants have been proposed, they are typically based on custom implementations that require expensive pretraining from scratch.In this work, we propose SLED: SLiding-Encoder and Decoder, a simple approach for processing long sequences that re-uses and leverages battle-tested short-text pretrained LMs. Specifically, we partition the input into overlapping chunks, encode each with a short-text LM encoder and use the pretrained decoder to fuse information across chunks (fusion-in-decoder). We illustrate through controlled experiments that SLED offers a viable strategy for long text understanding and evaluate our approach on SCROLLS, a benchmark with seven datasets across a wide range of language understanding tasks. We find that SLED is competitive with specialized models that are up to 50x larger and require a dedicated

and expensive pretraining step.

[TACL] Rank-Aware Negative Training for Semi-Supervised Text Classification
Murtadha Ahmed, Shengfeng Pan, Wen Bo, Jianlin Su, Xinxin Cao, Wenze Zhang and Yunfeng Liu   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Semi-supervised text classification-based paradigms (SSTC) typically employ the spirit of self-training. The key idea is to train a deep classifier on limited labeled texts and then iteratively predict the unlabeled texts as their pseudo-labels for further training. However, the performance is largely affected by the accuracy of pseudo-labels, which may not be significant in real-world scenarios. This paper presents a Rank-aware Negative Training (RNT) framework to address SSTC in learning with noisy label manner. To alleviate the noisy information, we adapt a reasoning with uncertainty-based approach to rank the unlabeled texts based on the evidential support received from the labeled texts. Moreover, we propose the use of negative training to train RNT based on the concept that "the input instance does not belong to the complementary label". A complementary label is randomly selected from all labels except the label on-target. Intuitively, the probability of a true label serving as a complementary label is low and thus provides less noisy information during the training, resulting in better performance on the test data. Finally, we evaluate the proposed solution on various text classification benchmark datasets. Our extensive experiments show that it consistently overcomes the state-of-the-art alternatives in most scenarios and achieves competitive performance in the others. The code of RNT will be publicly available.

[TACL] How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN
Richard McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao and Asli Celikyilmaz   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Current language models can generate high quality text. Are they simply copying text they have seen before, or have they learned generalizable linguistic abstractions? To tease apart these possibilities, we introduce RAVEN, a suite of analyses for assessing the novelty of generated text, focusing on sequential structure (n-grams) and syntactic structure. We apply these analyses to four neural language models (an LSTM, a Transformer, Transformer-XL, and GPT-2). For local structure - e.g., individual dependencies - model-generated text is substantially less novel than our baseline of human-generated text from each model's test set. For larger-scale structure - e.g., overall sentence structure - model-generated text is as novel or even more novel than the human-generated baseline, but models still sometimes copy substantially, in some cases duplicating passages over 1,000 words long from the training set. We also perform extensive manual analysis showing that GPT-2 uses both compositional and analogical generalization mechanisms and that GPT-2's novel text is usually well-formed morphologically and syntactically but has reasonably frequent semantic issues (e.g., being self-contradictory).

[TACL] Collective Human Opinions in Semantic Textual Similarity
Wang Yuxia , Shimin Tao , Ning Xie , Hao Yang , Karin Verspoor and Timothy Baldwin 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Despite the subjective nature of semantic textual similarity (STS) and pervasive disagreements in STS annotation, existing benchmarks have used averaged human ratings as gold standard. Averaging masks the true distribution of human opinions on examples of low agreement, and prevents models from capturing the semantic vagueness that the individual ratings represent. In this work, we introduce USTS, the first Uncertainty-aware STS dataset with 15,000 Chinese sentence pairs and 150,000 labels, to study collective human opinions in STS. Analysis reveals that neither a scalar nor a single Gaussian fits a set of observed judgements adequately. We further show that current STS models cannot capture the variance caused by human disagreement on individual instances, but rather reflect the predictive confidence over the aggregate dataset.

[Demo] Disco: a toolkit for Distributional Control of Generative Models
Marc Dymetman, Jos Rozen and Germán Kruszewski   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Pre-trained language models and other generative models have revolutionized NLP and beyond. However, these models tend to reproduce undesirable biases present in their training data. Also, they may overlook patterns that are important but challenging to capture. To address these limitations, researchers have introduced distributional control techniques. These techniques, not limited to language, allow controlling the prevalence (i.e. expectations) of any features of interest in the model's outputs. Despite their potential, the widespread adoption of these techniques has been hindered by the difficulty in adapting the complex, disconnected code. Here, we present disco, an open-source Python library that brings these techniques to the broader public

[Demo] SciLit: A Platform for Joint Scientific Literature Discovery, Summarization and Citation Generation
Richard H.R. Hahnloser and Nianlong Gu   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Scientific writing involves retrieving, summarizing, and citing relevant papers, which can be time-consuming processes. Although in many workflows these processes are serially linked, there are opportunities for natural language processing (NLP) to provide end-to-end assistive tools. We propose SciLit, a pipeline that automatically recommends relevant papers, extracts highlights, and suggests a reference sentence as a citation of a paper, taking into consideration the user-provided context and keywords. SciLit efficiently recommends papers from large databases of hundreds of millions of papers using a two-stage pre-fetching and re-ranking literature search system that flexibly deals with addition and removal of a paper database. We provide a convenient user interface that displays the recommended papers as extractive summaries and that offers abstractively-generated citing sentences which are aligned with the provided context and which mention the chosen keyword(s). Our assistive tool for literature discovery and scientific writing is available at https://scilit.vercel.app

[Demo] OpenDelta: A Plug-and-play Library for Parameter-efficient Adaptation of Pre-trained Models
Maosong Sun, Zhiyuan Liu, Zhen Zhang, Xingtai Lv, Weilin Zhao, Ning Ding and Shengding Hu   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The scale of large pre-trained models (PTMs) poses significant challenges in adapting to downstream tasks due to the high optimization overhead and storage costs associated with full-parameter fine-tuning. To address this, many studies explore parameter-efficient tuning methods, also framed as "delta tuning" in Ding et al. (2022), which updates only a small subset of parameters, known as "delta modules", while keeping the backbone model's parameters fixed. However, the practicality and flexibility of delta tuning have been limited due to existing implementations that directly modify the code of the backbone PTMs and hard-code specific delta tuning methods for each PTM. In this paper, we present OpenDelta, an open-source library that overcomes these limitations by providing a plug-and-play implementation of various delta tuning methods. Our novel techniques eliminate the need to modify the backbone PTMs' code, making OpenDelta compatible with different, even novel PTMs. OpenDelta is designed to be simple, modular, and extensible, providing a comprehensive platform for researchers and practitioners to adapt large PTMs efficiently.

[Demo] ESPnet-ST-v2: Multipurpose Spoken Language Translation Toolkit
Shinji Watanabe, Juan Pino, Soumi Maiti, Moto Hira, Zhaoheng Ni, Xiaohui Zhang, Tomoki Hayashi, Dan Berrebbi, Patrick Fernandes, Peter Polak, Siddharth Dalmia, Yifan Peng, Hirofumi Inaguma, Yun Tang, Jiatong Shi and Brian Yan   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
ESPnet-ST-v2 is a revamp of the open-source ESPnet-ST toolkit necessitated by the broadening interests of the spoken language translation community. ESPnet-ST-v2 supports 1) offline speech-to-text translation (ST), 2) simultaneous speech-to-text translation (SST), and 3) offline speech-to-speech translation (S2ST) – each task is supported with a wide variety of approaches, differentiating ESPnet-ST-v2 from other open

source spoken language translation toolkits. This toolkit offers state-of-the-art architectures such as transducers, hybrid CTC/attention, multi-decoders with searchable intermediates, time-synchronous blockwise CTC/attention, Translatotron models, and direct discrete unit models. In this paper, we describe the overall design, example models for each task, and performance benchmarking behind ESPnet-ST-v2, which is publicly available at https://github.com/espnet/espnet.

[Demo] CB2: Collaborative Natural Language Interaction Research Platform
Yoav Artzi, Mustafa Omer Gul and Jacob Sharf   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
CB2 is a multi-agent platform to study collaborative natural language interaction in a grounded task-oriented scenario. It includes a 3D game environment, a backend server designed to serve trained models to human agents, and various tools and processes to enable scalable studies. We deploy CB2 at https://cb2.ai as a system demonstration with a learned instruction following model.

[Demo] Fast Whitespace Correction with Encoder-Only Transformers
Sebastian Walter, Matthias Hertel and Hannah Bast   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The goal of whitespace correction is to fix space errors in arbitrary given text. For example, given the text "whi te space correctio nwithTransf or mers", produce "whitespace correction with Transformers". We compare two Transformer-based models, a character-level encoder-decoder model and a byte-level encoder-only model. We find that the encoder-only model is both faster and achieves higher quality. We provide an easy-to-use tool that is over 900 times faster than the previous best tool, with the same high quality. Our tool repairs text at a rate of over 200 kB/s on GPU, with a sequence-averaged F1-score ranging from 87.5

WikiBio: a Semantic Resource for the Intersectional Analysis of Biographical Events
Marco Antonio Stranisci, Rossana Damiano, Enrico Mensa, Viviana Patti, Daniele P. Radicioni and Tommaso Caselli   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Biographical event detection is a relevant task that allows for the exploration and comparison of the ways in which people's lives are told and represented. This may support several real-life applications in digital humanities and in works aimed at exploring bias about minoritized groups. Despite that, there are no corpora and models specifically designed for this task. In this paper we fill this gap by presenting a new corpus annotated for biographical event detection. The corpus, which includes 20 Wikipedia biographies, was aligned with 5 existing corpora in order to train a model for the biographical event detection task. The model was able to detect all mentions of the target-entity in a biography with an F-score of 0.808 and the entity-related events with an F-score of 0.859. Finally, the model was used for performing an analysis of biases about women and non-Western people in Wikipedia biographies.

Bhasa-Abhijnaanam: Native-script and romanized Language Identification for 22 Indic languages
Yash H. Madhani, Mitesh M. Khapra and Anoop Kunchukuttan   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
We create publicly available language identification (LID) datasets and models in all 22 Indian languages listed in the Indian constitution in both native-script and romanized text. First, we create Bhasha-Abhijnaanam, a language identification test set for native-script as well as romanized text which spans all 22 Indic languages. We also train IndicLID, a language identifier for all the above-mentioned languages in both native and romanized script. For native-script text, it has better language coverage than existing LIDs and is competitive or better than other LIDs. IndicLID is the first LID for romanized text in Indian languages. Two major challenges for romanized text LID are the lack of training data and low-LID performance when languages are similar. We provide simple and effective solutions to these problems. In general, there has been limited work on romanized text in any language, and our findings are relevant to other languages that need romanized language identification. Our models are publicly available at https://github.com/AI4Bharat/IndicLID under open-source licenses. Our training and test sets are also publicly available at https://huggingface.co/datasets/ai4bharat/Bhasha-Abhijnaanam under open-source licenses.

EPIC: Multi-Perspective Annotation of a Corpus of Irony
Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco and Davide Bernardi   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
We present EPIC (English Perspectivist Irony Corpus), the first annotated corpus for irony analysis based on the principles of data perspectivism. The corpus contains short conversations from social media in five regional varieties of English, and it is annotated by contributors from five countries corresponding to those varieties. We analyse the resource along the perspectives induced by the diversity of the annotators, in terms of origin, age, and gender, and the relationship between these dimensions, irony, and the topics of conversation. We validate EPIC by creating perspective-aware models that encode the perspectives of annotators grouped according to their demographic characteristics. Firstly, the performance of perspectivist models confirms that different annotators induce very different models. Secondly, in the classification of ironic and non-ironic texts, perspectivist models prove to be generally more confident than the non-perspectivist ones. Furthermore, comparing the performance on a perspective-based test set with those achieved on a gold standard test set, we can observe how perspectivist models tend to detect more precisely the positive class, showing their ability to capture the different perceptions of irony. Thanks to these models, we are moreover able to show interesting insights about the variation in the perception of irony by the different groups of annotators, such as among different generations and nationalities.

Do language models have coherent mental models of everyday things?
Yuling Gu, Bhavana Dalvi Mishra and Peter Clark   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
When people think of everyday things like an egg, they typically have a mental image associated with it. This allows them to correctly judge, for example, that "the yolk surrounds the shell" is a false statement. Do language models similarly have a coherent picture of such everyday things? To investigate this, we propose a benchmark dataset consisting of 100 everyday things, their parts, and the relationships between these parts, expressed as 11,720 "X relation Y?" true/false questions. Using these questions as probes, we observe that state-of-the-art pre-trained language models (LMs) like GPT-3 and Macaw have fragments of knowledge about these everyday things, but do not have fully coherent "parts mental models" (54-59% accurate, 19-43% conditional constraint violation). We propose an extension where we add a constraint satisfaction layer on top of the LM's raw predictions to apply commonsense constraints. As well as removing inconsistencies, we find that this also significantly improves accuracy (by 16-20%), suggesting how the incoherence of the LM's pictures of everyday things can be significantly reduced.

NLPeer: A Unified Resource for the Computational Study of Peer Review
Nils Dycke, Ilia Kuznetsov and Iryna Gurevych   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Peer review constitutes a core component of scholarly publishing; yet it demands substantial expertise and training, and is susceptible to errors and biases. Various applications of NLP for peer reviewing assistance aim to support reviewers in this complex process, but the lack of clearly licensed datasets and multi-domain corpora prevent the systematic study of NLP for peer review. To remedy this, we introduce NLPeer– the first ethically sourced multidomain corpus of more than 5k papers and 11k review reports from five different venues. In addition to the new datasets of paper drafts, camera-ready versions and peer reviews from the NLP community, we establish a unified data representation and augment previous peer review datasets to include parsed and structured paper representations, rich metadata and versioning information. We complement our resource with implementations and analysis of three reviewing assistance tasks, including a novel guided skimming task. Our work paves the path towards systematic, multi-faceted, evidence-based study of peer review in NLP and beyond. The data and code are publicly available.

Multilingual Multifaceted Understanding of Online News in Terms of Genre, Framing, and Persuasion Techniques

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino and Preslav Nakov 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

We present a new multilingual multifacet dataset of news articles, each annotated for genre (objective news reporting vs. opinion vs. satire), framing (what key aspects are highlighted), and persuasion techniques (logical fallacies, emotional appeals, ad hominem attacks, etc.). The persuasion techniques are annotated at the span level, using a taxonomy of 23 fine-grained techniques grouped into 6 coarse categories. The dataset contains 1,612 news articles covering recent news on current topics of public interest in six European languages (English, French, German, Italian, Polish, and Russian), with more than 37k annotated spans of persuasion techniques. We describe the dataset and the annotation process, and we report the evaluation results of multilabel classification experiments using state-of-the-art multilingual transformers at different levels of granularity: token-level, sentence-level, paragraph-level, and document-level.

Direct Fact Retrieval from Knowledge Graphs without Entity Linking

Jinheon Baek, Alham Fikri Aji, Jens Lehmann and Sung Ju Hwang 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

There has been a surge of interest in utilizing Knowledge Graphs (KGs) for various natural language processing/understanding tasks. The conventional mechanism to retrieve facts in KGs usually involves three steps: entity span detection, entity disambiguation, and relation classification. However, this approach requires additional labels for training each of the three subcomponents in addition to pairs of input texts and facts, and also may accumulate errors propagated from failures in previous steps. To tackle these limitations, we propose a simple knowledge retrieval framework, which directly retrieves facts from the KGs given the input text based on their representational similarities, which we refer to as Direct Fact Retrieval (DiFaR). Specifically, we first embed all facts in KGs onto a dense embedding space by using a language model trained by only pairs of input texts and facts, and then provide the nearest facts in response to the input text. Since the fact, consisting of only two entities and one relation, has little context to encode, we propose to further refine ranks of top-k retrieved facts with a reranker that contextualizes the input text and the fact jointly. We validate our DiFaR framework on multiple fact retrieval tasks, showing that it significantly outperforms relevant baselines that use the three-step approach.

Training Models to Generate, Recognize, and Reframe Unhelpful Thoughts

Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather M Foran and Y-Lan Boureau 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Many cognitive approaches to well-being, such as recognizing and reframing unhelpful thoughts, have received considerable empirical support over the past decades, yet still lack truly widespread adoption in self-help format. A barrier to that adoption is a lack of adequately specific and diverse dedicated practice material. This work examines whether current language models can be leveraged to both produce a virtually unlimited quantity of practice material illustrating standard unhelpful thought patterns matching specific given contexts, and generate suitable positive reframing proposals. We propose PATTERNREFRAME, a novel dataset of about 10k examples of thoughts containing unhelpful thought patterns conditioned on a given persona, accompanied by about 27k positive reframes. By using this dataset to train and/or evaluate current models, we show that existing models can already be powerful tools to help generate an abundance of tailored practice material and hypotheses, with no or minimal additional model training required.

ESCOXLM-R: Multilingual Taxonomy-driven Pre-training for the Job Market Domain

Mike Zhang, Rob van der Goot and Barbara Plank 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

The increasing number of benchmarks for Natural Language Processing (NLP) tasks in the computational job market domain highlights the demand for methods that can handle job-related tasks such as skill extraction, skill classification, job title classification, and de-identification. While some approaches have been developed that are specific to the job market domain, there is a lack of generalized, multilingual models and benchmarks for these tasks. In this study, we introduce a language model called ESCOXLM-R, based on XLM-R-large, which uses domain-adaptive pre-training on the European Skills, Competences, Qualifications and Occupations (ESCO) taxonomy, covering 27 languages. The pre-training objectives for ESCOXLM-R include dynamic masked language modeling and a novel additional objective for inducing multilingual taxonomical ESCO relations. We comprehensively evaluate the performance of ESCOXLM-R on 6 sequence labeling and 3 classification tasks in 4 languages and find that it achieves state-of-the-art results on 6 out of 9 datasets. Our analysis reveals that ESCOXLM-R performs better on short spans and outperforms XLM-R-large on entity-level and surface-level span-F1, likely due to ESCO containing short skill and occupation titles, and encoding information on the entity-level.

Multimodal Persona Based Generation of Comic Dialogs

Harsh Agrawal, Aditya M. Mishra, Manish Gupta and Mausam - 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

We focus on the novel problem of persona based dialogue generation for comic strips. Dialogs in comic strips is a unique and unexplored area where every strip contains utterances from various characters with each one building upon the previous utterances and the associated visual scene. Previous works like DialoGPT, PersonaGPT and other dialog generation models encode two-party dialogues and do not account for the visual information. To the best of our knowledge we are the first to propose the paradigm of multimodal persona based dialogue generation. We contribute a novel dataset, ComSet, consisting of 54K strips, harvested from 13 popular comics available online. Further, we propose a multimodal persona-based architecture, MPDialog, to generate dialogues for the next panel in the strip which decreases the perplexity score by 10 points over strong dialogue generation baseline models. We demonstrate that there is still ample opportunity for improvement, highlighting the importance of building stronger dialogue systems that are able to generate persona-consistent dialogues and understand the context through various modalities.

Contextual Knowledge Learning for Dialogue Generation

Wen Zheng, Natasa Milic-Frayling and Ke Zhou 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Incorporating conversational context and knowledge into dialogue generation models has been essential for improving the quality of the generated responses. The context, comprising utterances from previous dialogue exchanges, is used as a source of content for response generation and as a means of selecting external knowledge. However, to avoid introducing irrelevant content, it is key to enable fine-grained scoring of context and knowledge. In this paper, we present a novel approach to context and knowledge weighting as an integral part of model training. We guide the model training through a Contextual Knowledge Learning (CKL) process which involves Latent Vectors for context and knowledge, respectively. CKL Latent Vectors capture the relationship between context, knowledge, and responses through weak supervision and enable differential weighting of context utterances and knowledge sentences during the training process. Experiments with two standard datasets and human evaluation demonstrate that CKL leads to a significant improvement compared with the performance of six strong baseline models and shows robustness with regard to reduced sizes of training sets.

Improved Instruction Ordering in Recipe-Grounded Conversation

Duong Minh Le, Ruohao Guo, Wei Xu and Alan Ritter 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

In this paper, we study the task of instructional dialogue and focus on the cooking domain. Analyzing the generated output of the GPT-J model, we reveal that the primary challenge for a recipe-grounded dialog system is how to provide the instructions in the correct order. We hypothesize that this is due to the model's lack of understanding of user intent and inability to track the instruction state (i.e., which step was

last instructed). Therefore, we propose to explore two auxiliary subtasks, namely User Intent Detection and Instruction State Tracking, to support Response Generation with improved instruction grounding. Experimenting with our newly collected dataset, ChattyChef, shows that incorporating user intent and instruction state information helps the response generation model mitigate the incorrect order issue. Furthermore, to investigate whether ChatGPT has completely solved this task, we analyze its outputs and find that it also makes mistakes (10.7% of the responses), about half of which are out-of-order instructions. We will release ChattyChef to facilitate further research in this area at: https://github.com/octaviaguo/ChattyChef.

A Survey on Asking Clarification Questions Datasets in Conversational Systems
Hossein A. Rahmani, Xi Wang, Yue Feng, Qiang Zhang, Emine Yilmaz and Aldo Lipani 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The ability to understand a user's underlying needs is critical for conversational systems, especially with limited input from users in a conversation. Thus, in such a domain, Asking Clarification Questions (ACQs) to reveal users' true intent from their queries or utterances arise as an essential task. However, it is noticeable that a key limitation of the existing ACQs studies is their incomparability, from inconsistent use of data, distinct experimental setups and evaluation strategies. Therefore, in this paper, to assist the development of ACQs techniques, we comprehensively analyse the current ACQs research status, which offers a detailed comparison of publicly available datasets, and discusses the applied evaluation metrics, joined with benchmarks for multiple ACQs-related tasks. In particular, given a thorough analysis of the ACQs task, we discuss a number of corresponding research directions for the investigation of ACQs as well as the development of conversational systems.

FC-KBQA: A Fine-to-Coarse Composition Framework for Knowledge Base Question Answering
Lingxi Zhang, Jing Zhang, Yanling Wang, Shulin Cao, Xinmei Huang, Cuiping Li, Hong Chen and Juanzi Li 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The generalization problem on KBQA has drawn considerable attention. Existing research suffers from the generalization issue brought by the entanglement in the coarse-grained modeling of the logical expression, or inexecutability issues due to the fine-grained modeling of disconnected classes and relations in real KBs. We propose a Fine-to-Coarse Composition framework for KBQA (FC-KBQA) to both ensure the generalization ability and executability of the logical expression. The main idea of FC-KBQA is to extract relevant fine-grained knowledge components from KB and reformulate them into middle-grained knowledge pairs for generating the final logical expressions. FC-KBQA derives new state-of-the-art performance on GrailQA and WebQSP, and runs 4 times faster than the baseline. Our code is now available at GitHub https://github. com/RUCKBReasoning/FC-KBQA.

MeetingQA: Extractive Question-Answering on Meeting Transcripts
Archiki Prasad, Trung Bui, Seunghyun Yoon, Hanieh Deilamsalehy, Franck Dernoncourt and Mohit Bansal 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
With the ubiquitous use of online meeting platforms and robust automatic speech recognition systems, meeting transcripts have emerged as a promising domain for natural language tasks. Most recent works on meeting transcripts primarily focus on summarization and extraction of action items. However, meeting discussions also have a useful question-answering (QA) component, crucial to understanding the discourse or meeting content, and can be used to build interactive interfaces on top of long transcripts. Hence, in this work, we leverage this inherent QA component of meeting discussions and introduce MeetingQA, an extractive QA dataset comprising of questions asked by meeting participants and corresponding responses. As a result, questions can be open-ended and actively seek discussions, while the answers can be multi-span and distributed across multiple speakers. Our comprehensive empirical study of several robust baselines including long-context language models and recent instruction-tuned models reveals that models perform poorly on this task (F1 = 57.3) and severely lag behind human performance (F1 = 84.6), thus presenting a challenging new task for the community to improve upon.

MAD-TSC: A Multilingual Aligned News Dataset for Target-dependent Sentiment Classification
Evan Dufraisse, Adrian Popescu, Julien Tourille, Armelle Brun and Jerome Deshayes 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Target-dependent sentiment classification (TSC) enables a fine-grained automatic analysis of sentiments expressed in texts. Sentiment expression varies depending on the domain, and it is necessary to create domain-specific datasets. While socially important, TSC in the news domain remains relatively understudied. We introduce MAD-TSC, a new dataset which differs substantially from existing resources. First, it includes aligned examples in eight languages to facilitate a comparison of performance for individual languages, and a direct comparison of human and machine translation. Second, the dataset is sampled from a diversified parallel news corpus, and is diversified in terms of news sources and geographic spread of entities. Finally, MAD-TSC is more challenging than existing datasets because its examples are more complex. We exemplify the use of MAD-TSC with comprehensive monolingual and multilingual experiments. The latter show that machine translations can successfully replace manual ones, and that performance for all included languages can match that of English by automatically translating test examples.

Trigger Warning Assignment as a Multi-Label Document Classification Problem
Matti Wiegmann, Magdalena Wolska, Christopher Schröder, Ole Borchardt, Benno Stein and Martin Potthast 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
A trigger warning is used to warn people about potentially disturbing content. We introduce trigger warning assignment as a multi-label classification task, create the Webis Trigger Warning Corpus 2022, and with it the first dataset of 1 million fanfiction works from Archive of our Own with up to 36 different warnings per document. To provide a reliable catalog of trigger warnings, we organized 41 million of free-form tags assigned by fanfiction authors into the first comprehensive taxonomy of trigger warnings by mapping them to the 36 institutionally recommended warnings. To determine the best operationalization of trigger warnings, we explore state-of-the-art multi-label models, examining the trade-off between assigning coarse- and fine-grained warnings, open- and closed-set classification, document length, and label confidence. Our models achieve micro-F1 scores of about 0.5, which reveals the difficulty of the task. Tailored representations, long input sequences, and a higher recall on rare warnings would help.

MEMEX: Detecting Explanatory Evidence for Memes via Knowledge-Enriched Contextualization
Shivam Sharma, Ramaneswaran S, Udit Arora, Md. Shad Akhtar and Tanmoy Chakraborty 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Memes are a powerful tool for communication over social media. Their affinity for evolving across politics, history, and sociocultural phenomena renders them an ideal vehicle for communication. To comprehend the subtle message conveyed within a meme, one must understand the relevant background that facilitates its holistic assimilation. Besides digital archiving of memes and their metadata by a few websites like knowyourmeme.com, currently, there is no efficient way to deduce a meme's context dynamically. In this work, we propose a novel task, MEMEX - given a meme and a related document, the aim is to mine the context that succinctly explains the background of the meme. At first, we develop MCC (Meme Context Corpus), a novel dataset for MEMEX. Further, to benchmark MCC, we propose MIME (MultImodal Meme Explainer) that uses external knowledge-enriched meme representation and a multi-level approach to capture the cross-modal semantic dependencies between the meme and the context. MIME surpasses several unimodal and multimodal systems and yields an absolute improvement of 4% F1-score over the best baseline. Lastly, we conduct detailed analyses of MIME's performance, highlighting the aspects that could lead to optimal modeling of cross-modal contextual associations.

Downstream Datasets Make Surprisingly Good Pretraining Corpora

Kundan Krishna, Saurabh Garg, Jeffrey P. Bigham and Zachary Lipton   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
For most natural language processing tasks, the dominant practice is to finetune large pretrained transformer models (e.g., BERT) using smaller downstream datasets. Despite the success of this approach, it remains unclear to what extent these gains are attributable to the massive background corpora employed for pretraining versus to the pretraining objectives themselves. This paper introduces a large-scale study of self-pretraining, where the same (downstream) training data is used for both pretraining and finetuning. In experiments addressing both ELECTRA and RoBERTa models and 10 distinct downstream classification datasets, we observe that self-pretraining rivals standard pretraining on the BookWiki corpus (despite using around 10x–500x less data), outperforming the latter on 7 and 5 datasets, respectively. Surprisingly, these task-specific pretrained models often perform well on other tasks, including the GLUE benchmark. Besides classification tasks, self-pretraining also provides benefits on structured output prediction tasks such as span based question answering and commonsense inference, often providing more than 50% of the performance boosts provided by pretraining on the BookWiki corpus. Our results hint that in many scenarios, performance gains attributable to pretraining are driven primarily by the pretraining objective itself and are not always attributable to the use of external pretraining data in massive amounts. These findings are especially relevant in light of concerns about intellectual property and offensive content in web-scale pretraining data.

miCSE: Mutual Information Contrastive Learning for Low-shot Sentence Embeddings

Tassilo Klein and Moin Nabi   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
This paper presents miCSE, a mutual information-based contrastive learning framework that significantly advances the state-of-the-art in few-shot sentence embedding. The proposed approach imposes alignment between the attention pattern of different views during contrastive learning. Learning sentence embeddings with miCSE entails enforcing the structural consistency across augmented views for every sentence, making contrastive self-supervised learning more sample efficient. As a result, the proposed approach shows strong performance in the few-shot learning domain. While it achieves superior results compared to state-of-the-art methods on multiple benchmarks in few-shot learning, it is comparable in the full-shot scenario. This study opens up avenues for efficient self-supervised learning methods that are more robust than current contrastive methods for sentence embedding.

Randomized Positional Encodings Boost Length Generalization of Transformers

Anian Ruoss, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg and Joel Veness 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Transformers have impressive generalization capabilities on tasks with a fixed context length. However, they fail to generalize to sequences of arbitrary length, even for seemingly simple tasks such as duplicating a string. Moreover, simply training on longer sequences is inefficient due to the quadratic computation complexity of the global attention mechanism. In this work, we demonstrate that this failure mode is linked to positional encodings being out-of-distribution for longer sequences (even for relative encodings) and introduce a novel family of positional encodings that can overcome this problem. Concretely, our randomized positional encoding scheme simulates the positions of longer sequences and randomly selects an ordered subset to fit the sequence's length. Our large-scale empirical evaluation of 6000 models across 15 algorithmic reasoning tasks shows that our method allows Transformers to generalize to sequences of unseen length (increasing test accuracy by 12.0% on average).

Large-scale Lifelong Learning of In-context Instructions and How to Tackle It

Jisoo Mok, Jaeyoung Do, Sungjin Lee, Tara Taghavi, Seunghak Yu and Sungroh Yoon   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Jointly fine-tuning a Pre-trained Language Model (PLM) on a pre-defined set of tasks with in-context instructions has been proven to improve its generalization performance, allowing us to build a universal language model that can be deployed across task boundaries. In this work, we explore for the first time whether this attractive property of in-context instruction learning can be extended to a scenario in which tasks are fed to the target PLM in a sequential manner. The primary objective of so-called lifelong in-context instruction learning is to improve the target PLM's instance- and task-level generalization performance as it observes more tasks. DynaInst, the proposed method to lifelong in-context instruction learning, achieves noticeable improvements in both types of generalization, nearly reaching the upper bound performance obtained through joint training.

HyperMixer: An MLP-based Low Cost Alternative to Transformers

Florian Mai, Arnaud Pannatier, Fabio J Fehr, Haolin Chen, Francois Marelli, Francois Fleuret and James Henderson   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Transformer-based architectures are the model of choice for natural language understanding, but they come at a significant cost, as they have quadratic complexity in the input length, require a lot of training data, and can be difficult to tune. In the pursuit of lower costs, we investigate simple MLP-based architectures. We find that existing architectures such as MLPMixer, which achieves token mixing through a static MLP applied to each feature independently, are too detached from the inductive biases required for natural language understanding. In this paper, we propose a simple variant, HyperMixer, which forms the token mixing MLP dynamically using hypernetworks. Empirically, we demonstrate that our model performs better than alternative MLP-based models, and on par with Transformers. In contrast to Transformers, HyperMixer achieves these results at substantially lower costs in terms of processing time, training data, and hyperparameter tuning.

HuCurl: Human-induced Curriculum Discovery

Mohamed Elgaar and Hadi Amiri   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
We introduce the problem of curriculum discovery and describe a curriculum learning framework capable of discovering effective curricula in a curriculum space based on prior knowledge about sample difficulty. Using annotation entropy and loss as measures of difficulty, we show that (i): the top-performing discovered curricula for a given model and dataset are often non-monotonic as apposed to monotonic curricula in existing literature, (ii): the prevailing easy-to-hard or hard-to-easy transition curricula are often at the risk of underperforming, and (iii): the curricula discovered for smaller datasets and models perform well on larger datasets and models respectively. The proposed framework encompasses some of the existing curriculum learning approaches and can discover curricula that outperform them across several NLP tasks.

Training-free Neural Architecture Search for RNNs and Transformers

Aaron Serianni and Jugal Kalita   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Neural architecture search (NAS) has allowed for the automatic creation of new and effective neural network architectures, offering an alternative to the laborious process of manually designing complex architectures. However, traditional NAS algorithms are slow and require immense amounts of computing power. Recent research has investigated training-free NAS metrics for image classification architectures, drastically speeding up search algorithms. In this paper, we investigate training-free NAS metrics for recurrent neural network (RNN) and BERT-based transformer architectures, targeted towards language modeling tasks. First, we develop a new training-free metric, named hidden covariance, that predicts the trained performance of an RNN architecture and significantly outperforms existing training-free metrics. We experimentally evaluate the effectiveness of the hidden covariance metric on the NAS-Bench-NLP benchmark. Second, we find that the current search space paradigm for transformer architectures is not optimized for training-free neural architecture search. Instead, a simple qualitative analysis can effectively shrink the search space to the best performing architectures. This conclusion is based on our investigation

of existing training-free metrics and new metrics developed from recent transformer pruning literature, evaluated on our own benchmark of trained BERT architectures. Ultimately, our analysis shows that the architecture search space and the training-free metric must be developed together in order to achieve effective results. Our source code is available at https://github.com/aaronserianni/training-free-nas.

Rethinking the Role of Scale for In-Context Learning: An Interpretability-based Case Study at 66 Billion Scale

Hritik Bansal, Karthik Gopalakrishnan, Saket Dingliwal, Sravan Bodapati, Katrin Kirchhoff and Dan Roth 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Language models have been shown to perform better with an increase in scale on a wide variety of tasks via the in-context learning paradigm. In this paper, we investigate the hypothesis that the ability of a large language model to in-context learn-perform a task is not uniformly spread across all of its underlying components. Using a 66 billion parameter language model (OPT-66B) across a diverse set of 14 downstream tasks, we find this is indeed the case: 70% of the attention heads and 20% of the feed forward networks can be removed with minimal decline in task performance. We find substantial overlap in the set of attention heads (un)important for in-context learning across tasks and number of in-context examples. We also address our hypothesis through a task-agnostic lens, finding that a small set of attention heads in OPT-66B score highly on their ability to perform primitive induction operations associated with in-context learning, namely, prefix matching and copying. These induction heads overlap with task-specific important heads, reinforcing arguments by Olsson et al. (2022) regarding induction head generality to more sophisticated behaviors associated with in-context learning. Overall, our study provides several insights that indicate large language models may be under-trained for in-context learning and opens up questions on how to pre-train language models to more effectively perform in-context learning.

In-Context Analogical Reasoning with Pre-Trained Language Models

Xiaoyang Hu, Shane Storks, Richard L. Lewis and Joyce Chai 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Analogical reasoning is a fundamental capacity of human cognition that allows us to reason abstractly about novel situations by relating them to past experiences. While it is thought to be essential for robust reasoning in AI systems, conventional approaches require significant training and/or hard-coding of domain knowledge to be applied to benchmark tasks. Inspired by cognitive science research that has found connections between human language and analogy-making, we explore the use of intuitive language-based abstractions to support analogy in AI systems. Specifically, we apply large pre-trained language models (PLMs) to visual Raven's Progressive Matrices (RPM), a common relational reasoning test. By simply encoding the perceptual features of the problem into language form, we find that PLMs exhibit a striking capacity for zero-shot relational reasoning, exceeding human performance and nearing supervised vision-based methods. We explore different encodings that vary the level of abstraction over task features, finding that higher-level abstractions further strengthen PLMs' analogical reasoning. Our detailed analysis reveals insights on the role of model complexity, in-context learning, and prior knowledge in solving RPM tasks.

Symbolic Chain-of-Thought Distillation: Small Models Can Also "Think" Step-by-Step

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang and Yejin Choi 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Chain-of-thought prompting (e.g., "Let's think step-by-ste") primes large language models to verbalize rationalization for their predictions. While chain-of-thought can lead to dramatic performance gains, benefits appear to emerge only for sufficiently large models (beyond 50B parameters). We show that orders-of-magnitude smaller models (125M—1.3B parameters) can still benefit from chain-of-thought prompting. To achieve this, we introduce Symbolic Chain-of-Thought Distillation (SCoTD), a method to train a smaller student model on rationalizations sampled from a significantly larger teacher model. Experiments across several commonsense benchmarks show that: 1) SCoTD enhances the performance of the student model in both supervised and few-shot settings, and especially for challenge sets; 2) sampling many reasoning chains per instance from the teacher is paramount; and 3) after distillation, student chain-of-thoughts are judged by humans as comparable to the teacher, despite orders of magnitude fewer parameters. We test several hypotheses regarding what properties of chain-of-thought samples are important, e.g., diversity vs. teacher likelihood vs. open-endedness. We release our corpus of chain-of-thought samples and code.

Did You Read the Instructions? Rethinking the Effectiveness of Task Definitions in Instruction Learning

Fan Yin, Jesse Vig, Philippe Laban, Shafiq Joty, Caiming Xiong and Chien-Sheng Jason Wu 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Large language models (LLMs) have shown impressive performance in following natural language instructions to solve unseen tasks. However, it remains unclear whether models truly understand task definitions and whether the human-written definitions are optimal. In this paper, we systematically study the role of task definitions in instruction learning. We first conduct an ablation analysis informed by human annotations to understand which parts of a task definition are most important, and find that model performance only drops substantially when removing contents describing the task output, in particular label information. Next, we propose an automatic algorithm to compress task definitions to a minimal supporting set of tokens, and find that 60% of tokens can be removed while maintaining or even improving model performance. Based on these results, we propose two strategies to help models better leverage task instructions: (1) providing only key information for tasks in a common structured format, and (2) adding a meta-tuning stage to help the model better understand the definitions. With these two strategies, we achieve a 4.2 Rouge-L improvement over 119 unseen test tasks.

Text Style Transfer with Contrastive Transfer Pattern Mining

Jingxuan Han, Quan Wang, Licheng Zhang, Weidong Chen, Yan Song and Zhendong Mao 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Text style transfer (TST) is an important task in natural language generation, which aims to alter the stylistic attributes (e.g., sentiment) of a sentence and keep its semantic meaning unchanged. Most existing studies mainly focus on the transformation between styles, yet ignore that this transformation can be actually carried out via different hidden transfer patterns. To address this problem, we propose a novel approach, contrastive transfer pattern mining (CTPM), which automatically mines and utilizes inherent latent transfer patterns to improve the performance of TST. Specifically, we design an adaptive clustering module to automatically discover hidden transfer patterns from the data, and introduce contrastive learning based on the discovered patterns to obtain more accurate sentence representations, and thereby benefit the TST task. To the best of our knowledge, this is the first work that proposes the concept of transfer patterns in TST, and our approach can be applied in a plug-and-play manner to enhance other TST methods to further improve their performance. Extensive experiments on benchmark datasets verify the effectiveness and generality of our approach.

Faithful Low-Resource Data-to-Text Generation through Cycle Training

Zhuoer Wang, Marcus Collins, Nikhita Vedula, Simone Filice, Shervin Malmasi and Oleg Rokhlenko 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Methods to generate text from structured data have advanced significantly in recent years, primarily due to fine-tuning of pre-trained language models on large datasets. However, such models can fail to produce output faithful to the input data, particularly on out-of-domain data. Sufficient annotated data is often not available for specific domains, leading us to seek an unsupervised approach to improve the faithfulness of output text. Since the problem is fundamentally one of consistency between the representations of the structured data and text, we evaluate the effectiveness of cycle training in this work. Cycle training uses two models which are inverses of each other: one that generates text from structured data, and one which generates the structured data from natural language text. We show that cycle training, when initialized with

a small amount of supervised data (100 samples in our case), achieves nearly the same performance as fully supervised approaches for the data-to-text generation task on the WebNLG, E2E, WTQ, and WSQL datasets. We perform extensive empirical analysis with automated evaluation metrics and a newly designed human evaluation schema to reveal different cycle training strategies' effectiveness of reducing various types of generation errors. Our code is publicly available at https://github.com/Edillower/CycleNLG.

SSD-LM: Semi-autoregressive Simplex-based Diffusion Language Model for Text Generation and Modular Control
Xiaochuang Han, Sachin Kumar and Yulia Tsvetkov   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Despite the growing success of diffusion models in continuous-valued domains (e.g., images), similar efforts for discrete domains such as text have yet to match the performance of autoregressive language models. In this work, we present SSD-LM—a diffusion-based language model with two key design choices. First, SSD-LM is semi-autoregressive, iteratively generating blocks of text, allowing for flexible output length at decoding time while enabling local bidirectional context updates. Second, it is simplex-based, performing diffusion on the natural vocabulary space rather than a learned latent space, allowing us to incorporate classifier guidance and modular control using off-the-shelf classifiers without any adaptation. We evaluate SSD-LM on unconstrained text generation benchmarks, and show that it matches or outperforms strong autoregressive GPT-2 models across standard quality and diversity metrics, while vastly outperforming diffusion-based baselines. On controlled text generation, SSD-LM also outperforms competitive baselines, with an extra advantage in modularity.

DIONYSUS: A Pre-trained Model for Low-Resource Dialogue Summarization
Yu Li, Baolin Peng, Pengcheng He, Michel Galley, Zhou Yu and Jianfeng Gao   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Dialogue summarization has recently garnered significant attention due to its wide range of applications. However, existing methods for summarizing dialogues have limitations because they do not take into account the inherent structure of dialogue and rely heavily on labeled data, which can lead to poor performance in new domains. In this work, we propose DIONYSUS (dynamic input optimization in pre-training for dialogue summarization), a pre-trained encoder-decoder model for summarizing dialogues in any new domain. To pre-train DIONYSUS, we create two pseudo summaries for each dialogue example: one from a fine-tuned summarization model and the other from important dialogue turns. We then choose one of these pseudo summaries based on information distribution differences in different types of dialogues. This selected pseudo summary serves as the objective for pre-training DIONYSUS using a self-supervised approach on a large dialogue corpus. Our experiments show that DIONYSUS outperforms existing methods on six datasets, as demonstrated by its ROUGE scores in zero-shot and few-shot settings

Generating EDU Extracts for Plan-Guided Summary Re-Ranking
Griffin Adams, Alex Fabbri, Faisal Ladhak, Noémie Elhadad and Kathleen McKeown 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Two-step approaches, in which summary candidates are generated-then-reranked to return a single summary, can improve ROUGE scores over the standard single-step approach. Yet, standard decoding methods (i.e., beam search, nucleus sampling, and diverse beam search) produce candidates with redundant, and often low quality, content. In this paper, we design a novel method to generate candidates for re-ranking that addresses these issues. We ground each candidate abstract on its own unique content plan and generate distinct plan-guided abstracts using a model's top beam. More concretely, a standard language model (a BART LM) auto-regressively generates elemental discourse unit (EDU) content plans with an extractive copy mechanism. The top K beams from the content plan generator are then used to guide a separate LM, which produces a single abstractive candidate for each distinct plan. We apply an existing re-ranker (BRIO) to abstractive candidates generated from our method, as well as baseline decoding methods. We show large relevance improvements over previously published methods on widely used single document news article corpora, with ROUGE-2 F1 gains of 0.88, 2.01, and 0.38 on CNN / Dailymail, NYT, and Xsum, respectively. A human evaluation on CNN / DM validates these results. Similarly, on 1k samples from CNN / DM, we show that prompting GPT-3 to follow EDU plans outperforms sampling-based methods by by 1.05 ROUGE-2 F1 points. Code to generate and realize plans is available at https://github.com/griff4692/edu-sum.

Improving the Robustness of Summarization Systems with Dual Augmentation
Xiuying Chen, Guodong Long, Chongyang Tao, Mingzhe Li, Xin Gao, Chengqi Zhang and Xiangliang Zhang   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
A robust summarization system should be able to capture the gist of the document, regardless of the specific word choices or noise in the input. In this work, we first explore the summarization models' robustness against perturbations including word-level synonym substitution and noise. To create semantic-consistent substitutes, we propose a SummAttacker, which is an efficient approach to generating adversarial samples based on pre-trained language models. Experimental results show that state-of-the-art summarization models have a significant decrease in performance on adversarial and noisy test sets. Next, we analyze the vulnerability of the summarization systems and explore improving the robustness by data augmentation. Specifically, the first vulnerability factor we found is the low diversity of the training inputs. Correspondingly, we expose the encoder to more diverse cases created by SummAttacker in the input space. The second factor is the vulnerability of the decoder, and we propose an augmentation in the latent space of the decoder to improve its robustness. Concretely, we create virtual cases by manifold softmixing two decoder hidden states of similar semantic meanings. Experimental results on Gigaword and CNN/DM datasets demonstrate that our approach achieves significant improvements over strong baselines and exhibits higher robustness on noisy, attacked, and clean datasets

Peek Across: Improving Multi-Document Modeling via Cross-Document Question-Answering
Avi Caciularu, Matthew Peters, Jacob Goldberger, Ido Dagan and Arman Cohan   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The integration of multi-document pre-training objectives into language models has resulted in remarkable improvements in multi-document downstream tasks. In this work, we propose extending this idea by pre-training a generic multi-document model from a novel cross-document question answering pre-training objective. To that end, given a set (or cluster) of topically-related documents, we systematically generate semantically-oriented questions from a salient sentence in one document and challenge the model, during pre-training, to answer these questions while "peeking" into other topically-related documents. In a similar manner, the model is also challenged to recover the sentence from which the question was generated, again while leveraging cross-document information. This novel multi-document QA formulation directs the model to better recover cross-text informational relations, and introduces a natural augmentation that artificially increases the pre-training data. Further, unlike prior multi-document models that focus on either classification or summarization tasks, our pre-training objective formulation enables the model to perform tasks that involve both short text generation (e.g., QA) and long text generation (e.g., summarization). Following this scheme, we pre-train our model - termed QAmden - and evaluate its performance across several multi-document tasks, including multi-document QA, summarization, and query-focused summarization, yielding improvements of up to 7%, and significantly outperforms zero-shot GPT-3.5 and GPT-4.

Toward Expanding the Scope of Radiology Report Summarization to Multiple Anatomies and Modalities
Zhihong Chen, Maya Varma, Xiang Wan, Curtis Langlotz and Jean-Benoit Delbrouck 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Radiology report summarization (RRS) is a growing area of research. Given the Findings section of a radiology report, the goal is to generate a summary (called an Impression section) that highlights the key observations and conclusions of the radiology study. However, RRS currently faces essential limitations. First, many prior studies conduct experiments on private datasets, preventing reproduction of results and fair comparisons across different systems and solutions. Second, most prior approaches are evaluated solely on chest X-rays. To address these limitations, we propose a dataset (MIMIC-RRS) involving three new modalities and seven new anatomies based on the MIMIC-III

and MIMIC-CXR datasets. We then conduct extensive experiments to evaluate the performance of models both within and across modality-anatomy pairs in MIMIC-RRS. In addition, we evaluate their clinical efficacy via RadGraph, a factual correctness metric.

Automated Metrics for Medical Multi-Document Summarization Disagree with Human Evaluations
Lucy Lu Wang, Yulia Otmakhova, Jay DeYoung, Thinh Hung Truong, Bailey E. Kuehl, Erin A Bransom and Byron C. Wallace   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Evaluating multi-document summarization (MDS) quality is difficult. This is especially true in the case of MDS for biomedical literature reviews, where models must synthesize contradicting evidence reported across different documents. Prior work has shown that rather than performing the task, models may exploit shortcuts that are difficult to detect using standard n-gram similarity metrics such as ROUGE. Better automated evaluation metrics are needed, but few resources exist to assess metrics when they are proposed. Therefore, we introduce a dataset of human-assessed summary quality facets and pairwise preferences to encourage and support the development of better automated evaluation methods for literature review MDS. We take advantage of community submissions to the Multi-document Summarization for Literature Review (MSLR) shared task to compile a diverse and representative sample of generated summaries. We analyze how automated summarization evaluation metrics correlate with lexical features of generated summaries, to other automated metrics including several we propose in this work, and to aspects of human-assessed summary quality. We find that not only do automated metrics fail to capture aspects of quality as assessed by humans, in many cases the system rankings produced by these metrics are anti-correlated with rankings according to human annotators.

Concise Answers to Complex Questions: Summarization of Long-form Answers
Abhilash C Potluri, Fangyuan Xu and Eunsol Choi   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Long-form question answering systems provide rich information by presenting paragraph-level answers, often containing optional background or auxiliary information. While such comprehensive answers are helpful, not all information is required to answer the question (e.g. users with domain knowledge do not need an explanation of background). Can we provide a concise version of the answer by summarizing it, while still addressing the question? We conduct a user study on summarized answers generated from state-of-the-art models and our newly proposed extract-and-decontextualize approach. We find a large proportion of long-form answers (over 90%) in the ELI5 domain can be adequately summarized by at least one system, while complex and implicit answers are challenging to compress. We observe that decontextualization improves the quality of the extractive summary, exemplifying its potential in the summarization task. To promote future work, we provide an extractive summarization dataset covering 1K long-form answers and our user study annotations. Together, we present the first study on summarizing long-form answers, taking a step forward for QA agents that can provide answers at multiple granularities.

Extractive is not Faithful: An Investigation of Broad Unfaithfulness Problems in Extractive Summarization
Shiyue Zhang, David Wan and Mohit Bansal   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The problems of unfaithful summaries have been widely discussed under the context of abstractive summarization. Though extractive summarization is less prone to the common unfaithfulness issues of abstractive summaries, does that mean extractive is equal to faithful? Turns out that the answer is no. In this work, we define a typology with five types of broad unfaithfulness problems (including and beyond not-entailment) that can appear in extractive summaries, including incorrect coreference, incomplete coreference, incorrect discourse, incomplete discourse, as well as other misleading information. We ask humans to label these problems out of 1600 English summaries produced by 16 diverse extractive systems. We find that 30% of the summaries have at least one of the five issues. To automatically detect these problems, we find that 5 existing faithfulness evaluation metrics for summarization have poor correlations with human judgment. To remedy this, we propose a new metric, ExtEval, that is designed for detecting unfaithful extractive summaries and is shown to have the best performance. We hope our work can increase the awareness of unfaithfulness problems in extractive summarization and help future work to evaluate and resolve these issues.

Finding the Pillars of Strength for Multi-Head Attention
Jinjie Ni, Rui Mao, Zonglin Yang, Han Lei and Erik Cambria   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Recent studies have revealed some issues of Multi-Head Attention (MHA), e.g., redundancy and over-parameterization. Specifically, the heads of MHA were originally designed to attend to information from different representation subspaces, whereas prior studies found that some attention heads likely learn similar features and can be pruned without harming performance. Inspired by the minimum-redundancy feature selection, we assume that focusing on the most representative and distinctive features with minimum resources can mitigate the above issues and lead to more effective and efficient MHAs. In particular, we propose Grouped Head Attention, trained with a self-supervised group constraint that group attention heads, where each group focuses on an essential but distinctive feature subset. We additionally propose a Voting-to-Stay procedure to remove redundant heads, thus achieving a transformer with lighter weights. Extensive experiments are consistent with our hypothesis. Moreover, our method achieves significant performance gains on three well-established tasks while considerably compressing parameters.

Better Simultaneous Translation with Monotonic Knowledge Distillation
Shushu Wang, Jing Wu, Kai Fan, Wei Luo, Jun Xiao and Zhongqiang Huang   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Simultaneous machine translation (SiMT) presents a unique challenge as it requires generating target tokens before the source sentence is fully consumed. This can lead to the hallucination problem, where target tokens are generated without support from the source sentence. The prefix-to-prefix training data used to train SiMT models are not always parallel, due to divergent word order between the source and target languages, and can contribute to the problem. In this paper, we propose a novel approach that leverages traditional translation models as teachers and employs a two-stage beam search algorithm to generate monotonic yet accurate reference translations for sequence-level knowledge distillation. Experimental results demonstrate the significant improvements achieved by our approach over multiple strong SiMT baselines, leading to new state-of-the-art performance across various language pairs. Notably, when evaluated on a monotonic version of the WMT15 De-En test set, which includes references generated in a more monotonic style by professional translators, our approach achieves even more substantial improvement over the baselines. The source code and data are publicly available for further exploration.

Continual Knowledge Distillation for Neural Machine Translation
Yuanchi Zhang, Peng Li, Maosong Sun and Yang Liu   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
While many parallel corpora are not publicly accessible for data copyright, data privacy and competitive differentiation reasons, trained translation models are increasingly available on open platforms. In this work, we propose a method called continual knowledge distillation to take advantage of existing translation models to improve one model of interest. The basic idea is to sequentially transfer knowledge from each trained model to the distilled model. Extensive experiments on Chinese-English and German-English datasets show that our method achieves significant and consistent improvements over strong baselines under both homogeneous and heterogeneous trained model settings and is robust to malicious models.

Causality-aware Concept Extraction based on Knowledge-guided Prompting
Siyu Yuan, Deqing Yang, Jinxi Liu, Shuyu Tian, Jiaqing Liang, Yanghua Xiao and Rui Xie   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Concepts benefit natural language understanding but are far from complete in existing knowledge graphs (KGs). Recently, pre-trained language models (PLMs) have been widely used in text-based concept extraction (CE). However, PLMs tend to mine the co-occurrence associations from massive corpus as pre-trained knowledge rather than the real causal effect between tokens. As a result, the pre-trained knowledge confounds PLMs to extract biased concepts based on spurious co-occurrence correlations, inevitably resulting in low precision. In this paper, through the lens of a Structural Causal Model (SCM), we propose equipping the PLM-based extractor with a knowledge-guided prompt as an intervention to alleviate concept bias. The prompt adopts the topic of the given entity from the existing knowledge in KGs to mitigate the spurious co-occurrence correlations between entities and biased concepts. Our extensive experiments on representative multilingual KG datasets justify that our proposed prompt can effectively alleviate concept bias and improve the performance of PLM-based CE models.

Information Screening whilst Exploiting! Multimodal Relation Extraction with Feature Denoising and Multimodal Topic Modeling

Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing and Tat-Seng Chua   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Existing research on multimodal relation extraction (MRE) faces two co-existing challenges, internal-information over-utilization and external-information under-exploitation. To combat that, we propose a novel framework that simultaneously implements the idea of internal-information screening and external-information exploiting. First, we represent the fine-grained semantic structures of the input image and text with the visual and textual scene graphs, which are further fused into a unified cross-modal graph (CMG). Based on CMG, we perform structure refinement with the guidance of the graph information bottleneck principle, actively denoising the less-informative features. Next, we perform topic modeling over the input image and text, incorporating latent multimodal topic features to enrich the contexts. On the benchmark MRE dataset, our system outperforms the current best model significantly. With further in-depth analyses, we reveal the great potential of our method for the MRE task.

DiffusionNER: Boundary Diffusion for Named Entity Recognition

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu and Yueting Zhuang 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

In this paper, we propose DiffusionNER, which formulates the named entity recognition task as a boundary-denoising diffusion process and thus generates named entities from noisy spans. During training, DiffusionNER gradually adds noises to the golden entity boundaries by a fixed forward diffusion process and learns a reverse diffusion process to recover the entity boundaries. In inference, DiffusionNER first randomly samples some noisy spans from a standard Gaussian distribution and then generates the named entities by denoising them with the learned reverse diffusion process. The proposed boundary-denoising diffusion process allows progressive refinement and dynamic sampling of entities, empowering DiffusionNER with efficient and flexible entity generation capability. Experiments on multiple flat and nested NER datasets demonstrate that DiffusionNER achieves comparable or even better performance than previous state-of-the-art models.

Modeling Instance Interactions for Joint Information Extraction with Neural High-Order Conditional Random Field

Zixia Jia, Zhaohui Yan, Wenjuan Han, Zilong Zheng and Kewei Tu   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Prior works on joint Information Extraction (IE) typically model instance (e.g., event triggers, entities, roles, relations) interactions by representation enhancement, type dependencies scoring, or global decoding. We find that the previous models generally consider binary type dependency scoring of a pair of instances, and leverage local search such as beam search to approximate global solutions. To better integrate cross-instance interactions, in this work, we introduce a joint IE framework (CRFIE) that formulates joint IE as a high-order Conditional Random Field. Specifically, we design binary factors and ternary factors to directly model interactions between not only a pair of instances but also triplets. Then, these factors are utilized to jointly predict labels of all instances.To address the intractability problem of exact high-order inference, we incorporate a high-order neural decoder that is unfolded from a mean-field variational inference method, which achieves consistent learning and inference. The experimental results show that our approach achieves consistent improvements on three IE tasks compared with our baseline and prior work.

Split-NER: Named Entity Recognition via Two Question-Answering-based Classifications

Jatin Arora and Youngja Park   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

In this work, we address the NER problem by splitting it into two logical sub-tasks: (1) Span Detection which simply extracts entity mention spans irrespective of entity type; (2) Span Classification which classifies the spans into their entity types. Further, we formulate both sub-tasks as question-answering (QA) problems and produce two leaner models which can be optimized separately for each sub-task. Experiments with four cross-domain datasets demonstrate that this two-step approach is both effective and time efficient. Our system, SplitNER outperforms baselines on OntoNotes5.0, WNUT17 and a cybersecurity dataset and gives on-par performance on BioNLP13CG. In all cases, it achieves a significant reduction in training time compared to its QA baseline counterpart. The effectiveness of our system stems from fine-tuning the BERT model twice, separately for span detection and classification. The source code can be found at https://github.com/c3sr/split-ner.

RE-Matching: A Fine-Grained Semantic Matching Method for Zero-Shot Relation Extraction

*Jun Zhao, WenYu Zhan, Xin Zhao, Qi Zhang, Tao Gui, Zhongyu Wei, Junzhe Wang, Minlong Peng and Mingming Sun*11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Semantic matching is a mainstream paradigm of zero-shot relation extraction, which matches a given input with a corresponding label description. The entities in the input should exactly match their hypernyms in the description, while the irrelevant contexts should be ignored when matching. However, general matching methods lack explicit modeling of the above matching pattern. In this work, we propose a fine-grained semantic matching method tailored for zero-shot relation extraction. Guided by the matching pattern, we decompose the sentence-level similarity score into the entity matching score and context matching score. Considering that not all contextual words contribute equally to the relation semantics, we design a context distillation module to reduce the negative impact of irrelevant components on context matching. Experimental results show that our method achieves higher matching accuracy and more than 10 times faster inference speed, compared with the state-of-the-art methods.

Few-Shot Document-Level Event Argument Extraction

Xianjun Yang, Yujie Lu and Linda R Petzold   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Event argument extraction (EAE) has been well studied at the sentence level but under-explored at the document level. In this paper, we study to capture event arguments that actually spread across sentences in documents. Prior works usually assume full access to rich document supervision, ignoring the fact that the available argument annotation is limited in production. To fill this gap, we present FewDocAE, a Few-Shot Document-Level Event Argument Extraction benchmark, based on the existing document-level event extraction dataset. We first define the new problem and reconstruct the corpus by a novel N-Way-D-Doc sampling instead of the traditional N-Way-K-Shot strategy. Then we adjust the current document-level neural models into the few-shot setting to provide baseline results under in- and cross-domain settings. Since the argument extraction depends on the context from multiple sentences and the learning process is limited to very few examples, we find this novel task to be very challenging with substantively low performance. Considering FewDocAE is closely related to practical use under low-resource regimes, we hope this benchmark encourages more research in this direction. Our data and codes will be available online.

Few-shot Event Detection: An Empirical Study and a Unified View

Yubo Ma, Zehao Wang, Yixin Cao and Aixin Sun   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Few-shot event detection (ED) has been widely studied, while this brings noticeable discrepancies, e.g., various motivations, tasks, and experimental settings, that hinder the understanding of models for future progress. This paper presents a thorough empirical study, a unified view of ED models, and a better unified baseline. For fair evaluation, we compare 12 representative methods on three datasets, which are roughly grouped into prompt-based and prototype-based models for detailed analysis. Experiments consistently demonstrate that prompt-based methods, including ChatGPT, still significantly trail prototype-based methods in terms of overall performance. To investigate their superior performance, we break down their design elements along several dimensions and build a unified framework on prototype-based methods. Under such unified view, each prototype-method can be viewed a combination of different modules from these design elements. We further combine all advantageous modules and propose a simple yet effective baseline, which outperforms existing methods by a large margin (e.g., 2.7% F1 gains under low-resource setting).

Simple Augmentations of Logical Rules for Neuro-Symbolic Knowledge Graph Completion
Ananjan Nandi, Navdeep Kaur, Parag Singla and Mausam -   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
High-quality and high-coverage rule sets are imperative to the success of Neuro-Symbolic Knowledge Graph Completion (NS-KGC) models, because they form the basis of all symbolic inferences. Recent literature builds neural models for generating rule sets, however, preliminary experiments show that they struggle with maintaining high coverage. In this work, we suggest three simple augmentations to existing rule sets: (1) transforming rules to their abductive forms, (2) generating equivalent rules that use inverse forms of constituent relations and (3) random walks that propose new rules. Finally, we prune potentially low quality rules. Experiments over four datasets and five ruleset-baseline settings suggest that these simple augmentations consistently improve results, and obtain up to 7.1 pt MRR and 8.5 pt Hits@1 gains over using rules without augmentations.

Continual Contrastive Finetuning Improves Low-Resource Relation Extraction
Wenxuan Zhou, Sheng Zhang, Tristan Naumann, Muhao Chen and Hoifung Poon   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Relation extraction (RE), which has relied on structurally annotated corpora for model training, has been particularly challenging in low-resource scenarios and domains. Recent literature has tackled low-resource RE by self-supervised learning, where the solution involves pretraining the entity pair embedding by RE-based objective and finetuning on labeled data by classification-based objective. However, a critical challenge to this approach is the gap in objectives, which prevents the RE model from fully utilizing the knowledge in pretrained representations. In this paper, we aim at bridging the gap and propose to pretrain and finetune the RE model using consistent objectives of contrastive learning. Since in this kind of representation learning paradigm, one relation may easily form multiple clusters in the representation space, we further propose a multi-center contrastive loss that allows one relation to form multiple clusters to better align with pretraining. Experiments on two document-level RE datasets, BioRED and Re-DocRED, demonstrate the effectiveness of our method. Particularly, when using 1% end-task training data, our method outperforms PLM-based RE classifier by 10.5% and 6.1% on the two datasets, respectively.

ORGAN: Observation-Guided Radiology Report Generation via Tree Reasoning
Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li and Jiang Liu   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
This paper explores the task of radiology report generation, which aims at generating free-text descriptions for a set of radiographs. One significant challenge of this task is how to correctly maintain the consistency between the images and the lengthy report. Previous research explored solving this issue through planning-based methods, which generate reports only based on high-level plans. However, these plans usually only contain the major observations from the radiographs (e.g., lung opacity), lacking much necessary information, such as the observation characteristics and preliminary clinical diagnoses. To address this problem, the system should also take the image information into account together with the textual plan and perform stronger reasoning during the generation process. In this paper, we propose an Observation-guided radiology Report Generation framework (ORGan). It first produces an observation plan and then feeds both the plan and radiographs for report generation, where an observation graph and a tree reasoning mechanism are adopted to precisely enrich the plan information by capturing the multi-formats of each observation. Experimental results demonstrate that our framework outperforms previous state-of-the-art methods regarding text quality and clinical efficacy.

Measuring Progress in Fine-grained Vision-and-Language Understanding
Emanuele Bugliarello, Laurent Sartran, Aishwarya Agrawal, Lisa Anne Hendricks and Aida Nematzadeh   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
While pretraining on large-scale image–text data from the Web has facilitated rapid progress on many vision-and-language (V&L) tasks, recent work has demonstrated that pretrained models lack "fine-grained" understanding, such as the ability to recognise relationships, verbs, and numbers in images. This has resulted in an increased interest in the community to either develop new benchmarks or models for such capabilities. To better understand and quantify progress in this direction, we investigate four competitive V&L models on four fine-grained benchmarks. Through our analysis, we find that X-VLM (Zeng et al., 2022) consistently outperforms other baselines, and that modelling innovations can impact performance more than scaling Web data, which even degrades performance sometimes. Through a deeper investigation of X-VLM, we highlight the importance of both novel losses and rich data sources for learning fine-grained skills. Finally, we inspect training dynamics, and discover that for some tasks, performance peaks early in training or significantly fluctuates, never converging.

UnitY: Two-pass Direct Speech-to-speech Translation with Discrete Units
Hirofumi Inaguma, Sravya Popuri, Ilia Kulikov, Peng-Jen Chen, Changhan Wang, Yu-An Chung, Yun Tang, Ann Lee, Shinji Watanabe and Juan Pino   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Direct speech-to-speech translation (S2ST), in which all components can be optimized jointly, is advantageous over cascaded approaches to achieve fast inference with a simplified pipeline. We present a novel two-pass direct S2ST architecture, UnitY, which first generates textual representations and predicts discrete acoustic units subsequently. We enhance the model performance by subword prediction in the first-pass decoder, advanced two-pass decoder architecture design and search strategy, and better training regularization. To leverage large amounts of unlabeled text data, we pre-train the first-pass text decoder based on the self-supervised denoising auto-encoding task. Experimental evaluations on benchmark datasets at various data scales demonstrate that UnitY outperforms a single-pass speech-to-unit translation model by 2.5-4.2 ASR-BLEU with 2.83x decoding speed-up. We show that the proposed methods boost the performance even when predicting spectrogram in the second pass. However, predicting discrete units achieves 2.51x decoding speed-up compared to that case.

Efficient Semiring-Weighted Earley Parsing
Andreas Opedal, Ran Zmigrod, Tim Vieira, Ryan Cotterell and Jason Eisner   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
We present Earley's (1970) context-free parsing algorithm as a deduction system, incorporating several known and new speed-ups. In particular, our presentation supports a known worst-case runtime improvement from Earley's (1970) $O(N^3|G||R|)$, which is unworkable for the large grammars that arise in natural language processing, to $O(N^3|G|)$, which matches the complexity of CKY on a binarized version of the grammar G. Here N is the length of the sentence, $|R|$ is the number of productions in G, and $|G|$ is the total length of those productions. We also provide a version that achieves runtime of $O(N^3|M|)$ with $|M| \leq |G|$ when the grammar is represented compactly as a single finite-state automaton M (this is partly novel). We carefully treat the generalization to semiring-weighted deduction, preprocessing the grammar like Stolcke (1995) to eliminate the possibility of deduction cycles, and further generalize Stolcke's method to compute the weights of sentence

prefixes. We also provide implementation details for efficient execution, ensuring that on a preprocessed grammar, the semiring-weighted versions of our methods have the same asymptotic runtime and space requirements as the unweighted methods, including sub-cubic runtime on some grammars.

Decoding Symbolism in Language Models
Meiqi Guo, Rebecca Hwa and Adriana Kovashka   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
This work explores the feasibility of eliciting knowledge from language models (LMs) to decode symbolism, recognizing something (e.g.,roses) as a stand-in for another (e.g., love). We present our evaluative framework, Symbolism Analysis (SymbA), which compares LMs (e.g., RoBERTa, GPT-J) on different types of symbolism and analyze the outcomes along multiple metrics. Our findings suggest that conventional symbols are more reliably elicited from LMs while situated symbols are more challenging. Results also reveal the negative impact of the bias in pre-trained corpora. We further demonstrate that a simple re-ranking strategy can mitigate the bias and significantly improve model performances to be on par with human performances in some cases.

Word sense extension
Lei Yu and Yang Xu   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Humans often make creative use of words to express novel senses. A long-standing effort in natural language processing has been focusing on word sense disambiguation (WSD), but little has been explored about how the sense inventory of a word may be extended toward novel meanings. We present a paradigm of word sense extension (WSE) that enables words to spawn new senses toward novel context. We develop a framework that simulates novel word sense extension by first partitioning a polysemous word type into two pseudo-tokens that mark its different senses, and then inferring whether the meaning of a pseudo-token can be extended to convey the sense denoted by the token partitioned from the same word type. Our framework combines cognitive models of chaining with a learning scheme that transforms a language model embedding space to support various types of word sense extension. We evaluate our framework against several competitive baselines and show that it is superior in predicting plausible novel senses for over 7,500 English words. Furthermore, we show that our WSE framework improves performance over a range of transformer-based WSD models in predicting rare word senses with few or zero mentions in the training data.

XL-LEXEME: WiC Pretrained Model for Cross-Lingual LEXical sEMantic changE
Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro and Pierpaolo Basile   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The recent introduction of large-scale datasets for the WiC (Word in Context) task enables the creation of more reliable and meaningful contextualized word embeddings. However, most of the approaches to the WiC task use cross-encoders, which prevent the possibility of deriving comparable word embeddings. In this work, we introduce XL-LEXEME, a Lexical Semantic Change Detection model. XL-LEXEME extends SBERT, highlighting the target word in the sentence. We evaluate XL-LEXEME on the multilingual benchmarks for SemEval-2020 Task 1 - Lexical Semantic Change (LSC) Detection and the RuShiftEval shared task involving five languages: English, German, Swedish, Latin, and Russian. XL-LEXEME outperforms the state-of-the-art in English, German and Swedish with statistically significant differences from the baseline results and obtains state-of-the-art performance in the RuShiftEval shared task.

Going Beyond Sentence Embeddings: A Token-Level Matching Algorithm for Calculating Semantic Textual Similarity
Hongwei Wang and Dong Yu   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Semantic Textual Similarity (STS) measures the degree to which the underlying semantics of paired sentences are equivalent. State-of-the-art methods for STS task use language models to encode sentences into embeddings. However, these embeddings are limited in representing semantics because they mix all the semantic information together in fixed-length vectors, which are difficult to recover and lack explainability. This paper presents a token-level matching inference algorithm, which can be applied on top of any language model to improve its performance on STS task. Our method calculates pairwise token-level similarity and token matching scores, and then aggregates them with pretrained token weights to produce sentence similarity. Experimental results on seven STS datasets show that our method improves the performance of almost all language models, with up to 12.7% gain in Spearman's correlation. We also demonstrate that our method is highly explainable and computationally efficient.

APOLLO: A Simple Approach for Adaptive Pretraining of Language Models for Logical Reasoning
Soumya Sanyal, Yichong Xu, Shuohang Wang, Ziyi Yang, Reid Pryzant, Wenhao Yu, Chenguang Zhu and Xiang Ren   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Logical reasoning over text is an important ability that requires understanding the semantics of the text and reasoning through them to arrive at correct inferences. Prior works on pretraining language models to improve the logical reasoning ability require complex processing of training data (e.g., aligning symbolic knowledge to text), yielding task-specific data augmentation that is not easy to adapt to any general text corpus. In this work, we propose APOLLO, a simple adaptive pretraining approach to improve the logical reasoning skills of language models. We select a subset of Wikipedia for adaptive pretraining using a set of logical inference keywords as filter words. Further, we propose two self-supervised loss functions for training. First, we modify the masked language modeling loss only to mask specific parts-of-speech words that likely require higher-order reasoning to predict them. Second, we propose a sentence-level classification loss that teaches the model to distinguish between entailment and contradiction types of sentences. The proposed pretraining paradigm is both simple and independent of task formats. We demonstrate the effectiveness of APOLLO by comparing it with prior baselines on two logical reasoning datasets. APOLLO performs comparably on ReClor and outperforms baselines on LogiQA.

Evaluating Paraphrastic Robustness in Textual Entailment Models
Dhruv Verma, Yash Kumar Lal, Shreyashee Sinha, Benjamin Van Durme and Adam Poliak   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
We present PaRTE, a collection of 1,126 pairs of Recognizing Textual Entailment (RTE) examples to evaluate whether models are robust to paraphrasing. We posit that if RTE models understand language, their predictions should be consistent across inputs that share the same meaning. We use the evaluation set to determine if RTE models' predictions change when examples are paraphrased. In our experiments, contemporary models change their predictions on 8-16% of paraphrased examples, indicating that there is still room for improvement.

Label-Aware Hyperbolic Embeddings for Fine-grained Emotion Classification
Chih Yao Chen, Tun Min Hung, Yi-Li Hsu and Lun-Wei Ku   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Fine-grained emotion classification (FEC) is a challenging task. Specifically, FEC needs to handle subtle nuance between labels, which can be complex and confusing. Most existing models only address text classification problem in the euclidean space, which we believe may not be the optimal solution as labels of close semantic (e.g., afraid and terrified) may not be differentiated in such space, which harms the performance. In this paper, we propose HypEmo, a novel framework that can integrate hyperbolic embeddings to improve the FEC task. First, we learn label embeddings in the hyperbolic space to better capture their hierarchical structure, and then our model projects contextualized representations to the hyperbolic space to compute the distance between samples and labels. Experimental results show that incorporating such distance to weight cross entropy loss substantially improve the performance on two benchmark datasets, with around 3% improvement

compared to previous state-of-the-art, and could even improve up to 8.6% when the labels are hard to distinguish. Code is available at https://github.com/dinobby/HypEmo.

Learning Action Conditions from Instructional Manuals for Instruction Understanding

Te-Lin Wu, Caiqi ZHANG, Qingyuan Hu, Alexander Spangher and Nanyun Peng   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The ability to infer pre- and postconditions of an action is vital for comprehending complex instructions, and is essential for applications such as autonomous instruction-guided agents and assistive AI that supports humans to perform physical tasks. In this work, we propose a task dubbed action condition inference, which extracts mentions of preconditions and postconditions of actions in instructional manuals. We propose a weakly supervised approach utilizing automatically constructed large-scale training instances from online instructions, and curate a densely human-annotated and validated dataset to study how well the current NLP models do on the proposed task. We design two types of models differ by whether contextualized and global information is leveraged, as well as various combinations of heuristics to construct the weak supervisions. Our experiments show a > 20% F1-score improvement with considering the entire instruction contexts and a > 6% F1-score benefit with the proposed heuristics. However, the best performing model is still well-behind human performance.

A fine-grained comparison of pragmatic language understanding in humans and language models

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko and Edward Gibson 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Pragmatics and non-literal language understanding are essential to human communication, and present a long-standing challenge for artificial language models. We perform a fine-grained comparison of language models and humans on seven pragmatic phenomena, using zero-shot prompting on an expert-curated set of English materials. We ask whether models (1) select pragmatic interpretations of speaker utterances, (2) make similar error patterns as humans, and (3) use similar linguistic cues as humans to solve the tasks. We find that the largest models achieve high accuracy and match human error patterns: within incorrect responses, models favor literal interpretations over heuristic-based distractors. We also find preliminary evidence that models and humans are sensitive to similar linguistic cues. Our results suggest that pragmatic behaviors can emerge in models without explicitly constructed representations of mental states. However, models tend to struggle with phenomena relying on social expectation violations.

Crosslingual Generalization through Multitask Finetuning

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, KHALID ALMUBARAK, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff and Colin Raffel   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Multitask prompted finetuning (MTF) has been shown to help large language models generalize to new tasks in a zero-shot setting, but so far explorations of MTF have focused on English data and models. We apply MTF to the pretrained multilingual BLOOM and mT5 model families to produce finetuned variants called BLOOMZ and mT0. We find finetuning large multilingual language models on English tasks with English prompts allows for task genrealization to non-English languages that appear only in the pretraining corpus. Finetuning on multilingual tasks with English prompts further improves performance on English and non-English tasks leading to various state-of-the-art zero-shot results. We also investigate finetuning on multilingual tasks with prompts that have been machine-translated from English to match the language of each dataset. We find training on these machine-translated prompts leads to better performance on human-written prompts in the respective languages. Surprisingly, we find models are capable of zero-shot generalization to tasks in languages they have never intentionally seen. We conjecture that the models are learning higher-level capabilities that are both task- and language-agnostic. In addition, we introduce xP3, a composite of supervised datasets in 46 languages with English and machine-translated prompts. Our code, datasets and models are freely available at https://github.com/ bigscience-workshop/xmtf.

mPMR: A Multilingual Pre-trained Machine Reader at Scale

Weiwen Xu, Xin Li, Wai Lam and Lidong Bing   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
We present multilingual Pre-trained Machine Reader (mPMR), a novel method for multilingual machine reading comprehension (MRC)-style pre-training. mPMR aims to guide multilingual pre-trained language models (mPLMs) to perform natural language understanding (NLU) including both sequence classification and span extraction in multiple languages. To achieve cross-lingual generalization when only source-language fine-tuning data is available, existing mPLMs solely transfer NLU capability from a source language to target languages. In contrast, mPMR allows the direct inheritance of multilingual NLU capability from the MRC-style pre-training to downstream tasks. Therefore, mPMR acquires better NLU capability for target languages. mPMR also provides a unified solver for tackling cross-lingual span extraction and sequence classification, thereby enabling the extraction of rationales to explain the sentence-pair classification process.

BLOOM+1: Adding Language Support to BLOOM for Zero-Shot Prompting

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, KHALID ALMUBARAK, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Indra Winata, Stella Biderman, Edward Raff, Dragomir Radev and Vassilina Nikoulina
11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The BLOOM model is a large publicly available multilingual language model, but its pretraining was limited to 46 languages. To extend the benefits of BLOOM to other languages without incurring prohibitively large costs, it is desirable to adapt BLOOM to new languages not seen during pretraining. In this work, we apply existing language adaptation strategies to BLOOM and benchmark its zero-shot prompting performance on eight new languages in a resource-constrained setting. We find language adaptation to be effective at improving zero-shot performance in new languages. Surprisingly, we find that adapter-based finetuning is more effective than continued pretraining for large models. In addition, we discover that prompting performance is not significantly affected by language specifics, such as the writing system. It is primarily determined by the size of the language adaptation data. We also add new languages to BLOOMZ, which is a multitask finetuned version of BLOOM capable of following task instructions zero-shot. We find including a new language in the multitask fine-tuning mixture to be the most effective method to teach BLOOMZ a new language. We conclude that with sufficient training data language adaptation can generalize well to diverse languages. Our code is available at https://github.com/bigscience-workshop/multilingual-modeling.

Beyond Contrastive Learning: A Variational Generative Model for Multilingual Retrieval

John Wieting, Jonathan Clark, William Cohen, Graham Neubig and Taylor Berg-Kirkpatrick 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Contrastive learning has been successfully used for retrieval of semantically aligned sentences, but it often requires large batch sizes or careful engineering to work well. In this paper, we instead propose a generative model for learning multilingual text embeddings which can be used to retrieve or score sentence pairs. Our model operates on parallel data in N languages and, through an approximation we introduce, efficiently encourages source separation in this multilingual setting, separating semantic information that is shared between translations from stylistic or language-specific variation. We show careful large-scale comparisons between contrastive and generation-based approaches for learning multilingual text embeddings, a comparison that has not been done to the best of our knowledge despite the popularity of these approaches. We evaluate this method on a suite of tasks including semantic similarity, bitext mining, and cross-lingual question retrieval - the last of which we introduce in this paper. Overall, our model outperforms both a strong contrastive and generative baseline on these tasks.

Soft Language Clustering for Multilingual Model Pre-training

Jiali Zeng, Yufan Jiang, Yongjing Yin, Yi Jing, Fandong Meng, Binghuai Lin, Yunbo Cao and Jie Zhou 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Multilingual pre-trained language models have demonstrated impressive (zero-shot) cross-lingual transfer abilities, however, their performance is hindered when the target language has distant typology from the source language or when pre-training data is limited in size. In this paper, we propose XLM-P, a method that contextually retrieves prompts as flexible guidance for encoding instances conditionally. Our space-efficient and model-agnostic XLM-P approach enables (1) lightweight modeling of language-invariant and language-specific knowledge across languages, and (2) easy integration with other multilingual pre-training methods. On the tasks of XTREME, which include text classification, sequence labeling, question answering, and sentence retrieval, both base- and large-size language models pre-trained with our proposed method exhibit consistent performance improvement. Furthermore, it provides substantial advantages for low-resource languages in unsupervised sentence retrieval and for target languages that differ greatly from the source language in cross-lingual transfer.

Analyzing Transformers in Embedding Space
Guy Dar, Mor Geva, Ankit Gupta and Jonathan Berant 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Understanding Transformer-based models has attracted significant attention, as they lie at the heart of recent technological advances across machine learning. While most interpretability methods rely on running models over inputs, recent work has shown that a zero-pass approach, where parameters are interpreted directly without a forward/backward pass, is feasible for some Transformer parameters, and for two-layer attention networks. In this work, we present a theoretical analysis where all parameters of a trained Transformer are interpreted by projecting them into the embedding space, that is, the space of vocabulary items they operate on. We derive a simple theoretical framework to support our arguments and provide ample evidence for its validity. First, an empirical analysis showing that parameters of both pretrained and fine-tuned models can be interpreted in embedding space. Second, we present two applications of our framework: (a) aligning the parameters of different models that share a vocabulary, and (b) constructing a classifier without training by "translating" the parameters of a fine-tuned classifier to parameters of a different model that was only pretrained. Overall, our findings open the door to interpretation methods that, at least in part, abstract away from model specifics and operate in the embedding space only.

Explaining How Transformers Use Context to Build Predictions
Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas and Marta R. Costa-jussà 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Language Generation Models produce words based on the previous context. Although existing methods offer input attributions as explanations for a model's prediction, it is still unclear how prior words affect the model's decision throughout the layers. In this work, we leverage recent advances in explainability of the Transformer and present a procedure to analyze models for language generation. Using contrastive examples, we compare the alignment of our explanations with evidence of the linguistic phenomena, and show that our method consistently aligns better than gradient-based and perturbation-based baselines. Then, we investigate the role of MLPs inside the Transformer and show that they learn features that help the model predict words that are grammatically acceptable. Lastly, we apply our method to Neural Machine Translation models, and demonstrate that they generate human-like source-target alignments for building predictions.

Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios
Jiaxuan Li, Lang Yu and Allyson Ettinger 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Current pre-trained language models have enabled remarkable improvements in downstream tasks, but it remains difficult to distinguish effects of statistical correlation from more systematic logical reasoning grounded on the understanding of real world. We tease these factors apart by leveraging counterfactual conditionals, which force language models to predict unusual consequences based on hypothetical propositions. We introduce a set of tests from psycholinguistic experiments, as well as larger-scale controlled datasets, to probe counterfactual predictions from five pre-trained language models. We find that models are consistently able to override real-world knowledge in counterfactual scenarios, and that this effect is more robust in case of stronger baseline world knowledge—however, we also find that for most models this effect appears largely to be driven by simple lexical cues. When we mitigate effects of both world knowledge and lexical cues to test knowledge of linguistic nuances of counterfactuals, we find that only GPT-3 shows sensitivity to these nuances, though this sensitivity is also non-trivially impacted by lexical associative factors.

FairPrism: Evaluating Fairness-Related Harms in Text Generation
Eve Fleisig, Aubrie N Amstutz, Chad Atalla, Su Lin Blodgett, Hal Daumé III, Alexandra Olteanu, Emily Sheng, Dan Vann and Hanna Wallach
11:00-12:30 (Frontenac Ballroom and Queen's Quay)
It is critical to measure and mitigate fairness-related harms caused by AI text generation systems, including stereotyping and demeaning harms. To that end, we introduce FairPrism, a dataset of 5,000 examples of AI-generated English text with detailed human annotations covering a diverse set of harms relating to gender and sexuality. FairPrism aims to address several limitations of existing datasets for measuring and mitigating fairness-related harms, including improved transparency, clearer specification of dataset coverage, and accounting for annotator disagreement and harms that are context-dependent. FairPrism's annotations include the extent of stereotyping and demeaning harms, the demographic groups targeted, and appropriateness for different applications. The annotations also include specific harms that occur in interactive contexts and harms that raise normative concerns when the "speaker" is an AI system. Due to its precision and granularity, FairPrism can be used to diagnose (1) the types of fairness-related harms that AI text generation systems cause, and (2) the potential limitations of mitigation methods, both of which we illustrate through case studies. Finally, the process we followed to develop FairPrism offers a recipe for building improved datasets for measuring and mitigating harms caused by AI systems.

From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models
Shangbin Feng, Chan Young Park, Yuhan Liu and Yulia Tsvetkov 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Language models (LMs) are pretrained on diverse data sources—news, discussion forums, books, online encyclopedias. A significant portion of this data includes facts and opinions which, on one hand, celebrate democracy and diversity of ideas, and on the other hand are inherently socially biased. Our work develops new methods to (1) measure media biases in LMs trained on such corpora, along social and economic axes, and (2) measure the fairness of downstream NLP models trained on top of politically biased LMs. We focus on hate speech and misinformation detection, aiming to empirically quantify the effects of political (social, economic) biases in pretraining data on the fairness of high-stakes social-oriented tasks. Our findings reveal that pretrained LMs do have political leanings which reinforce the polarization present in pretraining corpora, propagating social biases into hate speech predictions and media biases into misinformation detectors. We discuss the implications of our findings for NLP research and propose future directions to mitigate unfairness.

Linear Classifier: An Often-Forgotten Baseline for Text Classification
Yu-Chen Lin, Si-An Chen, Jie-Jyun Liu and Chih-Jen Lin 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Large-scale pre-trained language models such as BERT are popular solutions for text classification. Due to the superior performance of these advanced methods, nowadays, people often directly train them for a few epochs and deploy the obtained model. In this opinion paper, we point out that this way may only sometimes get satisfactory results. We argue the importance of running a simple baseline like linear classifiers on bag-of-words features along with advanced methods. First, for many text data, linear methods show competitive performance, high efficiency, and robustness. Second, advanced models such as BERT may only achieve the best results if properly applied. Simple baselines

help to confirm whether the results of advanced models are acceptable. Our experimental results fully support these points.

Rogue Scores
Max Grusky   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Correct, comparable, and reproducible model evaluation is essential for progress in machine learning. Over twenty years, thousands of language and vision models have been evaluated with a popular metric called ROUGE. Does this widespread benchmark metric meet these three evaluation criteria? This systematic review of over two thousand publications using ROUGE finds: (A) Critical evaluation decisions and parameters are routinely omitted, making most reported scores irreproducible. (B) Differences in evaluation protocol are common, affect scores, and impact the comparability of results reported in many papers. (C) Thousands of papers use nonstandard evaluation packages with software defects that produce provably incorrect scores. Estimating the overall impact of these findings is difficult: because software citations are rare, it is nearly impossible to distinguish between correct ROUGE scores and incorrect "rogue scores."

Do Question Answering Modeling Improvements Hold Across Benchmarks?
Nelson F. Liu, Tony Lee, Robin Jia and Percy Liang   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Do question answering (QA) modeling improvements (e.g., choice of architecture and training procedure) hold consistently across the diverse landscape of QA benchmarks? To study this question, we introduce the notion of concurrence—two benchmarks have high concurrence on a set of modeling approaches if they rank the modeling approaches similarly. We measure the concurrence between 32 QA benchmarks on a set of 20 diverse modeling approaches and find that human-constructed benchmarks have high concurrence amongst themselves, even if their passage and question distributions are very different. Surprisingly, even downsampled human-constructed benchmarks (i.e., collecting less data) and programmatically-generated benchmarks (e.g., cloze-formatted examples) have high concurrence with human-constructed benchmarks. These results indicate that, despite years of intense community focus on a small number of benchmarks, the modeling improvements studied hold broadly.

Benchmarking Large Language Model Capabilities for Conditional Generation
Joshua Maynez, Priyanka Agrawal and Sebastian Gehrmann   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Pre-trained large language models (PLMs) underly most new developments in natural language processing. They have shifted the field from application-specific model pipelines to a single model that is adapted to a wide range of tasks. Autoregressive PLMs like GPT-3 or PaLM and associated techniques like fewshot learning, have additionally shifted the output modality to generation instead of classification or regression. Despite their ubiquitous use, the generation quality of language models is rarely evaluated when these models are introduced. Additionally, it is unclear how existing generation tasks–while they can be used to compare systems at a high level–relate to the real world use cases for which people have been adopting them. In this work, we discuss how to adapt existing application-specific generation benchmarks to PLMs and provide an in-depth, empirical study of the limitations and capabilities of PLMs in natural language generation tasks along dimensions such as scale, architecture, input and output language. Our results show that PLMs differ in their applicability to different data regimes and their generalization to multiple languages. They further inform practitioners as to which PLMs to use for a given generation task setup. We share best practices to be taken into consideration when benchmarking generation capabilities during the development of upcoming PLMs.

NLP Reproducibility For All: Understanding Experiences of Beginners
Shane Storks, Keunwoo Peter Yu, Ziqiao Ma and Joyce Chai   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
As natural language processing (NLP) has recently seen an unprecedented level of excitement, and more people are eager to enter the field, it is unclear whether current research reproducibility efforts are sufficient for this group of beginners to apply the latest developments. To understand their needs, we conducted a study with 93 students in an introductory NLP course, where students reproduced the results of recent NLP papers. Surprisingly, we find that their programming skill and comprehension of research papers have a limited impact on their effort spent completing the exercise. Instead, we find accessibility efforts by research authors to be the key to success, including complete documentation, better coding practice, and easier access to data files. Going forward, we recommend that NLP researchers pay close attention to these simple aspects of open-sourcing their work, and use insights from beginners' feedback to provide actionable ideas on how to better support them.

Ethics and NLP

11:00-12:30 (Pier 2&3)

[TACL] Hate Speech Classifiers Learn Normative Social Stereotypes
Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy and Morteza Dehghani   11:00-11:15 (Pier 2&3)
Social stereotypes negatively impact individuals' judgements about different groups and may have a critical role in understanding language directed toward marginalized groups. Here, we assess the role of social stereotypes in the automated detection of hate speech in the English language by examining the impact of social stereotypes on annotation behaviors, annotated datasets, and hate speech classifiers. Specifically, we first investigate the impact of novice annotators' stereotypes on their hate-speech-annotation behavior. Then, we examine the effect of normative stereotypes in language on the aggregated annotators' judgements in a large annotated corpus. Finally, we demonstrate how normative stereotypes embedded in language resources are associated with systematic prediction errors in a hate-speech-classifier. The results demonstrate that hate-speech classifiers reflect social stereotypes against marginalized groups which can perpetuate social inequalities when propagated at scale. This framework, combining social-psychological and computational-linguistic methods, provides insights into sources of bias in hate-speech moderation, informing ongoing debates regarding machine learning fairness.

NLPositionality: Characterizing Design Biases of Datasets and Models
Sebastin Santy, Jenny T Liang, Ronan Le Bras, Katharina Reinecke and Maarten Sap   11:15-11:30 (Pier 2&3)
Design biases in NLP systems, such as performance differences for different populations, often stem from their creator's positionality, i.e., views and lived experiences shaped by identity and background. Despite the prevalence and risks of design biases, they are hard to quantify because researcher, system, and dataset positionality is often unobserved. We introduce NLPositionality, a framework for characterizing design biases and quantifying the positionality of NLP datasets and models. Our framework continuously collects annotations from a diverse pool of volunteer participants on LabintheWild, and statistically quantifies alignment with dataset labels and model predictions. We apply NLPositionality to existing datasets and models for two tasks—social acceptability and hate speech detection. To date, we have collected 16,299 annotations in over a year for 600 instances from 1,096 annotators across 87 countries. We find that datasets and models align predominantly with Western, White, college-educated, and younger populations. Additionally, certain groups, such as non-binary people and non-native English speakers, are further marginalized by datasets and models as they rank least in alignment across all tasks. Finally, we draw from prior literature to discuss how researchers can examine their own positionality and that of their datasets and models, opening the door for more inclusive NLP systems.

What social attitudes about gender does BERT encode? Leveraging insights from psycholinguistics
Julia Watson, Barend Beekhuizen and Suzanne Stevenson   11:30-11:45 (Pier 2&3)
Much research has sought to evaluate the degree to which large language models reflect social biases. We complement such work with an approach to elucidating the connections between language model predictions and people's social attitudes. We show how word preferences in a large language model reflect social attitudes about gender, using two datasets from human experiments that found differences in gendered or gender neutral word choices by participants with differing views on gender (progressive, moderate, or conservative). We find that the language model BERT takes into account factors that shape human lexical choice of such language, but may not weigh those factors in the same way people do. Moreover, we show that BERT's predictions most resemble responses from participants with moderate to conservative views on gender. Such findings illuminate how a language model: (1) may differ from people in how it deploys words that signal gender, and (2) may prioritize some social attitudes over others.

The Elephant in the Room: Analyzing the Presence of Big Tech in Natural Language Processing Research
Mohamed Abdalla, Jan Philip Wahle, Terry Lima Ruas, Aurélie Névéol, Fanny Ducel, Saif M. Mohammad and Karen Fort 11:45-12:00 (Pier 2&3)
Recent advances in deep learning methods for natural language processing (NLP) have created new business opportunities and made NLP research critical for industry development. As one of the big players in the field of NLP, together with governments and universities, it is important to track the influence of industry on research. In this study, we seek to quantify and characterize industry presence in the NLP community over time. Using a corpus with comprehensive metadata of 78,187 NLP publications and 701 resumes of NLP publication authors, we explore the industry presence in the field since the early 90s. We find that industry presence among NLP authors has been steady before a steep increase over the past five years (180% growth from 2017 to 2022). A few companies account for most of the publications and provide funding to academic researchers through grants and internships. Our study shows that the presence and impact of the industry on natural language processing research are significant and fast-growing. This work calls for increased transparency of industry influence in the field.

WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models
Virginia K. Felkner, Ho-Chun Herbert Chang, Eugene Jang and Jonathan May   12:00-12:15 (Pier 2&3)
We present WinoQueer: a benchmark specifically designed to measure whether large language models (LLMs) encode biases that are harmful to the LGBTQ+ community. The benchmark is community-sourced, via application of a novel method that generates a bias benchmark from a community survey. We apply our benchmark to several popular LLMs and find that off-the-shelf models generally do exhibit considerable anti-queer bias. Finally, we show that LLM bias against a marginalized community can be somewhat mitigated by finetuning on data written about or by members of that community, and that social media text written by community members is more effective than news text written about the community by non-members. Our method for community-in-the-loop benchmark development provides a blueprint for future researchers to develop community-driven, harms-grounded LLM benchmarks for other marginalized communities.

ETHICIST: Targeted Training Data Extraction Through Loss Smoothed Soft Prompting and Calibrated Confidence Estimation
Zhexin Zhang, Jiaxin Wen and Minlie Huang   12:15-12:30 (Pier 2&3)
Large pre-trained language models achieve impressive results across many tasks. However, recent works point out that pre-trained language models may memorize a considerable fraction of their training data, leading to the privacy risk of information leakage. In this paper, we propose a method named Ethicist for targeted training data extraction through loss smoothed soft prompting and calibrated confidence estimation, investigating how to recover the suffix in the training data when given a prefix. To elicit memorization in the attacked model, we tune soft prompt embeddings while keeping the model fixed. We further propose a smoothing loss that smooths the loss distribution of the suffix tokens to make it easier to sample the correct suffix. In order to select the most probable suffix from a collection of sampled suffixes and estimate the prediction confidence, we propose a calibrated confidence estimation method, which normalizes the confidence of the generated suffixes with a local estimation. We show that Ethicist significantly improves the extraction performance on a recently proposed public benchmark. We also investigate several factors influencing the data extraction performance, including decoding strategy, model scale, prefix length, and suffix length. Our code is availabel at https://github.com/thu-coai/Targeted-Data-Extraction.

Multilingualism and Cross-Lingual NLP

11:00-12:30 (Pier 4&5)

Improving the Detection of Multilingual Online Attacks with Rich Social Media Data from Singapore
Janosch Haber, Bertie Vidgen, Matthew S Chapman, Vibhor Agarwal, Roy Ka-Wei Lee, Yong Keong Yap and Paul Röttger 11:00-11:15 (Pier 4&5)
Toxic content is a global problem, but most resources for detecting toxic content are in English. When datasets are created in other languages, they often focus exclusively on one language or dialect. In many cultural and geographical settings, however, it is common to code-mix languages, combining and interchanging them throughout conversations. To shine a light on this practice, and enable more research into code-mixed toxic content, we introduce SOA, a new multilingual dataset of online attacks. Using the multilingual city-state of Singapore as a starting point, we collect a large corpus of Reddit comments in Indonesian, Malay, Singlish, and other languages, and provide fine-grained hierarchical labels for online attacks. We publish the corpus with rich metadata, as well as additional unlabelled data for domain adaptation. We share comprehensive baseline results, show how the metadata can be used for granular error analysis, and demonstrate the benefits of domain adaptation for detecting multilingual online attacks.

MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages
Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur and Prem Natarajan   11:15-11:30 (Pier 4&5)
We present the MASSIVE dataset–Multilingual Amazon Slu resource package (SLURP) for Slot-filling, Intent classification, and Virtual assistant Evaluation. MASSIVE contains 1M realistic, parallel, labeled virtual assistant utterances spanning 51 languages, 18 domains, 60 intents, and 55 slots. MASSIVE was created by tasking professional translators to localize the English-only SLURP dataset into 50 typologically diverse languages from 29 genera. We also present modeling results on XLM-R and mT5, including exact match accuracy, intent classification accuracy, and slot-filling F1 score. We have released our dataset, modeling code, and models publicly.

Empowering Cross-lingual Behavioral Testing of NLP Models with Typological Features
Ester Hlavnova and Sebastian Ruder   11:30-11:45 (Pier 4&5)
A challenge towards developing NLP systems for the world's languages is understanding how they generalize to typological differences relevant for real-world applications. To this end, we propose M2C, a morphologically-aware framework for behavioral testing of NLP models. We use M2C to generate tests that probe models' behavior in light of specific linguistic features in 12 typologically diverse languages. We

evaluate state-of-the-art language models on the generated tests. While models excel at most tests in English, we highlight generalization failures to specific typological characteristics such as temporal expressions in Swahili and compounding possessives in Finish. Our findings motivate the development of models that address these blind spots.

[CL] Data-driven Cross-lingual Syntax: An Agreement Study with Massively Multilingual Models
Andrea Varda and Marco Marelli   11:45-12:00 (Pier 4&5)
Massively multilingual models such as mBERT and XLM-R are increasingly valued in Natural Language Processing research and applications, due to their ability to tackle the uneven distribution of resources available for different languages. The models' ability to process multiple languages relying on a shared set of parameters raises the question of whether the grammatical knowledge they extracted during pre-training can be considered as a data-driven cross-lingual grammar. The present work studies the inner workings of mBERT and XLM-R in order to test the cross-lingual consistency of the individual neural units that respond to a precise syntactic phenomenon, that is, number agreement, in five languages (English, German, French, Hebrew, Russian). We found that there is a significant overlap in the latent dimensions that encode agreement across the languages we considered. This overlap is larger (a) for long- vis-à-vis short-distance agreement and (b) when considering XLM-R as compared to mBERT, and peaks in the intermediate layers of the network. We further show that a small set of syntax-sensitive neurons can capture agreement violations across languages; however, their contribution is not decisive in agreement processing.

Towards Zero-Shot Multilingual Transfer for Code-Switched Responses
Ting-Wei Wu, Changsheng Zhao, Ernie Chang, Yangyang Shi, Pierce I-Jen Chuang, Vikas Chandra and Biing Juang 12:00-12:15 (Pier 4&5)
Recent task-oriented dialog systems have had great success in building English-based personal grammatists, but extending these systems to a global audience is challenging due to the need for annotated data in the target language. An alternative approach is to leverage existing data in a high-resource language to enable cross-lingual transfer in low-resource language models. However, this type of transfer has not been widely explored in natural language response generation. In this research, we investigate the use of state-of-the-art multilingual models such as mBART and T5 to facilitate zero-shot and few-shot transfer of code-switched responses. We propose a new adapter-based framework that allows for efficient transfer by learning task-specific representations and encapsulating source and target language representations. Our framework is able to successfully transfer language knowledge even when the target language corpus is limited. We present both quantitative and qualitative analyses to evaluate the effectiveness of our approach.

On Evaluating Multilingual Compositional Generalization with Translated Datasets
Zi Wang and Daniel Hershcovich   12:15-12:30 (Pier 4&5)
Compositional generalization allows efficient learning and human-like inductive biases. Since most research investigating compositional generalization in NLP is done on English, important questions remain underexplored. Do the necessary compositional generalization abilities differ across languages? Can models compositionally generalize cross-lingually? As a first step to answering these questions, recent work used neural machine translation to translate datasets for evaluating compositional generalization in semantic parsing. However, we show that this entails critical semantic distortion. To address this limitation, we craft a faithful rule-based translation of the MCWQ dataset from English to Chinese and Japanese. Even with the resulting robust benchmark, which we call MCWQ-R, we show that the distribution of compositions still suffers due to linguistic divergences, and that multilingual models still struggle with cross-lingual compositional generalization. Our dataset and methodology will serve as useful resources for the study of cross-lingual compositional generalization in other tasks.

Virtual Poster

11:00-12:30 (Pier 7&8)

[TACL] Minimum Description Length Recurrent Neural Networks
Nur Lan, Michal Geyer, Emmanuel Chemla and Roni Katzir   11:00-12:30 (Pier 7&8)
We train neural networks to optimize a Minimum Description Length score, i.e., to balance between the complexity of the network and its accuracy at a task. We show that networks optimizing this objective function master tasks involving memory challenges and go beyond context-free languages. These learners master languages such as $a^n b^n$, $a^n b^n c^n$, $a^n b^{2n}$, $a^n b^m c^{n+m}$, and they perform addition. Moreover, they often do so with 100

[TACL] Directed Acyclic Transformer Pre-training for High-quality Non-autoregressive Text Generation
Fei Huang, Pei Ke and Minlie Huang   11:00-12:30 (Pier 7&8)
Non-AutoRegressive (NAR) text generation models have drawn much attention because of their significantly faster decoding speed and good generation quality in machine translation. However, in a wider range of text generation tasks, existing NAR models lack proper pre-training, making them still far behind the pre-trained autoregressive models. In this paper, we propose Pre-trained Directed Acyclic Transformer (Pre-DAT) and a novel pre-training task to promote sentence-level prediction consistency in NAR generation. Experiments on five text generation tasks show that our PreDAT remarkably outperforms existing pre-trained NAR models (+4.2 scores on average) and even achieves better results than pre-trained autoregressive baselines in automatic evaluation, along with 17 times speedup in throughput. Further analysis shows that PreDAT benefits from the unbiased prediction order that alleviates the error accumulation problem in the autoregressive generation, which provides new insights into the advantages of NAR generation.

[SRW] ChatGPT vs Human-authored Text: Insights into Controllable Text Summarization and Sentence Style Transfer
Dongqi Pu and Vera Demberg   11:00-12:30 (Pier 7&8)
ChatGPT Analysis

[SRW] Prompt-based Zero-shot Text Classification with Conceptual Knowledge
Yuqi Wang, Wei Wang, Qi Chen, Kaizhu Huang, Anh Nguyen and Suparna De   11:00-12:30 (Pier 7&8)
The proposed framework incorporates conceptual knowledge for prompt-based text classification in the extreme zero-shot setting, which outperforms existing approaches in sentiment analysis and topic detection on four widely-used datasets.

[SRW] Improving Portfolio Management with Signals from Financial News
Zhilu Zhang   11:00-12:30 (Pier 7&8)

[SRW] Distractor Generation for Fill-in-the-Blank Exercises by Question Type
Nana Yoshimi, Tomoyuki Kajiwara, Satoru Uchida, Yuki Arase and Takashi Ninomiya   11:00-12:30 (Pier 7&8)
We define three types of questions (grammar, function word, and context) for fill-in-the-blank exercises and propose a method to generate

distractors according to the characteristics of each question type.

[SRW] "When Words Fail, Emojis Prevail": A Novel Architecture for Generating Sarcastic Sentences With Emoji Using Valence Reversal and Semantic Incongruity
Faria Binte Kader, Nafisa Hossain Nujat, Tasmia Binte Sogir, Mohsinul Kabir, Hasan Mahmud and Md Kamrul Hasan   11:00-12:30 (Pier 7&8)
A new framework in which when given a non-sarcastic text as input, the text is converted into a sarcastic one with emoji where the emoji will specifically help to identify the sarcastic intent of the text.

[Demo] OpenICL: An Open-Source Framework for In-context Learning
Yu Qiao, Jingjing Xu, Jiangtao Feng, Zhiyong Wu, Jiacheng Ye, Yaoxiang Wang and Zhenyu Wu   11:00-12:30 (Pier 7&8)
In recent years, In-context Learning (ICL) has gained increasing attention and emerged as the new paradigm for large language model (LLM) evaluation. Unlike traditional fine-tuning methods, ICL instead adapts the pre-trained models to unseen tasks without any parameter updates. However, the implementation of ICL is sophisticated due to the diverse retrieval and inference methods involved, as well as the varying pre-processing requirements for different models, datasets, and tasks. A unified and flexible framework for ICL is urgently needed to ease the implementation of the aforementioned components. To facilitate ICL research, we introduce OpenICL, an open-source toolkit for ICL and LLM evaluation. OpenICL is research-friendly with a highly flexible architecture that users can easily combine different components to suit their needs. It also provides various state-of-the-art retrieval and inference methods to streamline the process of adapting ICL to cutting-edge research. The effectiveness of OpenICL has been validated on a wide range of NLP tasks, including classification, QA, machine translation, and semantic parsing. As a side-product, we found OpenICL to be an efficient yet robust tool for LLMs evaluation. OpenICL is released at https://github.com/Shark-NLP/OpenICL.

[Demo] Human-in-the-loop Schema Induction
Chris Callison-Burch, Reece Suchocki, Susan Windisch Brown, Martha Palmer, Heng Ji, Sha Li, Rotem Dror, Lara Martin, Li Zhang, Hainiu Xu, Jiaxuan Ren, Zhaoyi Hou, Isaac Tham and Tianyi Zhang   11:00-12:30 (Pier 7&8)
Schema induction builds a graph representation explaining how events unfold in a scenario. Existing approaches have been based on information retrieval (IR) and information extraction (IE), often with limited human curation. We demonstrate a human-in-the-loop schema induction system powered by GPT-3. We first describe the different modules of our system, including prompting to generate schematic elements, manual edit of those elements, and conversion of those into a schema graph. By qualitatively comparing our system to previous ones, we show that our system not only transfers to new domains more easily than previous approaches, but also reduces efforts of human curation thanks to our interactive interface.

[Demo] LAVIS: A One-stop Library for Language-Vision Intelligence
Steven C.H. Hoi, silvio savarese, Guangsen Wang, Hung Le, Junnan Li and Dongxu Li   11:00-12:30 (Pier 7&8)
We introduce LAVIS, an open-source deep learning library for LAnguage-VISion research and applications. LAVIS aims to serve as a one-stop comprehensive library that brings recent advancements in the language-vision field accessible for researchers and practitioners, as well as fertilizing future research and development. It features a unified interface to easily access state-of-the-art image-language, video-language models and common datasets. LAVIS supports training, evaluation and benchmarking on a rich variety of tasks, including multimodal classification, retrieval, captioning, visual question answering, dialogue and pre-training. In the meantime, the library is also highly extensible and configurable, facilitating future development and customization. In this technical report, we describe design principles, key components and functionalities of the library, and also present benchmarking results across common language-vision tasks.

[Demo] TencentPretrain: A Scalable and Flexible Toolkit for Pre-training Models of Different Modalities
Kimmo Yan, Linlin Shen, Xiaoyong Du, Zhanhui Kang, Xingwu Sun, Xiaoshuai Chen, Feifei Li, Liqun Liu, Sihong Chen, Shan Huang, Chen Chen, Wenhang Shi, Tao Zhu, Taiqiang Wu, Weigang Gou, Han Guo, Weiquan Mao, Haoyan Liu, Ningyuan Sun, Yiren Chen, Weijie Liu, Rong Tian, Jing Zhao, Cheng Hou, Yudong Li and Zhe Zhao   11:00-12:30 (Pier 7&8)
Recently, the success of pre-training in text domain has been fully extended to vision, audio, and cross-modal scenarios. The proposed pre-training models of different modalities are showing a rising trend of homogeneity in their model structures, which brings the opportunity to implement different pre-training models within a uniform framework. In this paper, we present TencentPretrain, a toolkit supporting pre-training models of different modalities. The core feature of TencentPretrain is the modular design. The toolkit uniformly divides pre-training models into 5 components: embedding, encoder, target embedding, decoder, and target. As almost all of common modules are provided in each component, users can choose the desired modules from different components to build a complete pre-training model. The modular design enables users to efficiently reproduce existing pre-training models or build brand-new one. We test the toolkit on text, vision, and audio benchmarks and show that it can match the performance of the original implementations.

[Demo] Autodive: An Integrated Onsite Scientific Literature Annotation Tool
Yuanchun Zhou, Wenjuan Cui, Dongze Song, Mengyi Huang, Ludi Wang and Yi Du   11:00-12:30 (Pier 7&8)
Scientific literature is always available in Adobe's Portable Document Format (PDF), which is friendly for scientists to read. Compared with raw text, annotating directly on PDF documents can greatly improve the labeling efficiency of scientists whose annotation costs are very high. In this paper, we present Autodive, an integrated onsite scientific literature annotation tool for natural scientists and Natural Language Processing (NLP) researchers. This tool provides six core functions of annotation that support the whole lifecycle of corpus generation including i)annotation project management, ii)resource management, iii)ontology management, iv)manual annotation, v)onsite auto annotation, and vi)annotation task statistic. Two experiments are carried out to verify efficiency of the presented tool. A live demo of Autodive is available at http://autodive.sciwiki.cn. The source code is available at https://github.com/Autodive.

PropSegmEnt: A Large-Scale Corpus for Proposition-Level Segmentation and Entailment Recognition
Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, Dan Roth and Tal Schuster   11:00-12:30 (Pier 7&8)
The widely studied task of Natural Language Inference (NLI) requires a system to recognize whether one piece of text is textually entailed by another, i.e. whether the entirety of its meaning can be inferred from the other. In current NLI datasets and models, textual entailment relations are typically defined on the sentence- or paragraph-level. However, even a simple sentence often contains multiple propositions, i.e. distinct units of meaning conveyed by the sentence. As these propositions can carry different truth values in the context of a given premise, we argue for the need to recognize the textual entailment relation of each proposition in a sentence individually. We propose PropSegmEnt, a corpus of over 45K propositions annotated by expert human raters. Our dataset structure resembles the tasks of (1) segmenting sentences within a document to the set of propositions, and (2) classifying the entailment relation of each proposition with respect to a different yet topically-aligned document, i.e. documents describing the same event or entity. We establish strong baselines for the segmentation and entailment tasks. Through case studies on summary hallucination detection and document-level NLI, we demonstrate that our conceptual framework is potentially useful for understanding and explaining the compositionality of NLI labels.

LiveChat: A Large-Scale Personalized Dialogue Dataset Automatically Constructed from Live Streaming

Jingsheng Gao, Yixin Lian, Ziyi Zhou, yuzhuo fu and Baoyuan Wang   11:00-12:30 (Pier 7&8)
Open-domain dialogue systems have made promising progress in recent years. While the state-of-the-art dialogue agents are built upon large-scale social media data and large pre-trained models, there is no guarantee these agents could also perform well in fast-growing scenarios, such as live streaming, due to the bounded transferability of pre-trained models and biased distributions of public datasets from Reddit and Weibo, etc. To improve the essential capability of responding and establish a benchmark in the live open-domain scenario, we introduce the LiveChat dataset, composed of 1.33 million real-life Chinese dialogues with almost 3800 average sessions across 351 personas and fine-grained profiles for each persona. LiveChat is automatically constructed by processing numerous live videos on the Internet and naturally falls within the scope of multi-party conversations, where the issues of Who says What to Whom should be considered. Therefore, we target two critical tasks of response modeling and addressee recognition and propose retrieval-based baselines grounded on advanced techniques. Experimental results have validated the positive effects of leveraging persona profiles and larger average sessions per persona. In addition, we also benchmark the transferability of advanced generation-based models on LiveChat and pose some future directions for current challenges.

OpenPI-C: A Better Benchmark and Stronger Baseline for Open-Vocabulary State Tracking
Xueqing Wu, Sha Li and Heng Ji   11:00-12:30 (Pier 7&8)
Open-vocabulary state tracking is a more practical version of state tracking that aims to track state changes of entities throughout a process without restricting the state space and entity space. OpenPI (Tandon et al., 2020) is to date the only dataset annotated for open-vocabulary state tracking. However, we identify issues with the dataset quality and evaluation metric. For the dataset, we categorize 3 types of problems on the procedure level, step level and state change level respectively, and build a clean dataset OpenPI-C using multiple rounds of human judgment. For the evaluation metric, we propose a cluster-based metric to fix the original metric's preference for repetition.

Model-wise, we enhance the seq2seq generation baseline by reinstating two key properties for state tracking: temporal dependency and entity awareness. The state of the world after an action is inherently dependent on the previous state. We model this dependency through a dynamic memory bank and allow the model to attend to the memory slots during decoding. On the other hand, the state of the world is naturally a union of the states of involved entities. Since the entities are unknown in the open-vocabulary setting, we propose a two-stage model that refines the state change prediction conditioned on entities predicted from the first stage. Empirical results show the effectiveness of our proposed model, especially on the cleaned dataset and the cluster-based metric. The code and data are released at https://github.com/shirley-wu/openpi-c

Table and Image Generation for Investigating Knowledge of Entities in Pre-trained Vision and Language Models
Hidetaka Kamigaito, Katsuhiko Hayashi and Taro Watanabe   11:00-12:30 (Pier 7&8)
In this paper, we propose a table and image generation task to verify how the knowledge about entities acquired from natural language is retained in Vision & Language (V & L) models. This task consists of two parts: the first is to generate a table containing knowledge about an entity and its related image, and the second is to generate an image from an entity with a caption and a table containing related knowledge of the entity. In both tasks, the model must know the entities used to perform the generation properly. We created the Wikipedia Table and Image Generation (WikiTIG) dataset from about 200,000 infoboxes in English Wikipedia articles to perform the proposed tasks. We evaluated the performance on the tasks with respect to the above research question using the V & L model OFA, which has achieved state-of-the-art results in multiple tasks. Experimental results show that OFA forgets part of its entity knowledge by pre-training as a complement to improve the performance of image related tasks.

We Understand Elliptical Sentences, and Language Models should Too: A New Dataset for Studying Ellipsis and its Interaction with Thematic Fit
Davide Testa, Emmanuele Chersoni and Alessandro Lenci   11:00-12:30 (Pier 7&8)
Ellipsis is a linguistic phenomenon characterized by the omission of one or more sentence elements. Solving such a linguistic construction is not a trivial issue in natural language processing since it involves the retrieval of non-overtly expressed verbal material, which might in turn require the model to integrate human-like syntactic and semantic knowledge. In this paper, we explored the issue of how the prototypicality of event participants affects the ability of Language Models (LMs) to handle elliptical sentences and to identify the omitted arguments at different degrees of thematic fit, ranging from highly typical participants to semantically anomalous ones. With this purpose in mind, we built ELLie, the first dataset composed entirely of utterances containing different types of elliptical constructions, and structurally suited for evaluating the effect of argument thematic fit in solving ellipsis and reconstructing the missing element. Our tests demonstrated that the probability scores assigned by the models are higher for typical events than for atypical and impossible ones in different elliptical contexts, confirming the influence of prototypicality of the event participants in interpreting such linguistic structures. Finally, we conducted a retrieval task of the elided verb in the sentence in which the low performance of LMs highlighted a considerable difficulty in reconstructing the correct event.

Measuring Consistency in Text-based Financial Forecasting Models
Linyi Yang, Yingpeng Ma and Yue Zhang   11:00-12:30 (Pier 7&8)
Financial forecasting has been an important and active area of machine learning research, as even the most modest advantages in predictive accuracy can be parlayed into significant financial gains. Recent advances in natural language processing (NLP) bring the opportunity to leverage textual data, such as earnings reports of publicly traded companies, to predict the return rate for an asset. However, when dealing with such a sensitive task, the consistency of models – their invariance under meaning-preserving alternations in input – is a crucial property for building user trust. Despite this, current methods for financial forecasting do not take consistency into consideration. To address this issue, we propose FinTrust, an evaluation tool that assesses logical consistency in financial text. Using FinTrust, we show that the consistency of state-of-the-art NLP models for financial forecasting is poor. Our analysis of the performance degradation caused by meaning-preserving alternations suggests that current text-based methods are not suitable for robustly predicting market information.

Evaluate AMR Graph Similarity via Self-supervised Learning
Ziyi Shou and Fangzhen Lin   11:00-12:30 (Pier 7&8)
In work on AMR (Abstract Meaning Representation), similarity metrics are crucial as they are used to evaluate AMR systems such as AMR parsers. Current AMR metrics are all based on nodes or triples matching without considering the entire structures of AMR graphs. To address this problem, and inspired by learned similarity evaluation on plain text, we propose AMRSim, an automatic AMR graph similarity evaluation metric. To overcome the high cost of collecting human-annotated data, AMRSim automatically generates silver AMR graphs and utilizes self-supervised learning methods. We evaluated AMRSim on various datasets and found that AMRSim significantly improves the correlations with human semantic scores and remains robust under diverse challenges. We also discuss how AMRSim can be extended to multilingual cases.

Echoes from Alexandria: A Large Resource for Multilingual Book Summarization
Alessandro Scirè, Simone Conia, Simone Ciciliano and Roberto Navigli   11:00-12:30 (Pier 7&8)
In recent years, research in text summarization has mainly focused on the news domain, where texts are typically short and have strong layout features. The task of full-book summarization presents additional challenges which are hard to tackle with current resources, due to their limited size and availability in English only. To overcome these limitations, we present "Echoes from Alexandria", or in shortened form, "Echoes", a large resource for multilingual book summarization. Echoes features three novel datasets: i) Echo-Wiki, for multilingual book summarization, ii) Echo-XSum, for extremely-compressive multilingual book summarization, and iii) Echo-FairySum, for extractive book

summarization. To the best of our knowledge, Echoes – with its thousands of books and summaries – is the largest resource, and the first to be multilingual, featuring 5 languages and 25 language pairs. In addition to Echoes, we also introduce a new extractive-then-abstractive baseline, and, supported by our experimental results and manual analysis of the summaries generated, we argue that this baseline is more suitable for book summarization than purely-abstractive approaches. We release our resource and software at https://github.com/Babelscape/echoes-from-alexandria in the hope of fostering innovative research in multilingual book summarization.

Varta: A Large-Scale Headline-Generation Dataset for Indic Languages
Rahul Aralikatte, Ziling Cheng, Sumanth Doddapaneni and Jackie Chi Kit Cheung   11:00-12:30 (Pier 7&8)
We present Varta, a large-scale multilingual dataset for headline generation in Indic languages. This dataset includes more than 41 million pairs of headlines and articles in 14 different Indic languages (and English), which come from a variety of high-quality news sources. To the best of our knowledge, this is the largest collection of curated news articles for Indic languages currently available. We use the collected data in a series of experiments to answer important questions related to Indic NLP and multilinguality research in general. We show that the dataset is challenging even for state-of-the-art abstractive models and that they perform only slightly better than extractive baselines. Owing to its size, we also show that the dataset can be used to pre-train strong language models that outperform competitive baselines in both NLU and NLG benchmarks.

InfoSync: Information Synchronization across Multilingual Semi-structured Tables
Siddharth Hemant Khincha, Chelsi Jain, Vivek Gupta, Tushar Kataria and Shuo Zhang   11:00-12:30 (Pier 7&8)
Information Synchronization of semi-structured data across languages is challenging. For example, Wikipedia tables in one language need to be synchronized with others. To address this problem, we introduce a new dataset InfoSync and a two-step method for tabular synchronization. InfoSync contains 100K entity-centric tables (Wikipedia Infoboxes) across 14 languages, of which a subset (3.5K pairs) are manually annotated. The proposed method includes 1) Information Alignment to map rows and 2) Information Update for updating missing/outdated information for aligned tables across multilingual tables. When evaluated on InfoSync, information alignment achieves an F1 score of 87.91 (en <-> non-en). To evaluate information updation, we perform human-assisted Wikipedia edits on Infoboxes for 532 table pairs. Our approach obtains an acceptance rate of 77.28% on Wikipedia, showing the effectiveness of the proposed method.

ISLTranslate: Dataset for Translating Indian Sign Language
Abhinav Joshi, Susmit Agrawal and Ashutosh Modi   11:00-12:30 (Pier 7&8)
Sign languages are the primary means of communication for many hard-of-hearing people worldwide. Recently, to bridge the communication gap between the hard-of-hearing community and the rest of the population, several sign language translation datasets have been proposed to enable the development of statistical sign language translation systems. However, there is a dearth of sign language resources for the Indian sign language. This resource paper introduces ISLTranslate, a translation dataset for continuous Indian Sign Language (ISL) consisting of 31k ISL-English sentence/phrase pairs. To the best of our knowledge, it is the largest translation dataset for continuous Indian Sign Language. We provide a detailed analysis of the dataset. To validate the performance of existing end-to-end Sign language to spoken language translation systems, we benchmark the created dataset with a transformer-based model for ISL translation.

CrossSum: Beyond English-Centric Cross-Lingual Summarization for 1,500+ Language Pairs
Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang and Rifat Shahriyar   11:00-12:30 (Pier 7&8)
We present CrossSum, a large-scale cross-lingual summarization dataset comprising 1.68 million article-summary samples in 1,500+ language pairs. We create CrossSum by aligning parallel articles written in different languages via cross-lingual retrieval from a multilingual abstractive summarization dataset and a controlled human evaluation to validate its quality. We propose a multistage data sampling algorithm to effectively train a cross-lingual summarization model capable of summarizing an article in any target language. We also introduce LaSE, an embedding-based metric for automatically evaluating model-generated summaries. LaSE is strongly correlated with ROUGE and, unlike ROUGE, can be reliably measured even in the absence of references in the target language. Performance on ROUGE and LaSE indicate that our proposed model consistently outperforms baseline models. To the best of our knowledge, CrossSum is the largest cross-lingual summarization dataset and the first ever that is not centered around English. We are releasing the dataset, training and evaluation scripts, and models to spur future research on cross-lingual summarization. The resources can be found at https://github.com/csebuetnlp/CrossSum

Open-WikiTable : Dataset for Open Domain Question Answering with Complex Reasoning over Table
Sunjun Kweon, Yeonsu Kwon, Seonhee Cho, Yohan Jo and Edward Choi   11:00-12:30 (Pier 7&8)
Despite recent interest in open domain question answering (ODQA) over tables, many studies still rely on datasets that are not truly optimal for the task with respect to utilizing structural nature of table. These datasets assume answers reside as a single cell value and do not necessitate exploring over multiple cells such as aggregation, comparison, and sorting. Thus, we release Open-WikiTable, the first ODQA dataset that requires complex reasoning over tables. Open-WikiTable is built upon WikiSQL and WikiTableQuestions to be applicable in the open-domain setting. As each question is coupled with both textual answers and SQL queries, Open-WikiTable opens up a wide range of possibilities for future research, as both reader and parser methods can be applied. The dataset is publicly available.

I run as fast as a rabbit, can you? A Multilingual Simile Dialogues Datasets
Longxuan Ma, Wei-Nan Zhang, Shuhan Zhou, churui sun, Changxin Ke and Ting Liu   11:00-12:30 (Pier 7&8)
A simile is a figure of speech that compares two different things (called the tenor and the vehicle) via shared properties. The tenor and the vehicle are usually connected with comparator words such as "like" or "as". The simile phenomena are unique and complex in a real-life dialogue scene where the tenor and the vehicle can be nested phrases or sentences, mentioned by different speakers, exist in different sentences, or occur in reversed order. However, the current simile research usually focuses on similes in a triplet tuple (tenor, property, vehicle) or a single sentence where the tenor and vehicle are usually entities or noun phrases, which could not reflect complex simile phenomena in real scenarios. In this paper, we propose a novel and high-quality multilingual simile dialogue (MSD) dataset to facilitate the study of complex simile phenomena. The MSD is the largest manually annotated simile data (∼21K) and it contains both English and Chinese data. Meanwhile, the MSD data can also be used on dialogue tasks to test the ability of dialogue systems when using similes. We design 3 simile tasks (recognition, interpretation, and generation) and 2 dialogue tasks (retrieval and generation) with MSD. For each task, we provide experimental results from strong pre-trained or state-of-the-art models. The experiments demonstrate the challenge of MSD and we will release the data/code on GitHub.

Structure-Aware Language Model Pretraining Improves Dense Retrieval on Structured Data
Xinze Li, Zhenghao Liu, Chenyan Xiong, Shi Yu, Yu Gu, Zhiyuan Liu and Ge Yu   11:00-12:30 (Pier 7&8)
This paper presents Structure Aware Dense Retrieval (SANTA) model, which encodes user queries and structured data in one universal embedding space for retrieving structured data. SANTA proposes two pretraining methods to make language models structure-aware and learn effective representations for structured data: 1) Structured Data Alignment, which utilizes the natural alignment relations between structured data and unstructured data for structure-aware pretraining. It contrastively trains language models to represent multi-modal text data and teaches models to distinguish matched structured data for unstructured texts. 2) Masked Entity Prediction, which designs an entity-oriented mask strategy and asks language models to fill in the masked entities. Our experiments show that SANTA achieves state-of-the-art on code search and product search and conducts convincing results in the zero-shot setting. SANTA learns tailored representations for multi-modal

text data by aligning structured and unstructured data pairs and capturing structural semantics by masking and predicting entities in the structured data. All codes are available at https://github.com/OpenMatch/OpenMatch.

Enhancing Hierarchical Text Classification through Knowledge Graph Integration
Ye Liu, Kai Zhang, Zhenya Huang, Kehang Wang, Yanghai Zhang, Qi Liu and Enhong Chen 11:00-12:30 (Pier 7&8)
Hierarchical Text Classification (HTC) is an essential and challenging subtask of multi-label text classification with a taxonomic hierarchy. Recent advances in deep learning and pre-trained language models have led to significant breakthroughs in the HTC problem. However, despite their effectiveness, these methods are often restricted by a lack of domain knowledge, which leads them to make mistakes in a variety of situations. Generally, when manually classifying a specific document to the taxonomic hierarchy, experts make inference based on their prior knowledge and experience. For machines to achieve this capability, we propose a novel Knowledge-enabled Hierarchical Text Classification model (K-HTC), which incorporates knowledge graphs into HTC. Specifically, K-HTC innovatively integrates knowledge into both the text representation and hierarchical label learning process, addressing the knowledge limitations of traditional methods. Additionally, a novel knowledge-aware contrastive learning strategy is proposed to further exploit the information inherent in the data. Extensive experiments on two publicly available HTC datasets show the efficacy of our proposed method, and indicate the necessity of incorporating knowledge graphs in HTC tasks.

Discrete Prompt Optimization via Constrained Generation for Zero-shot Re-ranker
Sukmin Cho, Soyeong Jeong, Jeong yeon Seo and Jong Park 11:00-12:30 (Pier 7&8)
Re-rankers, which order retrieved documents with respect to the relevance score on the given query, have gained attention for the information retrieval (IR) task. Rather than fine-tuning the pre-trained language model (PLM), the large-scale language model (LLM) is utilized as a zero-shot re-ranker with excellent results. While LLM is highly dependent on the prompts, the impact and the optimization of the prompts for the zero-shot re-ranker are not explored yet. Along with highlighting the impact of optimization on the zero-shot re-ranker, we propose a novel discrete prompt optimization method, Constrained Prompt generation (Co-Prompt), with the metric estimating the optimum for re-ranking. Co-Prompt guides the generated texts from PLM toward optimal prompts based on the metric without parameter update. The experimental results demonstrate that Co-Prompt leads to outstanding re-ranking performance against the baselines. Also, Co-Prompt generates more interpretable prompts for humans against other prompt optimization methods.

UniTRec: A Unified Text-to-Text Transformer and Joint Contrastive Learning Framework for Text-based Recommendation
Zhiming Mao, Huimin Wang, Yiming Du and Kam-Fai Wong 11:00-12:30 (Pier 7&8)
Prior study has shown that pretrained language models (PLM) can boost the performance of text-based recommendation. In contrast to previous works that either use PLM to encode user history as a whole input text, or impose an additional aggregation network to fuse multi-turn history representations, we propose a unified local- and global-attention Transformer encoder to better model two-level contexts of user history. Moreover, conditioned on user history encoded by Transformer encoders, our framework leverages Transformer decoders to estimate the language perplexity of candidate text items, which can serve as a straightforward yet significant contrastive signal for user-item text matching. Based on this, our framework, UniTRec, unifies the contrastive objectives of discriminative matching scores and candidate text perplexity to jointly enhance text-based recommendation. Extensive evaluation shows that UniTRec delivers SOTA performance on three text-based recommendation tasks.

Dynamic Structured Neural Topic Model with Self-Attention Mechanism
Nozomu Miyamoto, Masaru Isonuma, Sho Takase, Junichiro Mori and Ichiro Sakata 11:00-12:30 (Pier 7&8)
This study presents a dynamic structured neural topic model, which can handle the time-series development of topics while capturing the dependencies. Our model captures the topic branching and merging processes by modeling topic dependencies based on a self-attention mechanism. Additionally, we introduce citation regularization, which induces attention weights to represent citation relations by modeling text and citations jointly. Our model outperforms a prior dynamic embedded topic model regarding perplexity and coherence, while maintaining sufficient diversity across topics. Furthermore, we confirm that our model can potentially predict emerging topics from academic literature.

Boosting Zero-shot Cross-lingual Retrieval by Training on Artificially Code-Switched Data
Robert Litschko, Ekaterina Artemova and Barbara Plank 11:00-12:30 (Pier 7&8)
Transferring information retrieval (IR) models from a high-resource language (typically English) to other languages in a zero-shot fashion has become a widely adopted approach. In this work, we show that the effectiveness of zero-shot rankers diminishes when queries and documents are present in different languages. Motivated by this, we propose to train ranking models on artificially code-switched data instead, which we generate by utilizing bilingual lexicons. To this end, we experiment with lexicons induced from (1) cross-lingual word embeddings and (2) parallel Wikipedia page titles. We use the mMARCO dataset to extensively evaluate reranking models on 36 language pairs spanning Monolingual IR (MoIR), Cross-lingual IR (CLIR), and Multilingual IR (MLIR). Our results show that code-switching can yield consistent and substantial gains of 5.1 MRR@10 in CLIR and 3.9 MRR@10 in MLIR, while maintaining stable performance in MoIR. Encouragingly, the gains are especially pronounced for distant languages (up to 2x absolute gain). We further show that our approach is robust towards the ratio of code-switched tokens and also extends to unseen languages. Our results demonstrate that training on code-switched data is a cheap and effective way of generalizing zero-shot rankers for cross-lingual and multilingual retrieval.

Cross-lingual Science Journalism: Select, Simplify and Rewrite Summaries for Non-expert Readers
Mehwish Fatima and Michael Strube 11:00-12:30 (Pier 7&8)
Automating Cross-lingual Science Journalism (CSJ) aims to generate popular science summaries from English scientific texts for non-expert readers in their local language. We introduce CSJ as a downstream task of text simplification and cross-lingual scientific summarization to facilitate science journalists' work. We analyze the performance of possible existing solutions as baselines for the CSJ task. Based on these findings, we propose to combine the three components - SELECT, SIMPLIFY and REWRITE (SSR) to produce cross-lingual simplified science summaries for non-expert readers. Our empirical evaluation on the Wikipedia dataset shows that SSR significantly outperforms the baselines for the CSJ task and can serve as a strong baseline for future work. We also perform an ablation study investigating the impact of individual components of SSR. Further, we analyze the performance of SSR on a high-quality, real-world CSJ dataset with human evaluation and in-depth analysis, demonstrating the superior performance of SSR for CSJ.

A Two-Stage Decoder for Efficient ICD Coding
Thanh-Tung Nguyen, Viktor Schlegel, Abhinav Ramesh Kashyap and Stefan Winkler 11:00-12:30 (Pier 7&8)
Clinical notes in healthcare facilities are tagged with the International Classification of Diseases (ICD) code; a list of classification codes for medical diagnoses and procedures. ICD coding is a challenging multilabel text classification problem due to noisy clinical document inputs and long-tailed label distribution. Recent automated ICD coding efforts improve performance by encoding medical notes and codes with additional data and knowledge bases. However, most of them do not reflect how human coders generate the code: first, the coders select general code categories and then look for specific subcategories that are relevant to a patient's condition. Inspired by this, we propose a two-stage decoding mechanism to predict ICD codes. Our model uses the hierarchical properties of the codes to split the prediction into two steps: At first, we predict the parent code and then predict the child code based on the previous prediction. Experiments on the public MIMIC-III data

set have shown that our model performs well in single-model settings without external data or knowledge.

Score It All Together: A Multi-Task Learning Study on Automatic Scoring of Argumentative Essays

Yuning Ding, Marie Bexte and Andrea Horbach   11:00-12:30 (Pier 7&8)

When scoring argumentative essays in an educational context, not only the presence or absence of certain argumentative elements but also their quality is important. On the recently published student essay dataset PERSUADE, we first show that the automatic scoring of argument quality benefits from additional information about context, writing prompt and argument type. We then explore the different combinations of three tasks: automated span detection, type and quality prediction. Results show that a multi-task learning approach combining the three tasks outperforms sequential approaches that first learn to segment and then predict the quality/type of a segment.

Wukong-Reader: Multi-modal Pre-training for Fine-grained Visual Document Understanding

Haoli Bai, Zhiguang Liu, Xiaojun Meng, li wentao, Shuang Liu, Yifeng LUO, nian xie, Rongfu Zheng, Liangwei Wang, Lu Hou, Jiansheng Wei, Xin Jiang and Qun Liu   11:00-12:30 (Pier 7&8)

Unsupervised pre-training on millions of digital-born or scanned documents has shown promising advances in visual document understanding (VDU). While various vision-language pre-training objectives are studied in existing solutions, the document textline, as an intrinsic granularity in VDU, has seldom been explored so far. A document textline usually contains words that are spatially and semantically correlated, which can be easily obtained from OCR engines. In this paper, we propose Wukong-Reader, trained with new pre-training objectives to leverage the structural knowledge nested in document textlines. We introduce textline-region contrastive learning to achieve fine-grained alignment between the visual regions and texts of document textlines. Furthermore, masked region modeling and textline-grid matching are also designed to enhance the visual and layout representations of textlines. Experiments show that Wukong-Reader brings superior performance on various VDU tasks in both English and Chinese. The fine-grained alignment over textlines also empowers Wukong-Reader with promising localization ability.

FEDLEGAL: The First Real-World Federated Learning Benchmark for Legal NLP

Zhuo Zhang, Xiangjing Hu, Jingyuan Zhang, Yating Zhang, Hui Wang, Lizhen Qu and Zenglin Xu   11:00-12:30 (Pier 7&8)

The inevitable private information in legal data necessitates legal artificial intelligence to study privacy-preserving and decentralized learning methods. Federated learning (FL) has merged as a promising technique for multiple participants to collaboratively train a shared model while efficiently protecting the sensitive data of participants. However, to the best of our knowledge, there is no work on applying FL to legal NLP. To fill this gap, this paper presents the first real-world FL benchmark for legal NLP, coined FEDLEGAL, which comprises five legal NLP tasks and one privacy task based on the data from Chinese courts. Based on the extensive experiments on these datasets, our results show that FL faces new challenges in terms of real-world non-IID data. The benchmark also encourages researchers to investigate privacy protection using real-world data in the FL setting, as well as deploying models in resource-constrained scenarios. The code and datasets of FEDLEGAL are available here.

Replace and Report: NLP Assisted Radiology Report Generation

Kaveri Kale, Pushpak Bhattacharyya and Kshitij Jadhav   11:00-12:30 (Pier 7&8)

Clinical practice frequently uses medical imaging for diagnosis and treatment. A significant challenge for automatic radiology report generation is that the radiology reports are long narratives consisting of multiple sentences for both abnormal and normal findings. Therefore, applying conventional image captioning approaches to generate the whole report proves to be insufficient, as these are designed to briefly describe images with short sentences. We propose a template-based approach to generate radiology reports from radiographs. Our approach involves the following: i) using a multilabel image classifier, produce the tags for the input radiograph; ii) using a transformer-based model, generate pathological descriptions (a description of abnormal findings seen on radiographs) from the tags generated in step (i); iii) using a BERT-based multi-label text classifier, find the spans in the normal report template to replace with the generated pathological descriptions; and iv) using a rule-based system, replace the identified span with the generated pathological description. We performed experiments with the two most popular radiology report datasets, IU Chest X-ray and MIMIC-CXR and demonstrated that the BLEU-1, ROUGE-L, METEOR, and CIDEr scores are better than the State-of-the-Art models by 25%, 36%, 44% and 48% respectively, on the IU X-RAY dataset. To the best of our knowledge, this is the first attempt to generate chest X-ray radiology reports by first creating small sentences for abnormal findings and then replacing them in the normal report template.

Similarity-Based Content Scoring - A more Classroom-Suitable Alternative to Instance-Based Scoring?

Marie Bexte, Andrea Horbach and Torsten Zesch   11:00-12:30 (Pier 7&8)

Automatically scoring student answers is an important task that is usually solved using instance-based supervised learning. Recently, similarity-based content scoring has been proposed as an alternative approach yielding similar perfor- mance. It has hypothetical advantages such as a lower need for annotated training data and better zero-shot performance, both of which are properties that would be highly beneficial when applying content scoring in a realistic classroom setting. In this paper we take a closer look at these alleged advantages by comparing different instance-based and similarity-based methods on multiple data sets in a number of learning curve experiments. We find that both the demand on data and cross-prompt performance is similar, thus not confirming the former two suggested advantages. The by default more straightforward possibility to give feedback based on a similarity-based approach may thus tip the scales in favor of it, although future work is needed to explore this advantage in practice.

The Mechanical Bard: An Interpretable Machine Learning Approach to Shakespearean Sonnet Generation

Edwin Agnew, Michelle Qiu, Lily Zhu, Sam Wiseman and Cynthia Rudin   11:00-12:30 (Pier 7&8)

We consider the automated generation of sonnets, a poetic form constrained according to meter, rhyme scheme, and length. Sonnets generally also use rhetorical figures, expressive language, and a consistent theme or narrative. Our constrained decoding approach allows for the generation of sonnets within preset poetic constraints, while using a relatively modest neural backbone. Human evaluation confirms that our approach produces Shakespearean sonnets that resemble human-authored sonnets, and which adhere to the genre's defined constraints and contain lyrical language and literary devices.

Contrastive Learning with Generated Representations for Inductive Knowledge Graph Embedding

Qian Li, Shafiq Joty, Daling Wang, Shi Feng, Yifei Zhang and Chengwei Qin   11:00-12:30 (Pier 7&8)

With the evolution of Knowledge Graphs (KGs), new entities emerge which are not seen before. Representation learning of KGs in such an inductive setting aims to capture and transfer the structural patterns from existing entities to new entities. However, the performance of existing methods in inductive KGs are limited by sparsity and implicit transfer. In this paper, we propose VMCL, a Contrastive Learning (CL) framework with graph guided Variational autoencoder on Meta-KGs in the inductive setting. We first propose representation generation to capture the encoded and generated representations of entities, where the generated variations can densify representations with complementary features. Then, we design two CL objectives that work across entities and meta-KGs to simulate the transfer mode. With extensive experiments we demonstrate that our proposed VMCL can significantly outperform previous state-of-the-art baselines.

Counterfactual Debiasing for Fact Verification

Weizhi Xu, Qiang Liu, Shu Wu and Liang Wang   11:00-12:30 (Pier 7&8)

Fact verification aims to automatically judge the veracity of a claim according to several pieces of evidence. Due to the manual construction of datasets, spurious correlations between claim patterns and its veracity (i.e., biases) inevitably exist. Recent studies show that models usually learn such biases instead of understanding the semantic relationship between the claim and evidence. Existing debiasing works can be roughly divided into data-augmentation-based and weight-regularization-based pipeline, where the former is inflexible and the latter relies on the uncertain output on the training stage. Unlike previous works, we propose a novel method from a counterfactual view, namely CLEVER, which is augmentation-free and mitigates biases on the inference stage. Specifically, we train a claim-evidence fusion model and a claim-only model independently. Then, we obtain the final prediction via subtracting output of the claim-only model from output of the claim-evidence fusion model, which counteracts biases in two outputs so that the unbiased part is highlighted. Comprehensive experiments on several datasets have demonstrated the effectiveness of CLEVER.

Hard Sample Aware Prompt-Tuning
Yuanjian Xu, Qi An, Jiahuan Zhang, Peng Li and Zaiqing Nie   11:00-12:30 (Pier 7&8)

Prompt-tuning based few-shot learning has garnered increasing attention in recent years due to its efficiency and promising capability. To achieve the best performance for NLP tasks with just a few samples, it is vital to include as many informative samples as possible and to avoid misleading ones. However, there is no work in prompt-tuning literature addressing the problem of differentiating informative hard samples from misleading ones in model training, which is challenging due to the lack of supervision signals about the quality of the samples to train a well-performed model. We propose a Hard Sample Aware Prompt-Tuning framework (i.e. HardPT) to solve the non-differentiable problem in hard sample identification with reinforcement learning, and to strengthen the discrimination of the feature space without changing the original data distribution via an adaptive contrastive learning method. An extensive empirical study on a series of NLP tasks demonstrates the capability of HardPT in few-shot scenarios. HardPT obtains new SOTA results on all evaluated NLP tasks, including pushing the SST-5 accuracy to 49.5% (1.1% point absolute improvement), QNLI accuracy to 74.6% (1.9% absolute improvement), NMLI accuracy to 71.5 (0.7% absolute improvement), TACREV F_1-score to 28.2 (1.0 absolute improvement), and i2b2/VA F_1-score to 41.2 (1.3 absolute improvement).

RMLM: A Flexible Defense Framework for Proactively Mitigating Word-level Adversarial Attacks
Zhaoyang Wang, Zhiyue Liu, Xiaopeng Zheng, Qinliang Su and Jiahai Wang   11:00-12:30 (Pier 7&8)

Adversarial attacks on deep neural networks keep raising security concerns in natural language processing research. Existing defenses focus on improving the robustness of the victim model in the training stage. However, they often neglect to proactively mitigate adversarial attacks during inference. Towards this overlooked aspect, we propose a defense framework that aims to mitigate attacks by confusing attackers and correcting adversarial contexts that are caused by malicious perturbations. Our framework comprises three components: (1) a synonym-based transformation to randomly corrupt adversarial contexts in the word level, (2) a developed BERT defender to correct abnormal contexts in the representation level, and (3) a simple detection method to filter out adversarial examples, any of which can be flexibly combined. Additionally, our framework helps improve the robustness of the victim model during training. Extensive experiments demonstrate the effectiveness of our framework in defending against word-level adversarial attacks.

Understanding Programs by Exploiting (Fuzzing) Test Cases
Jianyu Zhao, Yuyang Rong, Yiwen Guo, Yifeng He and Hao Chen   11:00-12:30 (Pier 7&8)

Semantic understanding of programs has attracted great attention in the community. Inspired by recent successes of large language models (LLMs) in natural language understanding, tremendous progress has been made by treating programming language as another sort of natural language and training LLMs on corpora of program code. However, programs are essentially different from texts after all, in a sense that they are normally heavily structured and syntax-strict. In particular, programs and their basic units (i.e., functions and subroutines) are designed to demonstrate a variety of behaviors and/or provide possible outputs, given different inputs. The relationship between inputs and possible outputs/behaviors represents the functions/subroutines and profiles the program as a whole. Hence, we propose to incorporate such a relationship into learning, for achieving a deeper semantic understanding of programs. To obtain inputs that are representative enough to trigger the execution of most part of the code, we resort to fuzz testing and propose fuzz tuning to boost the performance of program understanding and code representation learning, given a pre-trained LLM. The effectiveness of the proposed method is verified on two program understanding tasks including code clone detection and code classification, and it outperforms current state-of-the-arts by large margins. Code is available at https://github.com/rabbitjy/FuzzTuning.

Distantly Supervised Course Concept Extraction in MOOCs with Academic Discipline
Mengying Lu, Yuquan Wang, Jifan Yu, Yexing Du, Lei Hou and Juanzi Li   11:00-12:30 (Pier 7&8)

With the rapid growth of Massive Open Online Courses (MOOCs), it is expensive and time-consuming to extract high-quality knowledgeable concepts taught in the course by human effort to help learners grasp the essence of the course. In this paper, we propose to automatically extract course concepts using distant supervision to eliminate the heavy work of human annotations, which generates labels by matching them with an easily accessed dictionary. However, this matching process suffers from noise and incomplete annotations because of the limited dictionary and diverse MOOCs. To tackle these challenges, we present a novel three-stage framework DS-MOCE, which leverages the power of pre-trained language models explicitly and implicitly and employs discipline-embedding models with a self-train strategy based on label generation refinement across different domains. We also provide an expert-labeled dataset spanning 20 academic disciplines. Experimental results demonstrate the superiority of DS-MOCE over the state-of-the-art distantly supervised methods (with 7% absolute F1 score improvement). Code and data are now available at https://github.com/THU-KEG/MOOC-NER.

Dynamic Routing Transformer Network for Multimodal Sarcasm Detection
Yuan Tian, Nan Xu, Ruike Zhang and Wenji Mao   11:00-12:30 (Pier 7&8)

Multimodal sarcasm detection is an important research topic in natural language processing and multimedia computing, and benefits a wide range of applications in multiple domains. Most existing studies regard the incongruity between image and text as the indicative clue in identifying multimodal sarcasm. To capture cross-modal incongruity, previous methods rely on fixed architectures in network design, which restricts the model from dynamically adjusting to diverse image-text pairs. Inspired by routing-based dynamic network, we model the dynamic mechanism in multimodal sarcasm detection and propose the Dynamic Routing Transformer Network (DynRT-Net). Our method utilizes dynamic paths to activate different routing transformer modules with hierarchical co-attention adapting to cross-modal incongruity. Experimental results on a public dataset demonstrate the effectiveness of our method compared to the state-of-the-art methods. Our codes are available at https://github.com/TIAN-viola/DynRT.

TECHS: Temporal Logical Graph Networks for Explainable Extrapolation Reasoning
Qika Lin, Jun Liu, Rui Mao, Fangzhi Xu and Erik Cambria   11:00-12:30 (Pier 7&8)

Extrapolation reasoning on temporal knowledge graphs (TKGs) aims to forecast future facts based on past counterparts. There are two main challenges: (1) incorporating the complex information, including structural dependencies, temporal dynamics, and hidden logical rules; (2) implementing differentiable logical rule learning and reasoning for explainability. To this end, we propose an explainable extrapolation reasoning framework TEmporal logiCal grapH networkS (TECHS), which mainly contains a temporal graph encoder and a logical decoder. The former employs a graph convolutional network with temporal encoding and heterogeneous attention to embed topological structures and

temporal dynamics. The latter integrates propositional reasoning and first-order reasoning by introducing a reasoning graph that iteratively expands to find the answer. A forward message-passing mechanism is also proposed to update node representations, and their propositional and first-order attention scores. Experimental results demonstrate that it outperforms state-of-the-art baselines.

UniEvent: Unified Generative Model with Multi-Dimensional Prefix for Zero-Shot Event-Relational Reasoning

Zhengwei Tao, Zhi Jin, Haiyan Zhao, Chengfeng Dou, yongqiang zhao, Tao Shen and Chongyang Tao   11:00-12:30 (Pier 7&8)

Reasoning about events and their relations attracts surging research efforts since it is regarded as an indispensable ability to fulfill various event-centric or common-sense reasoning tasks. However, these tasks often suffer from limited data availability due to the labor-intensive nature of their annotations. Consequently, recent studies have explored knowledge transfer approaches within a multi-task learning framework to address this challenge. Although such methods have achieved acceptable results, such brute-force solutions struggle to effectively transfer event-relational knowledge due to the vast array of inter-event relations (e.g. temporal, causal, conditional) and reasoning formulations (e.g. discriminative, abductive, ending prediction). To enhance knowledge transfer and enable zero-shot generalization among various combinations, in this work we propose a novel unified framework, called UNIEVENT. Inspired by prefix-based multitask learning, our approach organizes event relational reasoning tasks into a coordinate system with multiple axes, representing inter-event relations and reasoning formulations. We then train a unified text-to-text generative model that utilizes coordinate-assigning prefixes for each task. By leveraging our adapted prefixes, our unified model achieves state-of-the-art or competitive performance on both zero-shot and supervised reasoning tasks, as demonstrated in extensive experiments

Learning Query Adaptive Anchor Representation for Inductive Relation Prediction

Zhiwen Xie, Yi Zhang, Jin Liu, Guangyou Zhou and Jimmy Xiangji Huang   11:00-12:30 (Pier 7&8)

Relation prediction on knowledge graphs (KGs) attempts to infer the missing links between entities. Most previous studies are limited to the transductive setting where all entities must be seen during the training, making them unable to perform reasoning on emerging entities. Recently, the inductive setting is proposed to handle the entities in the test phase to be unseen during training, However, it suffers from the inefficient reasoning under the enclosing subgraph extraction issue and the lack of effective entity-independent feature modeling. To this end, we propose a novel Query Adaptive Anchor Representation (QAAR) model for inductive relation prediction. First, we extract one opening subgraph and perform reasoning by one time for all candidate triples, which is more efficient when the number of candidate triples is large. Second, we define some query adaptive anchors which are independent on any specific entity. Based on these anchors, we take advantage of the transferable entity-independent features (relation-aware, structure-aware and distance features) that can be used to produce entity embeddings for emerging unseen entities. Such entity-independent features is modeled by a query-aware graph attention network on the opening subgraph. Experimental results demonstrate that our proposed QAAR outperforms state-of-the-art baselines in inductive relation prediction task.

ClaimDiff: Comparing and Contrasting Claims on Contentious Issues

Miyoung Ko, Ingyu Seong, Hwaran Lee, Joonsuk Park, Minsuk Chang and Minjoon Seo   11:00-12:30 (Pier 7&8)

With the growing importance of detecting misinformation, many studies have focused on verifying factual claims by retrieving evidence. However, canonical fact verification tasks do not apply to catching subtle differences in factually consistent claims, which might still bias the readers, especially on contentious political or economic issues. Our underlying assumption is that among the trusted sources, one's argument is not necessarily more true than the other, requiring comparison rather than verification. In this study, we propose ClaimDIff, a novel dataset that primarily focuses on comparing the nuance between claim pairs. In ClaimDiff, we provide human-labeled 2,941 claim pairs from 268 news articles. We observe that while humans are capable of detecting the nuances between claims, strong baselines struggle to detect them, showing over a 19% absolute gap with the humans. We hope this initial study could help readers to gain an unbiased grasp of contentious issues through machine-aided comparison.

Query Enhanced Knowledge-Intensive Conversation via Unsupervised Joint Modeling

Mingzhu Cai, Siqi Bao, Xin Tian, Huang He, Fan Wang and Hua Wu   11:00-12:30 (Pier 7&8)

In this paper, we propose an unsupervised query enhanced approach for knowledge-intensive conversations, namely QKConv. There are three modules in QKConv: a query generator, an off-the-shelf knowledge selector, and a response generator. QKConv is optimized through joint training, which produces the response by exploring multiple candidate queries and leveraging corresponding selected knowledge. The joint training solely relies on the dialogue context and target response, getting exempt from extra query annotations or knowledge provenances. To evaluate the effectiveness of the proposed QKConv, we conduct experiments on three representative knowledge-intensive conversation datasets: conversational question-answering, task-oriented dialogue, and knowledge-grounded conversation. Experimental results reveal that QKConv performs better than all unsupervised methods across three datasets and achieves competitive performance compared to supervised methods.

Boosting Distress Support Dialogue Responses with Motivational Interviewing Strategy

Anuradha Welivita and Pearl Pu   11:00-12:30 (Pier 7&8)

AI-driven chatbots have become an emerging solution to address psychological distress. Due to the lack of psychotherapeutic data, researchers use dialogues scraped from online peer support forums to train them. But since the responses in such platforms are not given by professionals, they contain both conforming and non-conforming responses. In this work, we attempt to recognize these conforming and non-conforming response types present in online distress-support dialogues using labels adapted from a well-established behavioral coding scheme named Motivational Interviewing Treatment Integrity (MITI) code and show how some response types could be rephrased into a more MI adherent form that can, in turn, enable chatbot responses to be more compliant with the MI strategy. As a proof of concept, we build several rephrasers by fine-tuning Blender and GPT3 to rephrase MI non-adherent Advise without permission responses into Advise with permission. We show how this can be achieved with the construction of pseudo-parallel corpora avoiding costs for human labor. Through automatic and human evaluation we show that in the presence of less training data, techniques such as prompting and data augmentation can be used to produce substantially good rephrasings that reflect the intended style and preserve the content of the original text.

Leveraging Explicit Procedural Instructions for Data-Efficient Action Prediction

Julia Isabel White, Arushi Raghuvanshi and Yada Pruksachatkun   11:00-12:30 (Pier 7&8)

Task-oriented dialogues often require agents to enact complex, multi-step procedures in order to meet user requests. While large language models have found success automating these dialogues in constrained environments, their widespread deployment is limited by the substantial quantities of task-specific data required for training. The following paper presents a data-efficient solution to constructing dialogue systems, leveraging explicit instructions derived from agent guidelines, such as company policies or customer service manuals. Our proposed Knowledge-Augmented Dialogue System (KADS) combines a large language model with a knowledge retrieval module that pulls documents outlining relevant procedures from a predefined set of policies, given a user-agent interaction. To train this system, we introduce a semi-supervised pre-training scheme that employs dialogue-document matching and action-oriented masked language modeling with partial parameter freezing. We evaluate the effectiveness of our approach on prominent task-oriented dialogue datasets, Action-Based Conversations Dataset and Schema-Guided Dialogue, for two dialogue tasks: action state tracking and workflow discovery. Our results demonstrate that procedural knowledge augmentation improves accuracy predicting in- and out-of-distribution actions while preserving high performance in

settings with low or sparse data.

CASE: Aligning Coarse-to-Fine Cognition and Affection for Empathetic Response Generation

Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang and Minlie Huang   11:00-12:30 (Pier 7&8)

Empathetic conversation is psychologically supposed to be the result of conscious alignment and interaction between the cognition and affection of empathy. However, existing empathetic dialogue models usually consider only the affective aspect or treat cognition and affection in isolation, which limits the capability of empathetic response generation. In this work, we propose the CASE model for empathetic dialogue generation. It first builds upon a commonsense cognition graph and an emotional concept graph and then aligns the user's cognition and affection at both the coarse-grained and fine-grained levels. Through automatic and manual evaluation, we demonstrate that CASE outperforms state-of-the-art baselines of empathetic dialogues and can generate more empathetic and informative responses.

End-to-End Task-Oriented Dialogue Systems Based on Schema

Wiradee Imrattanatrai and Ken Fukuda   11:00-12:30 (Pier 7&8)

This paper presents a schema-aware end-to-end neural network model for handling task-oriented dialogues based on a dynamic set of slots within a schema. Contrary to existing studies that proposed end-to-end approaches for task-oriented dialogue systems by relying on a unified schema across domains, we design our approach to support a domain covering multiple services where diverse schemas are available. To enable better generalizability among services and domains with different schemas, we supply the schema's context information including slot descriptions and value constraints to the model. The experimental results on a well-known Schema-Guided Dialogue (SGD) dataset demonstrated the performance improvement by the proposed model compared to state-of-the-art baselines in terms of end-to-end modeling, dialogue state tracking task, and generalization on new services and domains using a limited number of dialogues.

Ask an Expert: Leveraging Language Models to Improve Strategic Reasoning in Goal-Oriented Dialogue Models

Qiang Zhang, Jason Naradowsky and Yusuke Miyao   11:00-12:30 (Pier 7&8)

Existing dialogue models may encounter scenarios which are not well-represented in the training data, and as a result generate responses that are unnatural, inappropriate, or unhelpful. We propose the "Ask an Expert" framework in which the model is trained with access to an "expert" which it can consult at each turn. Advice is solicited via a structured dialogue with the expert, and the model is optimized to selectively utilize (or ignore) it given the context and dialogue history. In this work the expert takes the form of an LLM. We evaluate this framework in a mental health support domain, where the structure of the expert conversation is outlined by pre-specified prompts which reflect a reasoning strategy taught to practitioners in the field. Blenderbot models utilizing "Ask an Expert" show quality improvements across all expert sizes, including those with fewer parameters than the dialogue model itself. Our best model provides a 10% improvement over baselines, approaching human-level scores on "engingingness" and "helpfulness" metrics.

NewsDialogues: Towards Proactive News Grounded Conversation

Siheng Li, Yichun Yin, Cheng Yang, Wangjie Jiang, Yiwei Li, Zesen Cheng, Lifeng Shang, Xin Jiang, Qun Liu and Yujiu Yang   11:00-12:30 (Pier 7&8)

Hot news is one of the most popular topics in daily conversations. However, news grounded conversation has long been stymied by the lack of well-designed task definition and scarce data. In this paper, we propose a novel task, Proactive News Grounded Conversation, in which a dialogue system can proactively lead the conversation based on some key topics of the news. In addition, both information-seeking and chit-chat scenarios are included realistically, where the user may ask a series of questions about the news details or express their opinions and be eager to chat. To further develop this novel task, we collect a human-to-human Chinese dialogue dataset NewsDialogues, which includes 1K conversations with a total of 14.6K utterances and detailed annotations for target topics and knowledge spans. Furthermore, we propose a method named Predict-Generate-Rank, consisting of a generator for grounded knowledge prediction and response generation, and a ranker for the ranking of multiple responses to alleviate the exposure bias. We conduct comprehensive experiments to demonstrate the effectiveness of the proposed method and further present several key findings and challenges to prompt future research.

PaCE: Unified Multi-modal Dialogue Pre-training with Progressive and Compositional Experts

Yunshui Li, Binyuan Hui, ZhiChao Yin, Min Yang, Fei Huang and Yongbin Li   11:00-12:30 (Pier 7&8)

Perceiving multi-modal information and fulfilling dialogues with humans is a long-term goal of artificial intelligence. Multi-modal dialogue is commonly regarded as an effective approach for multi-modal dialogue. However, due to the limited availability of multi-modal dialogue data, there is still scarce research on multi-modal dialogue pre-training. Yet another intriguing challenge emerges from the encompassing nature of multi-modal dialogue, which involves various modalities and tasks. Moreover, new forms of tasks may arise at unpredictable points in the future. Hence, it is essential for designed multi-modal dialogue models to possess sufficient flexibility to adapt to such scenarios. This paper proposes PaCE, a unified, structured, compositional multi-modal dialogue pre-training framework. It utilizes a combination of several fundamental experts to accommodate multiple dialogue-related tasks and can be pre-trained using limited dialogue and extensive non-dialogue multi-modal data. Furthermore, we propose a progressive training method where old experts from the past can assist new experts, facilitating the expansion of their capabilities. Experimental results demonstrate that PaCE achieves state-of-the-art results on eight multi-modal dialog benchmarks.

Speech-Text Pre-training for Spoken Dialog Understanding with Explicit Cross-Modal Alignment

Tianshu Yu, haoyu gao, Ting-En Lin, Min Yang, Yuchuan Wu, Wentao Ma, chao wang, Fei Huang and Yongbin Li   11:00-12:30 (Pier 7&8)

Recently, speech-text pre-training methods have shown remarkable success in many speech and natural language processing tasks. However, most previous pre-trained models are usually tailored for one or two specific tasks, but fail to conquer a wide range of speech-text tasks. In addition, existing speech-text pre-training methods fail to explore the contextual information within a dialogue to enrich utterance representations. In this paper, we propose Speech-text Pre-training for spoken dialog understanding with ExpliCiT cRoss-Modal Alignment (SPECTRA), which is the first-ever speech-text dialog pre-training model. Concretely, to consider the temporality of speech modality, we design a novel temporal position prediction task to capture the speech-text alignment. This pre-training task aims to predict the start and end time of each textual word in the corresponding speech waveform. In addition, to learn the characteristics of spoken dialogs, we generalize a response selection task from textual dialog pre-training to speech-text dialog pre-training scenarios. Experimental results on four different downstream speech-text tasks demonstrate the superiority of SPECTRA in learning speech-text alignment and multi-turn dialog context.

Robust Learning for Multi-party Addressee Recognition with Discrete Addressee Codebook

Pengcheng Zhu, Wei Zhou, Kuncai Zhang, Yuankai Ma and Haiqing Chen   11:00-12:30 (Pier 7&8)

Addressee recognition aims to identify addressees in multi-party conversations. While state-of-the-art addressee recognition models have achieved promising performance, they still suffer from the issue of robustness when applied in real-world scenes. When exposed to a noisy environment, these models regard the noise as input and identify the addressee in a pre-given addressee closed set, while the addressees of the noise do not belong to this closed set, thus leading to the wrong identification of addressee. To this end, we propose a Robust Addressee Recognition (RAR) method, which discrete the addressees into a character codebook, making it able to represent open set addressees and robust in a noisy environment. Experimental results show that the introduction of the addressee character codebook helps to represent the open set addressees and highly improves the robustness of addressee recognition even if the input is noise.

CausalDialogue: Modeling Utterance-level Causality in Conversations

Yi-Lin Tuan, Alon Albalak, Wenda Xu, Michael S Saxon, Connor F Pryor, Lise Getoor and William Yang Wang   11:00-12:30 (Pier 7&8)

Despite their widespread adoption, neural conversation models have yet to exhibit natural chat capabilities with humans. In this research, we examine user utterances as causes and generated responses as effects, recognizing that changes in a cause should produce a different effect. To further explore this concept, we have compiled and expanded upon a new dataset called CausalDialogue through crowd-sourcing. This dataset includes multiple cause-effect pairs within a directed acyclic graph (DAG) structure. Our analysis reveals that traditional loss functions struggle to effectively incorporate the DAG structure, leading us to propose a causality-enhanced method called Exponential Maximum Average Treatment Effect (ExMATE) to enhance the impact of causality at the utterance level in training neural conversation models. To evaluate the needs of considering causality in dialogue generation, we built a comprehensive benchmark on CausalDialogue dataset using different models, inference, and training methods. Through experiments, we find that a causality-inspired loss like ExMATE can improve the diversity and agility of conventional loss function and there is still room for improvement to reach human-level quality on this new dataset.

Intent Discovery with Frame-guided Semantic Regularization and Augmentation

yajing sun, Rui Zhang, Jingyuan Yang and Wei Peng   11:00-12:30 (Pier 7&8)

Most existing intent discovery methods leverage representation learning and clustering to transfer the prior knowledge of known intents to unknown ones. The learned representations are limited to the syntactic forms of sentences, therefore, fall short of recognizing adequate variations under the same meaning of unknown intents. This paper proposes an approach utilizing frame knowledge as conceptual semantic guidance to bridge the gap between known intents representation learning and unknown intents clustering. Specifically, we employ semantic regularization to minimize the bidirectional KL divergence between model predictions for frame-based and sentence-based samples. Moreover, we construct a frame-guided data augmenter to capture intent-friendly semantic information and implement contrastive clustering learning for unsupervised sentence embedding. Extensive experiments on two benchmark datasets show that our method achieves substantial improvements in accuracy (5%+) compared to solid baselines.

DEnsity: Open-domain Dialogue Evaluation Metric using Density Estimation

ChaeHun Park, Seungil Chad Lee, Daniel Rim and Jaegul Choo   11:00-12:30 (Pier 7&8)

Despite the recent advances in open-domain dialogue systems, building a reliable evaluation metric is still a challenging problem. Recent studies proposed learnable metrics based on classification models trained to distinguish the correct response. However, neural classifiers are known to make overly confident predictions for examples from unseen distributions. We propose DENSITY, which evaluates a response by utilizing density estimation on the feature space derived from a neural classifier. Our metric measures how likely a response would appear in the distribution of human conversations. Moreover, to improve the performance of DENSITY, we utilize contrastive learning to further compress the feature space. Experiments on multiple response evaluation datasets show that DENSITY correlates better with human evaluations than the existing metrics.

Retrieval-free Knowledge Injection through Multi-Document Traversal for Dialogue Models

Rui Wang, Jianzhu Bao, Fei Mi, Yi Chen, Hongru Wang, Yasheng Wang, Yitong Li, Lifeng Shang, Kam-Fai Wong and Ruifeng Xu 11:00-12:30 (Pier 7&8)

Dialogue models are often enriched with extensive external knowledge to provide informative responses through a retrieval-augmented pipeline. Nevertheless, retrieval-augmented approaches rely on finely annotated retrieval training data and knowledge-grounded response generation data, making it costly to transfer. To tackle this challenge, this paper proposed a retrieval-free approach, KiDG, by automatically turning knowledge documents into simulated multi-turn dialogues through a Multi-Document Traversal algorithm. The simulated knowledge-intensive dialogues constructed by KiDG in one domain can be easily used to train and enhance pre-trained dialogue models' knowledge w.r.t. this domain without costly annotation. We conduct extensive experiments comparing retrieval-augmented models and a variety of retrieval-free models. We found that dialogue models enhanced with data simulated with KiDG largely outperform state-of-the-art retrieval-free methods, and it achieves comparable performance compared to retrieval-augmented methods while being better, and cheaper at domain transfer.

Multi3NLU++: A Multilingual, Multi-Intent, Multi-Domain Dataset for Natural Language Understanding in Task-Oriented Dialogue

Nikita Moghe, Evgeniia Razumovskaia, Liane K Guillou, Ivan Vulić, Anna Korhonen and Alexandra Birch   11:00-12:30 (Pier 7&8)

Task-oriented dialogue (ToD) systems have been widely deployed in many industries as they deliver more efficient customer support. These systems are typically constructed for a single domain or language and do not generalise well beyond this. To support work on Natural Language Understanding (NLU) in ToD across multiple languages and domains simultaneously, we constructed Multi3NLU++, a multilingual, multi-intent, multi-domain dataset. Multi3NLU++ extends the English-only NLU++ dataset to include manual translations into a range of high, medium, and low resource languages (Spanish, Marathi, Turkish and Amharic), in two domains (banking and hotels). Because of its multi-intent property, Multi3NLU++ represents complex and natural user goals, and therefore allows us to measure the realistic performance of ToD systems in a varied set of the world's languages. We use Multi3NLU++ to benchmark state-of-the-art multilingual models for the NLU tasks of intent detection and slot labeling for ToD systems in the multilingual setting. The results demonstrate the challenging nature of the dataset, particularly in the low-resource language setting, offering ample room for future experimentation in multi-domain multilingual ToD setups.

Prompted LLMs as Chatbot Modules for Long Open-domain Conversation

Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos and Kangwook Lee   11:00-12:30 (Pier 7&8)

In this paper, we propose MPC (Modular Prompted Chatbot), a new approach for creating high-quality conversational agents without the need for fine-tuning. Our method utilizes pre-trained large language models (LLMs) as individual modules for long-term consistency and flexibility, by using techniques such as few-shot prompting, chain-of-thought (CoT), and external memory. Our human evaluation results show that MPC is on par with fine-tuned chatbot models in open-domain conversations, making it an effective solution for creating consistent and engaging chatbots.

Zero-Shot Prompting for Implicit Intent Prediction and Recommendation with Commonsense Reasoning

Hui-Chi Kuo and Yun-Nung Chen   11:00-12:30 (Pier 7&8)

The current generation of intelligent assistants require explicit user requests to perform tasks or services, often leading to lengthy and complex conversations. In contrast, human assistants can infer multiple implicit intents from utterances via their commonsense knowledge, thereby simplifying interactions. To bridge this gap, this paper proposes a framework for multi-domain dialogue systems. This framework automatically infers implicit intents from user utterances, and prompts a large pre-trained language model to suggest suitable task-oriented bots. By leveraging commonsense knowledge, our framework recommends associated bots in a zero-shot manner, enhancing interaction efficiency and effectiveness. This approach substantially reduces interaction complexity, seamlessly integrates various domains and tasks, and represents a significant step towards creating more human-like intelligent assistants that can reason about implicit intents, offering a superior user experience.

CORE: Cooperative Training of Retriever-Reranker for Effective Dialogue Response Selection
Chongyang Tao, Jiazhan Feng, Tao Shen, Chang Liu, Juntao Li, Xiubo Geng and Daxin Jiang 11:00-12:30 (Pier 7&8)
Establishing retrieval-based dialogue systems that can select appropriate responses from the pre-built index has gained increasing attention. Recent common practice is to construct a two-stage pipeline with a fast retriever (e.g., bi-encoder) for first-stage recall followed by a smart response reranker (e.g., cross-encoder) for precise ranking. However, existing studies either optimize the retriever and reranker in independent ways, or distill the knowledge from a pre-trained reranker into the retriever in an asynchronous way, leading to sub-optimal performance of both modules. Thus, an open question remains about how to train them for a better combination of the best of both worlds. To this end, we present a cooperative training of the response retriever and the reranker whose parameters are dynamically optimized by the ground-truth labels as well as list-wise supervision signals from each other. As a result, the two modules can learn from each other and evolve throughout the training. Experimental results on two benchmarks demonstrate the superiority of our method.

Diverse Retrieval-Augmented In-Context Learning for Dialogue State Tracking
Brendan King and Jeffrey Flanigan 11:00-12:30 (Pier 7&8)
There has been significant interest in zero and few-shot learning for dialogue state tracking (DST) due to the high cost of collecting and anno- tating task-oriented dialogues. Recent work has demonstrated that in-context learning requires very little data and zero parameter updates, and even outperforms trained methods in the few-shot setting. We propose RefPyDST, which advances the state of the art with three advancements to in-context learning for DST. First, we formulate DST as a Python programming task, explicitly modeling language coreference as variable reference in Python. Second, since in-context learning depends highly on the context examples, we propose a method to retrieve a diverse set of relevant examples to improve performance. Finally, we introduce a novel re-weighting method during decoding that takes into account probabilities of competing surface forms, and produces a more accurate dialogue state prediction. We evaluate our approach using MultiWOZ and achieve state-of-the-art multi-domain joint-goal accuracy in zero and few-shot settings.

One Cannot Stand for Everyone! Leveraging Multiple User Simulators
to train Task-oriented Dialogue Systems
Yajiao LIU, Xin Jiang, Yichun Yin, Yasheng Wang, Fei Mi, Qun Liu, Xiang Wan and Benyou Wang 11:00-12:30 (Pier 7&8)
User simulators are agents designed to imitate human users; recent advances have found that Task-oriented Dialogue (ToD) systems optimized toward a user simulator could better satisfy the need of human users. However, this might result in a sub-optimal ToD system if it is tailored to only one *ad hoc* user simulator, since human users can behave differently. In this paper, we propose a framework called MUST to optimize ToD systems via leveraging Multiple User SimulaTors.
The main challenges of implementing MUST fall in 1) how to adaptively determine which user simulator to interact with the ToD system at each optimization step, since the ToD system might be over-fitted to some specific user simulators, and simultaneously under-fitted to some others; 2) how to avoid catastrophic forgetting of the adaption for a simulator that is not selected for several consecutive optimization steps. To tackle these challenges, we formulate MUST as a Multi-armed bandits (MAB) problem and provide a method called $MUST_{adaptive}$ that
balances *i*) the *boosting adaption* for adaptive interactions between different user simulators and the ToD system and *ii*) the *uniform adaption* to avoid the catastrophic forgetting issue. With both automatic evaluations and human evaluations, our experimental results on MultiWOZ show that the dialogue system trained by MUST achieves a better performance than those trained by a single user simulator. It also has a better generalization ability when testing with unseen user simulators.

Towards Fewer Hallucinations in Knowledge-Grounded Dialogue Generation via Augmentative and Contrastive Knowledge-Dialogue
Bin Sun, Yitong Li, Fei Mi, fanhu bie, Yiwei Li and Kan Li 11:00-12:30 (Pier 7&8)
Existing knowledge-grounded open-domain dialogue generation models often face the hallucination problem, i.e. the dialogue generative model will persist in an inappropriate knowledge and generate responses that inconsistent with the facts. We argue that this problem mainly stems from the polarized optimization objectives and weak knowledge generation ability. To mitigate the hallucination, we take inspiration from human communicating that people will replay euphemistic responses for the unclear or unrecognizable knowledge, and propose an Augmentative and Contrastive Knowledge Dialogue Expansion Framework (ACK-DEF). ACK-DEF constructs the augmentative and con- trastive knowledge dialogue samples, which consist of the knowledge of different degrees of errors and the response of manual design, to expand the original training set and smooth the polarized optimization objective that enables models to generate ground-truth with or without gold knowledge. Not only the knowledge, ACK-DEF also provides the tactful responses of manual design corresponding to the incomplete correct knowledge. Experimental results on the Wikipedia of Wizard dataset show that employing the ACK-DEF is effective to alleviate the hallucination problem.

EM Pre-training for Multi-party Dialogue Response Generation
Yiyang Li and Hai Zhao 11:00-12:30 (Pier 7&8)
Dialogue response generation requires an agent to generate a response according to the current dialogue history, in terms of which two-party dialogues have been well studied, but leaving a great gap for multi-party dialogues at the same time. Different from two-party dialogues where each response is a direct reply to its previous utterance, the addressee of a response utterance should be specified before it is generated in the multi-party scenario. Thanks to the huge amount of two-party conversational data, various pre-trained language models for two-party dialogue response generation have been proposed. However, due to the lack of annotated addressee labels in multi-party dialogue datasets, it is hard to use them to pre-train a response generation model for multi-party dialogues. To tackle this obstacle, we propose an Expectation-Maximization (EM) approach that iteratively performs the expectation steps to generate addressee labels, and the maximization steps to optimize a response generation model. Theoretical analyses and extensive experiments have justified the feasibility and effectiveness of our proposed method. The official implementation of this paper is available at https://github.com/EricLee8/MPDRG.

Medical Dialogue Generation via Dual Flow Modeling
Kaishuai Xu, Wenjun Hou, Yi Cheng, Jian Wang and Wenjie Li 11:00-12:30 (Pier 7&8)
Medical dialogue systems (MDS) aim to provide patients with medical services, such as diagnosis and prescription. Since most patients can- not precisely describe their symptoms, dialogue understanding is challenging for MDS. Previous studies mainly addressed this by extracting the mentioned medical entities as critical dialogue history information. In this work, we argue that it is also essential to capture the transitions of the medical entities and the doctor's dialogue acts in each turn, as they help the understanding of how the dialogue flows and enhance the prediction of the entities and dialogue acts to be adopted in the following turn. Correspondingly, we propose a Dual Flow enhanced Medical (DFMed) dialogue generation framework. It extracts the medical entities and dialogue acts used in the dialogue history and models their transitions with an entity-centric graph flow and a sequential act flow, respectively. We employ two sequential models to encode them and devise an interweaving component to enhance their interactions. Experiments on two datasets demonstrate that our method exceeds baselines in both automatic and manual evaluations.

Towards Open Environment Intent Prediction
Yunhua Zhou, Jiawei Hong and Xipeng Qiu 11:00-12:30 (Pier 7&8)
Out-of-Domain (OOD) Intent Classification and New Intent Discovering are two basic and critical tasks in the Task-Oriented Dialogue Sys-

tem, which are typically treated two independent tasks. Classification focuses on identifying intents beyond the predefined set of the dialog system, but it will not further differentiate detected OOD intents in fine granularity. Discovering focuses on how to cluster unlabeled samples according to their semantic representation, which relies heavily on prior knowledge and can not provide label information for the formed clusters. To be closer to the real user-facing scenarios, we introduce a task paradigm to extend Classification with Discovering referred as Open Environment Intent Prediction, which is to make a further fine-grained discovery of OOD based on OOD Intent Classification. Using various widely-used generative models as an archetype, we propose a general scheme for Open Environment Intent Prediction. In a nutshell, we first perform intent detection to identify the In-domain (IND) samples and then generate labels for those identified as OOD. With these generated labels, we can discover new general intents and provide label information for them. We develop a suite of benchmarks on the existing intent datasets and present a simple yet effective implementation. Extensive experiments demonstrate that our method establishes substantial improvement compared to the baselines.

A Synthetic Data Generation Framework for Grounded Dialogues
Jianzhu Bao, Rui Wang, Yasheng Wang, Aixin Sun, Yitong Li, Fei Mi and Ruifeng Xu   11:00-12:30 (Pier 7&8)
Training grounded response generation models often requires a large collection of grounded dialogues. However, it is costly to build such dialogues. In this paper, we present a synthetic data generation framework (SynDG) for grounded dialogues. The generation process utilizes large pre-trained language models and freely available knowledge data (e.g., Wikipedia pages, persona profiles, etc.). The key idea of designing SynDG is to consider dialogue flow and coherence in the generation process. Specifically, given knowledge data, we first heuristically determine a dialogue flow, which is a series of knowledge pieces. Then, we employ T5 to incrementally turn the dialogue flow into a dialogue. To ensure coherence of both the dialogue flow and the synthetic dialogue, we design a two-level filtering strategy, at the flow-level and the utterance-level respectively. Experiments on two public benchmarks show that the synthetic grounded dialogue data produced by our framework is able to significantly boost model performance in both full training data and low-resource scenarios.

Faithful Question Answering with Monte-Carlo Planning
Ruixin Hong, Hongming Zhang, Hong Zhao, Dong Yu and Changshui Zhang   11:00-12:30 (Pier 7&8)
Although large language models demonstrate remarkable question-answering performances, revealing the intermediate reasoning steps that the models faithfully follow remains challenging. In this paper, we propose FAME (FAithful question answering with MontE-carlo planning) to answer questions based on faithful reasoning steps. The reasoning steps are organized as a structured entailment tree, which shows how premises are used to produce intermediate conclusions that can prove the correctness of the answer. We formulate the task as a discrete decision-making problem and solve it through the interaction of a reasoning environment and a controller. The environment is modular and contains several basic task-oriented modules, while the controller proposes actions to assemble the modules. Since the search space could be large, we introduce a Monte-Carlo planning algorithm to do a look-ahead search and select actions that will eventually lead to high-quality steps. FAME achieves advanced performance on the standard benchmark. It can produce valid and faithful reasoning steps compared with large language models with a much smaller model size.

Expand, Rerank, and Retrieve: Query Reranking for Open-Domain Question Answering
Yung-Sung Chuang, Wei Fang, Shang-Wen Li, Wen-tau Yih and James Glass   11:00-12:30 (Pier 7&8)
We propose EAR, a query Expansion And Reranking approach for improving passage retrieval, with the application to open-domain question answering. EAR first applies a query expansion model to generate a diverse set of queries, and then uses a query reranker to select the ones that could lead to better retrieval results. Motivated by the observation that the best query expansion often is not picked by greedy decoding, EAR trains its reranker to predict the rank orders of the gold passages when issuing the expanded queries to a given retriever. By connecting better the query expansion model and retriever, EAR significantly enhances a traditional sparse retrieval method, BM25. Empirically, EAR improves top-5/20 accuracy by 3-8 and 5-10 points in in-domain and out-of-domain settings, respectively, when compared to a vanilla query expansion model, GAR, and a dense retrieval model, DPR.

An Empirical Comparison of LM-based Question and Answer Generation Methods
Asahi Ushio, Fernando Alva-Manchego and Jose Camacho-Collados   11:00-12:30 (Pier 7&8)
Question and answer generation (QAG) consists of generating a set of question-answer pairs given a context (e.g. a paragraph). This task has a variety of applications, such as data augmentation for question answering (QA) models, information retrieval and education. In this paper, we establish baselines with three different QAG methodologies that leverage sequence-to-sequence language model (LM) fine-tuning. Experiments show that an end-to-end QAG model, which is computationally light at both training and inference times, is generally robust and outperforms other more convoluted approaches. However, there are differences depending on the underlying generative LM. Finally, our analysis shows that QA models fine-tuned solely on generated question-answer pairs can be competitive when compared to supervised QA models trained on human-labeled data.

Combo of Thinking and Observing for Outside-Knowledge VQA
Qingyi Si, Yuchen Mo, Zheng Lin, HUISHAN JI and Weiping Wang   11:00-12:30 (Pier 7&8)
Outside-knowledge visual question answering is a challenging task that requires both the acquisition and the use of open-ended real-world knowledge. Some existing solutions draw external knowledge into the cross-modality space which overlooks the much vaster textual knowledge in natural-language space, while others transform the image into a text which further fuses with the textual knowledge into the natural-language space and completely abandons the use of visual features. In this paper, we are inspired to constrain the cross-modality space into the same space of natural-language space which makes the visual features preserved directly, and the model still benefits from the vast knowledge in natural-language space. To this end, we propose a novel framework consisting of a multimodal encoder, a textual encoder and an answer decoder. Such structure allows us to introduce more types of knowledge including explicit and implicit multimodal and textual knowledge. Extensive experiments validate the superiority of the proposed method which outperforms the state-of-the-art by 6.17% accuracy. We also conduct comprehensive ablations of each component, and systematically study the roles of varying types of knowledge. Codes and knowledge data are to be released.

Chain-of-Skills: A Configurable Model for Open-Domain Question Answering
Kaixin Ma, Hao Cheng, Yu Zhang, Xiaodong Liu, Eric Nyberg and Jianfeng Gao   11:00-12:30 (Pier 7&8)
The retrieval model is an indispensable component for real-world knowledge-intensive tasks, e.g., open-domain question answering (ODQA). As separate retrieval skills are annotated for different datasets, recent work focuses on customized methods, limiting the model transfer-ability and scalability. In this work, we propose a modular retriever where individual modules correspond to key skills that can be reused across datasets. Our approach supports flexible skill configurations based on the target domain to boost performance. To mitigate task interference, we design a novel modularization parameterization inspired by sparse Transformer. We demonstrate that our model can benefit from self-supervised pretraining on Wikipedia and fine-tuning using multiple ODQA datasets, both in a multi-task fashion. Our approach outperforms recent self-supervised retrievers in zero-shot evaluations and achieves state-of-the-art fine-tuned retrieval performance on NQ, HotpotQA and OTT-QA.

MultiTabQA: Generating Tabular Answers for Multi-Table Question Answering

Vaishali Pal, Andrew Yates, Evangelos Kanoulas and Maarten de Rijke 11:00-12:30 (Pier 7&8)

Recent advances in tabular question answering (QA) with large language models are constrained in their coverage and only answer questions over a single table. However, real-world queries are complex in nature, often over multiple tables in a relational database or web page. Single table questions do not involve common table operations such as set operations, Cartesian products (joins), or nested queries. Furthermore, multi-table operations often result in a tabular output, which necessitates table generation capabilities of tabular QA models. To fill this gap, we propose a new task of answering questions over multiple tables. Our model, MultiTabQA, not only answers questions over multiple tables, but also generalizes to generate tabular answers. To enable effective training, we build a pre-training dataset comprising of 132,645 SQL queries and tabular answers. Further, we evaluate the generated tables by introducing table-specific metrics of varying strictness assessing various levels of granularity of the table structure. MultiTabQA outperforms state-of-the-art single table QA models adapted to a multi-table QA setting by finetuning on three datasets: Spider, Atis and GeoQuery.

Using counterfactual contrast to improve compositional generalization for multi-step quantitative reasoning
Armineh Nourbakhsh, Sameena Shah and Carolyn Rosé 11:00-12:30 (Pier 7&8)

In quantitative question answering, compositional generalization is one of the main challenges of state of the art models, especially when longer sequences of reasoning steps are required. In this paper we propose CounterComp, a method that uses counterfactual scenarios to generate samples with compositional contrast. Instead of a data augmentation approach, CounterComp is based on metric learning, which allows for direct sampling from the training set and circumvents the need for additional human labels. Our proposed auxiliary metric learning loss improves the performance of three state of the art models on four recently released datasets. We also show how the approach can improve OOD performance on unseen domains, as well as unseen compositions. Lastly, we demonstrate how the method can lead to better compositional attention patterns during training.

The Dangers of trusting Stochastic Parrots: Faithfulness and Trust in Open-domain Conversational Question Answering
Sabrina Chiesurin, Dimitris Dimakopoulos, Marco Antonio Sobrevilla Cabezudo, Arash Eshghi, Ioannis Papaioannou, Verena Rieser and Ioannis Konstas 11:00-12:30 (Pier 7&8)

Large language models are known to produce output which sounds fluent and convincing, but is also often wrong, e.g. "unfaithful" with respect to a rationale as retrieved from a knowledge base. In this paper, we show that task-based systems which exhibit certain advanced linguistic dialog behaviors, such as lexical alignment (repeating what the user said), are in fact preferred and trusted more, whereas other phenomena, such as pronouns and ellipsis are dis-preferred. We use open-domain question answering systems as our test-bed for task based dialog generation and compare several open- and closed-book models. Our results highlight the danger of systems that appear to be trustworthy by parroting user input while providing an unfaithful response.

Hybrid Hierarchical Retrieval for Open-Domain Question Answering
Manoj Ghuhan Arivazhagan, Lan Liu, Peng Qi, Xinchi Chen, William Yang Wang and zhiheng huang 11:00-12:30 (Pier 7&8)

Retrieval accuracy is crucial to the performance of open-domain question answering (ODQA) systems. Recent work has demonstrated that dense hierarchical retrieval (DHR), which retrieves document candidates first and then relevant passages from the refined document set, can significantly outperform the single stage dense passage retriever (DPR). While effective, this approach requires document structure information to learn document representation and is hard to adopt to other domains without this information. Additionally, the dense retrievers tend to generalize poorly on out-of-domain data comparing with sparse retrievers such as BM25. In this paper, we propose Hybrid Hierarchical Retrieval (HHR) to address the existing limitations. Instead of relying solely on dense retrievers, we can apply sparse retriever, dense retriever, and a combination of them in both stages of document and passage retrieval. We perform extensive experiments on ODQA benchmarks and observe that our framework not only brings in-domain gains, but also generalizes better to zero-shot TriviaQA and Web Questions datasets with an average of 4.69% improvement on recall@100 over DHR. We also offer practical insights to trade off between retrieval accuracy, latency, and storage cost. The code is available on github.

Exploiting Abstract Meaning Representation for Open-Domain Question Answering
Cunxiang Wang, Zhikun Xu, Qipeng Guo, Xiangkun Hu, Xuefeng Bai, Zheng Zhang and Yue Zhang 11:00-12:30 (Pier 7&8)

The Open-Domain Question Answering (ODQA) task involves retrieving and subsequently generating answers from fine-grained relevant passages within a database. Current systems leverage Pretrained Language Models (PLMs) to model the relationship between questions and passages. However, the diversity in surface form expressions can hinder the model's ability to capture accurate correlations, especially within complex contexts. Therefore, we utilize Abstract Meaning Representation (AMR) graphs to assist the model in understanding complex semantic information. We introduce a method known as Graph-as-Token (GST) to incorporate AMRs into PLMs. Results from Natural Questions (NQ) and TriviaQA (TQ) demonstrate that our GST method can significantly improve performance, resulting in up to 2.44/3.17 Exact Match score improvements on NQ/TQ respectively. Furthermore, our method enhances robustness and outperforms alternative Graph Neural Network (GNN) methods for integrating AMRs. To the best of our knowledge, we are the first to employ semantic graphs in ODQA.

IDOL: Indicator-oriented Logic Pre-training for Logical Reasoning
Zihang Xu, Ziqing Yang, Yiming Cui and Shijin Wang 11:00-12:30 (Pier 7&8)

In the field of machine reading comprehension (MRC), existing systems have surpassed the average performance of human beings in many tasks like SQuAD. However, there is still a long way to go when it comes to logical reasoning. Although some methods for it have been put forward, they either are designed in a quite complicated way or rely too much on external structures. In this paper, we proposed IDOL (InDicator-Oriented Logic Pre-training), an easy-to-understand but highly effective further pre-training task which logically strengthens the pre-trained models with the help of 6 types of logical indicators and a logically rich dataset LoGic Pre-training (LGP). IDOL achieves state-of-the-art performance on ReClor and LogiQA, the two most representative benchmarks in logical reasoning MRC, and is proven to be capable of generalizing to different pre-trained models and other types of MRC benchmarks like RACE and SQuAD 2.0 while keeping competitive general language understanding ability through testing on tasks in GLUE. Besides, at the beginning of the era of large language models, we take several of them like ChatGPT into comparison and find that IDOL still shows its advantage.

A Survey for Efficient Open Domain Question Answering
Qin Zhang, Shangsi Chen, Dongkuan Xu, Qingqing Cao, Xiaojun Chen, Trevor Cohn and Meng Fang 11:00-12:30 (Pier 7&8)

Open domain question answering (ODQA) is a longstanding task aimed at answering factual questions from a large knowledge corpus without any explicit evidence in natural language processing (NLP). Recent works have predominantly focused on improving the answering accuracy and have achieved promising progress. However, higher accuracy often requires more memory consumption and inference latency, which might not necessarily be efficient enough for direct deployment in the real world. Thus, a trade-off between accuracy, memory consumption and processing speed is pursued. In this paper, we will survey recent advancements in the efficiency of ODQA models and conclude core techniques for achieving efficiency. Additionally, we will provide a quantitative analysis of memory cost, query speed, accuracy, and overall performance comparison. Our goal is to keep scholars informed of the latest advancements and open challenges in ODQA efficiency research and contribute to the further development of ODQA efficiency.

Generating Deep Questions with Commonsense Reasoning Ability from the Text by Disentangled Adversarial Inference

Jianxing Yu, Shiqi Wang, Libin Zheng, Qinliang Su, Wei Liu, Baoquan Zhao and Jian Yin   11:00-12:30 (Pier 7&8)

This paper proposes a new task of commonsense question generation, which aims to yield deep-level and to-the-point questions from the text. Their answers need to reason over disjoint relevant contexts and external commonsense knowledge, such as encyclopedic facts and causality. The knowledge may not be explicitly mentioned in the text but is used by most humans for problem-shooting. Such complex reasoning with hidden contexts involves deep semantic understanding. Thus, this task has great application value, such as making high-quality quizzes in advanced exams. Due to the lack of modeling complexity, existing methods may produce shallow questions that can be answered by simple word matching. To address these challenges, we propose a new QG model by simultaneously considering asking contents, expressive ways, and answering complexity. We first retrieve text-related commonsense context. Then we disentangle the key factors that control questions in terms of reasoning content and verbalized way. Independence priors and constraints are imposed to facilitate disentanglement. We further develop a discriminator to promote the deep results by considering their answering complexity. Through adversarial inference, we learn the latent factors from data. By sampling the expressive factor from the data distributions, diverse questions can be yielded. Evaluations of two typical data sets show the effectiveness of our approach.

Sentiment Knowledge Enhanced Self-supervised Learning for Multimodal Sentiment Analysis
Fan Qian, Jiqing Han, Yongjun He, Tieran Zheng and Guibin Zheng   11:00-12:30 (Pier 7&8)

Multimodal Sentiment Analysis (MSA) has made great progress that benefits from extraordinary fusion scheme. There is a lack of labeled data, resulting in severe overfitting and poor generalization for supervised models applied in this field. In this paper, we propose Sentiment Knowledge Enhanced Self-supervised Learning (SKESL) to capture common sentimental patterns in unlabeled videos, which facilitates further learning on limited labeled data. Specifically, with the help of sentiment knowledge and non-verbal behavior, SKESL conducts sentiment word masking and predicts fine-grained word sentiment intensity, so as to embed sentiment information at the word level into pre-trained multimodal representation. In addition, a non-verbal injection method is also proposed to integrate non-verbal information into the word semantics. Experiments on two standard benchmarks of MSA clearly show that SKESL significantly outperforms the baseline, and achieves new State-Of-The-Art (SOTA) results.

Multilingual Multi-Figurative Language Detection
Huiyuan Lai, Antonio Toral and Malvina Nissim   11:00-12:30 (Pier 7&8)

Figures of speech help people express abstract concepts and evoke stronger emotions than literal expressions, thereby making texts more creative and engaging. Due to its pervasive and fundamental character, figurative language understanding has been addressed in Natural Language Processing, but it's highly understudied in a multilingual setting and when considering more than one figure of speech at the same time. To bridge this gap, we introduce multilingual multi-figurative language modelling, and provide a benchmark for sentence-level figurative language detection, covering three common figures of speech and seven languages. Specifically, we develop a framework for figurative language detection based on template-based prompt learning. In so doing, we unify multiple detection tasks that are interrelated across multiple figures of speech and languages, without requiring task- or language-specific modules. Experimental results show that our framework outperforms several strong baselines and may serve as a blueprint for the joint modelling of other interrelated tasks.

Estimating the Uncertainty in Emotion Attributes using Deep Evidential Regression
Wen Wu, Chao Zhang and Philip C. Woodland   11:00-12:30 (Pier 7&8)

In automatic emotion recognition (AER), labels assigned by different human annotators to the same utterance are often inconsistent due to the inherent complexity of emotion and the subjectivity of perception. Though deterministic labels generated by averaging or voting are often used as the ground truth, it ignores the intrinsic uncertainty revealed by the inconsistent labels. This paper proposes a Bayesian approach, deep evidential emotion regression (DEER), to estimate the uncertainty in emotion attributes. Treating the emotion attribute labels of an utterance as samples drawn from an unknown Gaussian distribution, DEER places an utterance-specific normal-inverse gamma prior over the Gaussian likelihood and predicts its hyper-parameters using a deep neural network model. It enables a joint estimation of emotion attributes along with the aleatoric and epistemic uncertainties. AER experiments on the widely used MSP-Podcast and IEMOCAP datasets showed DEER produced state-of-the-art results for both the mean values and the distribution of emotion attributes.

Bidirectional Generative Framework for Cross-domain Aspect-based Sentiment Analysis
Yue Deng, Wenxuan Zhang, Sinno Jialin Pan and Lidong Bing   11:00-12:30 (Pier 7&8)

Cross-domain aspect-based sentiment analysis (ABSA) aims to perform various fine-grained sentiment analysis tasks on a target domain by transferring knowledge from a source domain. Since labeled data only exists in the source domain, a model is expected to bridge the domain gap for tackling cross-domain ABSA. Though domain adaptation methods have proven to be effective, most of them are based on a discriminative model, which needs to be specifically designed for different ABSA tasks. To offer a more general solution, we propose a unified bidirectional generative framework to tackle various cross-domain ABSA tasks. Specifically, our framework trains a generative model in both text-to-label and label-to-text directions. The former transforms each task into a unified format to learn domain-agnostic features, and the latter generates natural sentences from noisy labels for data augmentation, with which a more accurate model can be trained. To investigate the effectiveness and generality of our framework, we conduct extensive experiments on four cross-domain ABSA tasks and present new state-of-the-art results on all tasks. Our data and code are publicly available at https://github.com/DAMO-NLP-SG/BGCA.

NormBank: A Knowledge Bank of Situational Social Norms
Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy and Diyi Yang   11:00-12:30 (Pier 7&8)

We present NormBank, a knowledge bank of 155k situational norms. This resource is designed to ground flexible normative reasoning for interactive, assistive, and collaborative AI systems. Unlike prior commonsense resources, NormBank grounds each inference within a multivalent sociocultural frame, which includes the setting (e.g., restaurant), the agents' contingent roles (waiter, customer), their attributes (age, gender), and other physical, social, and cultural constraints (e.g., the temperature or the country of operation). In total, NormBank contains 63k unique constraints from a taxonomy that we introduce and iteratively refine here. Constraints then apply in different combinations to frame social norms. Under these manipulations, norms are non-monotonic — one can cancel an inference by updating its frame even slightly. Still, we find evidence that neural models can help reliably extend the scope and coverage of NormBank. We further demonstrate the utility of this resource with a series of transfer experiments. For data and code, see https://github.com/SALT-NLP/normbank.

Counterfactual Probing for the Influence of Affect and Specificity on Intergroup Bias
Venkata Subrahmanyan Govindarajan, David I Beaver, Kyle Mahowald and Junyi Jessy Li   11:00-12:30 (Pier 7&8)

While existing work on studying bias in NLP focues on negative or pejorative language use, Govindarajan et al. (2023) offer a revised framing of bias in terms of intergroup social context, and its effects on language behavior. In this paper, we investigate if two pragmatic features (specificity and affect) systematically vary in different intergroup contexts — thus connecting this new framing of bias to language output. Preliminary analysis finds modest correlations between specificity and affect of tweets with supervised intergroup relationship (IGR) labels. Counterfactual probing further reveals that while neural models finetuned for predicting IGR reliably use affect in classification, the model's usage of specificity is inconclusive.

Modeling Cross-Cultural Pragmatic Inference with Codenames Duet

Omar Shaikh, Caleb Ziems, William Held, Aryan J Pariani, Fred Morstatter and Diyi Yang   11:00-12:30 (Pier 7&8)

Pragmatic reference enables efficient interpersonal communication. Prior work uses simple reference games to test models of pragmatic reasoning, often with unidentified speakers and listeners. In practice, however, speakers' sociocultural background shapes their pragmatic assumptions. For example, readers of this paper assume NLP refers to Natural Language Processing, and not "Neuro-linguistic Programming." This work introduces the Cultural Codes dataset, which operationalizes sociocultural pragmatic inference in a simple word reference game. Cultural Codes is based on the multi-turn collaborative two-player game, Codenames Duet. Our dataset consists of 794 games with 7,703 turns, distributed across 153 unique players. Alongside gameplay, we collect information about players' personalities, values, and demographics. Utilizing theories of communication and pragmatics, we predict each player's actions via joint modeling of their sociocultural priors and the game context. Our experiments show that accounting for background characteristics significantly improves model performance for tasks related to both clue-giving and guessing, indicating that sociocultural priors play a vital role in gameplay decisions.

Interactive Concept Learning for Uncovering Latent Themes in Large Text Collections
Maria Leonor Pacheco, Tunazzina Islam, Lyle Ungar, Ming Yin and Dan Goldwasser   11:00-12:30 (Pier 7&8)

Experts across diverse disciplines are often interested in making sense of large text collections. Traditionally, this challenge is approached either by noisy unsupervised techniques such as topic models, or by following a manual theme discovery process. In this paper, we expand the definition of a theme to account for more than just a word distribution, and include generalized concepts deemed relevant by domain experts. Then, we propose an interactive framework that receives and encodes expert feedback at different levels of abstraction. Our framework strikes a balance between automation and manual coding, allowing experts to maintain control of their study while reducing the manual effort required.

Grounded Multimodal Named Entity Recognition on Social Media
Jianfei Yu, Ziyan Li, Jieming Wang and Rui Xia   11:00-12:30 (Pier 7&8)

In recent years, Multimodal Named Entity Recognition (MNER) on social media has attracted considerable attention. However, existing MNER studies only extract entity-type pairs in text, which is useless for multimodal knowledge graph construction and insufficient for entity disambiguation. To solve these issues, in this work, we introduce a Grounded Multimodal Named Entity Recognition (GMNER) task. Given a text-image social post, GMNER aims to identify the named entities in text, their entity types, and their bounding box groundings in image (i.e. visual regions). To tackle the GMNER task, we construct a Twitter dataset based on two existing MNER datasets. Moreover, we extend four well-known MNER methods to establish a number of baseline systems and further propose a Hierarchical Index generation framework named H-Index, which generates the entity-type-region triples in a hierarchical manner with a sequence-to-sequence model. Experiment results on our annotated dataset demonstrate the superiority of our H-Index framework over baseline systems on the GMNER task.

UPPAM: A Unified Pre-training Architecture for Political Actor Modeling based on Language
Xinyi Mou, Zhongyu Wei, Qi Zhang and Xuanjing Huang   11:00-12:30 (Pier 7&8)

Modeling political actors is at the core of quantitative political science. Existing works have incorporated contextual information to better learn the representation of political actors for specific tasks through graph models. However, they are limited to the structure and objective of training settings and can not be generalized to all politicians and other tasks. In this paper, we propose a Unified Pre-training Architecture for Political Actor Modeling based on language (UPPAM). In UPPAM, we aggregate statements to represent political actors and learn the mapping from languages to representation, instead of learning the representation of particular persons. We further design structure-aware contrastive learning and behavior-driven contrastive learning tasks, to inject multidimensional information in the political context into the mapping. In this framework, we can profile political actors from different aspects and solve various downstream tasks. Experimental results demonstrate the effectiveness and capability of generalization of our method.

Additive manifesto decomposition: A policy domain aware method for understanding party positioning
Tanise Ceron, Dmitry Nikolaev and Sebastian Padó   11:00-12:30 (Pier 7&8)

Automatic extraction of party (dis)similarities from texts such as party election manifestos or parliamentary speeches plays an increasing role in computational political science. However, existing approaches are fundamentally limited to targeting only global party (dis)-similarity: they condense the relationship between a pair of parties into a single figure, their similarity. In aggregating over all policy domains (e.g., health or foreign policy), they do not provide any qualitative insights into which domains parties agree or disagree on. This paper proposes a workflow for estimating policy domain aware party similarity that overcomes this limitation. The workflow covers (a) definition of suitable policy domains; (b) automatic labeling of domains, if no manual labels are available; (c) computation of domain-level similarities and aggregation at a global level; (d) extraction of interpretable party positions on major policy axes via multidimensional scaling. We evaluate our workflow on manifestos from the German federal elections. We find that our method (a) yields high correlation when predicting party similarity at a global level and (b) provides accurate party-specific positions, even with automatically labelled policy domains.

COBRA Frames: Contextual Reasoning about Effects and Harms of Offensive Statements
Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta and Maarten Sap   11:00-12:30 (Pier 7&8)

Warning: This paper contains content that may be offensive or upsetting. Understanding the harms and offensiveness of statements requires reasoning about the social and situational context in which statements are made. For example, the utterance "your English is very good" may implicitly signal an insult when uttered by a white man to a non-white colleague, but uttered by an ESL teacher to their student would be interpreted as a genuine compliment. Such contextual factors have been largely ignored by previous approaches to toxic language detection. We introduce COBRA frames, the first context-aware formalism for explaining the intents, reactions, and harms of offensive or biased statements grounded in their social and situational context. We create COBRACORPUS, a dataset of 33k potentially offensive statements paired with machine-generated contexts and free-text explanations of offensiveness, implied biases, speaker intents, and listener reactions. To study the contextual dynamics of offensiveness, we train models to generate COBRA explanations, with and without access to the context. We find that explanations by context-agnostic models are significantly worse than by context-aware ones, especially in situations where the context inverts the statement's offensiveness (29% accuracy drop). Our work highlights the importance and feasibility of contextualized NLP by modeling social factors.

MultiEMO: An Attention-Based Correlation-Aware Multimodal Fusion Framework for Emotion Recognition in Conversations
Tao Shi and Shao-Lun Huang   11:00-12:30 (Pier 7&8)

Emotion Recognition in Conversations (ERC) is an increasingly popular task in the Natural Language Processing community, which seeks to achieve accurate emotion classifications of utterances expressed by speakers during a conversation. Most existing approaches focus on modeling speaker and contextual information based on the textual modality, while the complementarity of multimodal information has not been well leveraged, few current methods have sufficiently captured the complex correlations and mapping relationships across different modalities. Furthermore, existing state-of-the-art ERC models have difficulty classifying minority and semantically similar emotion categories. To address these challenges, we propose a novel attention-based correlation-aware multimodal fusion framework named MultiEMO, which effectively integrates multimodal cues by capturing cross-modal mapping relationships across textual, audio and visual modalities based on bidirectional multi-head cross-attention layers. The difficulty of recognizing minority and semantically hard-to-distinguish emotion classes is alleviated by our proposed Sample-Weighted Focal Contrastive (SWFC) loss. Extensive experiments on two benchmark ERC datasets

demonstrate that our MultiEMO framework consistently outperforms existing state-of-the-art approaches in all emotion categories on both datasets, the improvements in minority and semantically similar emotions are especially significant.

Scale-Invariant Infinite Hierarchical Topic Model
Shusei Eshima and Daichi Mochihashi   11:00-12:30 (Pier 7&8)
Hierarchical topic models have been employed to organize a large number of diverse topics from corpora into a latent tree structure. However, existing models yield fragmented topics with overlapping themes whose expected probability becomes exponentially smaller along the depth of the tree.
To solve this intrinsic problem, we propose a scale-invariant infinite hierarchical topic model (ihLDA). The ihLDA adaptively adjusts the topic creation to make the expected topic probability decay considerably slower than that in existing models. Thus, it facilitates the estimation of deeper topic structures encompassing diverse topics in a corpus. Furthermore, the ihLDA extends a widely used tree-structured prior (Adams et al., 2010) in a hierarchical Bayesian way, which enables drawing an infinite topic tree from the base tree while efficiently sampling the topic assignments for the words.
Experiments demonstrate that the ihLDA has better topic uniqueness and hierarchical diversity than existing approaches, including state-of-the-art neural models.

Zero-Shot Classification by Logical Reasoning on Natural Language Explanations
Chi Han, Hengzhi Pei, Xinya Du and Heng Ji   11:00-12:30 (Pier 7&8)
Humans can classify data of an unseen category by reasoning on its language explanations. This ability is owing to the compositional nature of language: we can combine previously seen attributes to describe the new category. For example, we might describe a sage thrasher as "it has a slim straight relatively short bill, yellow eyes and a long tail", so that others can use their knowledge of attributes "slim straight relatively short bill", "yellow eyes" and "long tail" to recognize a sage thrasher. Inspired by this observation, in this work we tackle zero-shot classification task by logically parsing and reasoning on natural language explanations. To this end, we propose the framework CLORE (Classification by LOgical Reasoning on Explanations). While previous methods usually regard textual information as implicit features, CLORE parses explanations into logical structures and then explicitly reasons along this structure on the input to produce a classification score. Experimental results on explanation-based zero-shot classification benchmarks demonstrate that CLORE is superior to baselines, which we show is mainly due to higher scores on tasks requiring more logical reasoning. We also demonstrate that our framework can be extended to zero-shot classification on visual modality. Alongside classification decisions, CLORE can provide the logical parsing and reasoning process as a clear form of rationale. Through empirical analysis we demonstrate that CLORE is also less affected by linguistic biases than baselines.

EmbedTextNet: Dimension Reduction with Weighted Reconstruction and Correlation Losses for Efficient Text Embedding
Dae Yon Hwang, Bilal Taha and Yaroslav Nechaev   11:00-12:30 (Pier 7&8)
The size of embeddings generated by large language models can negatively affect system latency and model size in certain downstream practical applications (e.g. KNN search). In this work, we propose EmbedTextNet, a light add-on network that can be appended to an arbitrary language model to generate a compact embedding without requiring any changes in its architecture or training procedure. Specifically, we use a correlation penalty added to the weighted reconstruction loss that better captures the informative features in the text embeddings, which improves the efficiency of the language models. We evaluated EmbedTextNet on three different downstream tasks: text similarity, language modelling, and text retrieval. Empirical results on diverse benchmark datasets demonstrate the effectiveness and superiority of EmbedTextNet compared to state-of-art methodologies in recent works, especially in extremely low dimensional embedding sizes. The developed code for reproducibility is included in the supplementary material.

A Memory Model for Question Answering from Streaming Data Supported by Rehearsal and Anticipation of Coreference Information
Vladimir Araujo, Alvaro M Soto and Marie-Francine Moens   11:00-12:30 (Pier 7&8)
Existing question answering methods often assume that the input content (e.g., documents or videos) is always accessible to solve the task. Alternatively, memory networks were introduced to mimic the human process of incremental comprehension and compression of the information in a fixed-capacity memory. However, these models only learn how to maintain memory by backpropagating errors in the answers through the entire network. Instead, it has been suggested that humans have effective mechanisms to boost their memorization capacities, such as rehearsal and anticipation. Drawing inspiration from these, we propose a memory model that performs rehearsal and anticipation while processing inputs to memorize important information for solving question answering tasks from streaming data. The proposed mechanisms are applied self-supervised during training through masked modeling tasks focused on coreference information. We validate our model on a short-sequence (bAbI) dataset as well as large-sequence textual (NarrativeQA) and video (ActivityNet-QA) question answering datasets, where it achieves substantial improvements over previous memory network approaches. Furthermore, our ablation study confirms the proposed mechanisms' importance for memory models.

Class-Incremental Learning based on Label Generation
Yijia Shao, Yiduo Guo, Dongyan Zhao and Bing Liu   11:00-12:30 (Pier 7&8)
Despite the great success of pre-trained language models, it is still a challenge to use these models for continual learning, especially for the class-incremental learning (CIL) setting due to catastrophic forgetting (CF). This paper reports our finding that if we formulate CIL as a continual label generation problem, CF is drastically reduced and the generalizable representations of pre-trained models can be better retained. We thus propose a new CIL method (VAG) that also leverages the sparsity of vocabulary to focus the generation and creates pseudo-replay samples by using label semantics. Experimental results show that VAG outperforms baselines by a large margin.

Enhancing Out-of-Vocabulary Estimation with Subword Attention
Raj Patel and Carlotta Domeniconi   11:00-12:30 (Pier 7&8)
Word embedding methods like word2vec and GloVe have been shown to learn strong representations of words. However, these methods only learn representations for words in the training corpus and therefore struggle to handle unknown and new words, known as out-of-vocabulary (OOV) words. As a result, there have been multiple attempts to learn OOV word representations in a similar fashion to how humans learn new words, using word roots/subwords and/or surrounding words. However, while most of these approaches use advanced architectures like attention on the context of the OOV word, they tend to use simple structures like ngram addition or character based convolutional neural networks (CNN) to handle processing subword information. In response to this, we propose SubAtt, a transformer based OOV estimation model that uses attention mechanisms on both the context and the subwords. In addition to attention, we also show that pretraining subword representations also leads to improvement in OOV estimation. We show SubAtt outperforms current state-of-the-art OOV estimation models.

PreQuant: A Task-agnostic Quantization Approach for Pre-trained Language Models
Zhuocheng Gong, Jiahao Liu, Qifan Wang, Yang Yang, Jingang Wang, Wei Wu, Yunsen Xian, Dongyan Zhao and Rui Yan   11:00-12:30 (Pier 7&8)
While transformer-based pre-trained language models (PLMs) have dominated a number of NLP applications, these models are heavy to deploy and expensive to use. Therefore, effectively compressing large-scale PLMs becomes an increasingly important problem. Quantization,

which represents high-precision tensors with low-bit fix-point format, is a viable solution. However, most existing quantization methods are task-specific, requiring customized training and quantization with a large number of trainable parameters on each individual task. Inspired by the observation that the over-parameterization nature of PLMs makes it possible to freeze most of the parameters during the fine-tuning stage, in this work, we propose a novel "quantize before fine-tuning" framework, PreQuant, that differs from both quantization-aware training and post-training quantization. {pasted macro 'OUR'} is compatible with various quantization strategies, with outlier-aware parameter-efficient fine-tuning to correct the induced quantization error. We demonstrate the effectiveness of PreQuant on the GLUE benchmark using BERT, RoBERTa, and T5. We also provide an empirical investigation into the workflow of PreQuant, which sheds light on its efficacy.

Domain Aligned Prefix Averaging for Domain Generalization in Abstractive Summarization
Pranav Ajit Nair, Sukomal Pal and Pradeepika Verma   11:00-12:30 (Pier 7&8)
Domain generalization is hitherto an underexplored area applied in abstractive summarization. Moreover, most existing works on domain generalization have sophisticated training algorithms. In this paper, we propose a lightweight, weight averaging based, Domain Aligned Prefix Averaging approach to domain generalization for abstractive summarization. Given a number of source domains, our method first trains a prefix for each one of them. These source prefixes generate summaries for a small number of target domain documents. The similarity of the generated summaries to their corresponding source documents is used for calculating weights required to average source prefixes. In DAPA, prefix tuning allows for lightweight finetuning, and weight averaging allows for the computationally efficient addition of new source domains. When evaluated on four diverse summarization domains, DAPA shows comparable or better performance against the baselines demonstrating the effectiveness of its prefix averaging scheme.

Not Enough Data to Pre-train Your Language Model? MT to the Rescue!
Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi and Ander Corral   11:00-12:30 (Pier 7&8)
In recent years, pre-trained transformer-based language models (LM) have become a key resource for implementing most NLP tasks. However, pre-training such models demands large text collections not available in most languages. In this paper, we study the use of machine-translated corpora for pre-training LMs. We answer the following research questions: RQ1: Is MT-based data an alternative to real data for learning a LM?; RQ2: Can real data be complemented with translated data and improve the resulting LM? In order to validate these two questions, several BERT models for Basque have been trained, combining real data and synthetic data translated from Spanish. The evaluation carried out on 9 NLU tasks indicates that models trained exclusively on translated data offer competitive results. Furthermore, models trained with real data can be improved with synthetic data, although further research is needed on the matter.

Label Agnostic Pre-training for Zero-shot Text Classification
Christopher Clarke, Yuzhao Heng, Yiping Kang, Krisztian Flautner, Lingjia Tang and Jason Mars   11:00-12:30 (Pier 7&8)
Conventional approaches to text classification typically assume the existence of a fixed set of predefined labels to which a given text can be classified. However, in real-world applications, there exists an infinite label space for describing a given text. In addition, depending on the aspect (sentiment, topic, etc.) and domain of the text (finance, legal, etc.), the interpretation of the label can vary greatly. This makes the task of text classification, particularly in the zero-shot scenario, extremely challenging. In this paper, we investigate the task of zero-shot text classification with the aim of improving the ability of pre-trained language models (PLMs) to generalize to both seen and unseen data across varying aspects and domains. To solve this we introduce two new simple yet effective pre-training strategies, Implicit and Explicit pre-training. These methods inject aspect-level understanding into the model at train time with the goal of conditioning the model to build task-level understanding. To evaluate this, we construct and release UTCD, a new benchmark dataset for evaluating text classification in zero-shot settings. Experimental results on UTCD show that our approach achieves improved zero-shot generalization on a suite of challenging datasets across an array of zero-shot formalizations.

MANNER: A Variational Memory-Augmented Model for Cross Domain Few-Shot Named Entity Recognition
Jinyuan Fang, Xiaobin Wang, Zaiqiao Meng, Pengjun Xie, Fei Huang and Yong Jiang   11:00-12:30 (Pier 7&8)
This paper focuses on the task of cross domain few-shot named entity recognition (NER), which aims to adapt the knowledge learned from source domain to recognize named entities in target domain with only a few labeled examples. To address this challenging task, we propose MANNER, a variational memory-augmented few-shot NER model. Specifically, MANNER uses a memory module to store information from the source domain and then retrieve relevant information from the memory to augment few-shot task in the target domain. In order to effectively utilize the information from memory, MANNER uses optimal transport to retrieve and process information from memory, which can explicitly adapt the retrieved information from source domain to target domain and improve the performance in the cross domain few-shot setting. We conduct experiments on English and Chinese cross domain few-shot NER datasets, and the experimental results demonstrate that MANNER can achieve superior performance.

TeAST: Temporal Knowledge Graph Embedding via Archimedean Spiral Timeline
Jiang Li, Xiangdong Su and Guanglai Gao   11:00-12:30 (Pier 7&8)
Temporal knowledge graph embedding (TKGE) models are commonly utilized to infer the missing facts and facilitate reasoning and decision-making in temporal knowledge graph based systems. However, existing methods fuse temporal information into entities, potentially leading to the evolution of entity information and limiting the link prediction performance of TKG. Meanwhile, current TKGE models often lack the ability to simultaneously model important relation patterns and provide interpretability, which hinders their effectiveness and potential applications. To address these limitations, we propose a novel TKGE model which encodes **T**emporal knowledge graph **e**mbeddings via **A**rchimedean **S**piral **T**imeline (TeAST), which maps relations onto the corresponding Archimedean spiral timeline and transforms the quadruples completion to 3th-order tensor completion problem. Specifically, the Archimedean spiral timeline ensures that relations that occur simultaneously are placed on the same timeline, and all relations evolve over time. Meanwhile, we present a novel temporal spiral regularizer to make the spiral timeline orderly. In addition, we provide mathematical proofs to demonstrate the ability of TeAST to encode various relation patterns. Experimental results show that our proposed model significantly outperforms existing TKGE methods. Our code is available at https://github.com/IMU-MachineLearningSXD/TeAST.

DSEE: Dually Sparsity-embedded Efficient Tuning of Pre-trained Language Models
Xuxi Chen, Tianlong Chen, Weizhu Chen, Ahmed Hassan Awadallah, Zhangyang Wang and Yu Cheng   11:00-12:30 (Pier 7&8)
Gigantic pre-trained models have become central to natural language processing (NLP), serving as the starting point for fine-tuning towards a range of downstream tasks. However, two pain points persist for this paradigm: (a) as the pre-trained models grow bigger (e.g., 175B parameters for GPT-3), even the fine-tuning process can be time-consuming and computationally expensive; (b) the fine-tuned model has the same size as its starting point by default, which is neither sensible due to its more specialized functionality, nor practical since many fine-tuned models will be deployed in resource-constrained environments. To address these pain points, we propose a framework for resource- and parameter-efficient fine-tuning by leveraging the sparsity prior in both weight updates and the final model weights. Our proposed framework, dubbed Dually Sparsity-Embedded Efficient Tuning (DSEE), aims to achieve two key objectives: (i) parameter efficient fine-tuning - by enforcing sparsity-aware low-rank updates on top of the pre-trained weights; and (ii) resource-efficient inference - by encouraging a sparse weight structure towards the final fine-tuned model. We leverage sparsity in these two directions by exploiting both unstructured and structured sparse patterns in pre-trained language models via a unified approach. Extensive experiments and in-depth investigations, with diverse

network backbones (i.e., BERT, RoBERTa, and GPT-2) on dozens of datasets, consistently demonstrate impressive parameter-/inference-efficiency, while maintaining competitive downstream performance. For instance, DSEE saves about 25% inference FLOPs while achieving comparable performance, with 0.5% trainable parameters on BERT. Codes are available at https://github.com/VITA-Group/DSEE.

Text Adversarial Purification as Defense against Adversarial Attacks
Linyang Li, Demin Song and Xipeng Qiu   11:00-12:30 (Pier 7&8)
Adversarial purification is a successful defense mechanism against adversarial attacks without requiring knowledge of the form of the incoming attack. Generally, adversarial purification aims to remove the adversarial perturbations therefore can make correct predictions based on the recovered clean samples. Despite the success of adversarial purification in the computer vision field that incorporates generative models such as energy-based models and diffusion models, using purification as a defense strategy against textual adversarial attacks is rarely explored. In this work, we introduce a novel adversarial purification method that focuses on defending against textual adversarial attacks. With the help of language models, we can inject noise by masking input texts and reconstructing the masked texts based on the masked language models. In this way, we construct an adversarial purification process for textual models against the most widely used word-substitution adversarial attacks. We test our proposed adversarial purification method on several strong adversarial attack methods including Textfooler and BERT-Attack and experimental results indicate that the purification algorithm can successfully defend against strong word-substitution attacks.

TART: Improved Few-shot Text Classification Using Task-Adaptive Reference Transformation
Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen and Chang-Tien Lu   11:00-12:30 (Pier 7&8)
Meta-learning has emerged as a trending technique to tackle few-shot text classification and achieve state-of-the-art performance. However, the performance of existing approaches heavily depends on the inter-class variance of the support set. As a result, it can perform well on tasks when the semantics of sampled classes are distinct while failing to differentiate classes with similar semantics. In this paper, we propose a novel Task-Adaptive Reference Transformation (TART) network, aiming to enhance the generalization by transforming the class prototypes to per-class fixed reference points in task-adaptive metric spaces. To further maximize divergence between transformed prototypes in task-adaptive metric spaces, TART introduces a discriminative reference regularization among transformed prototypes. Extensive experiments are conducted on four benchmark datasets and our method demonstrates clear superiority over the state-of-the-art models in all the datasets. In particular, our model surpasses the state-of-the-art method by 7.4% and 5.4% in 1-shot and 5-shot classification on the 20 Newsgroups dataset, respectively.

DaMSTF: Domain Adversarial Learning Enhanced Meta Self-Training for Domain Adaptation
Menglong Lu, Zhen Huang, Yunxiang Zhao, Zhiliang Tian, Yang Liu and Dongsheng Li   11:00-12:30 (Pier 7&8)
Self-training emerges as an important research line on domain adaptation. By taking the model's prediction as the pseudo labels of the unlabeled data, self-training bootstraps the model with pseudo instances in the target domain. However, the prediction errors of pseudo labels (label noise) challenge the performance of self-training. To address this problem, previous approaches only use reliable pseudo instances, i.e., pseudo instances with high prediction confidence, to retrain the model. Although these strategies effectively reduce the label noise, they are prone to miss the hard examples. In this paper, we propose a new self-training framework for domain adaptation, namely Domain adversarial learning enhanced Self-Training Framework (DaMSTF). Firstly, DaMSTF involves meta-learning to estimate the importance of each pseudo instance, so as to simultaneously reduce the label noise and preserve hard examples. Secondly, we design a meta constructor for constructing the meta-validation set, which guarantees the effectiveness of the meta-learning module by improving the quality of the meta-validation set. Thirdly, we find that the meta-learning module suffers from the training guidance vanish- ment and tends to converge to an inferior optimal. To this end, we employ domain adversarial learning as a heuristic neural network initialization method, which can help the meta-learning module converge to a better optimal. Theoretically and experimentally, we demonstrate the effectiveness of the proposed DaMSTF. On the cross-domain sentiment classification task, DaMSTF improves the performance of BERT with an average of nearly 4%.

AutoMoE: Heterogeneous Mixture-of-Experts with Adaptive Computation for Efficient Neural Machine Translation
Ganesh Jawahar, Subhabrata Mukherjee, Xiaodong Liu, Young Jin Kim, Muhammad Abdul-Mageed, Laks Lakshmanan, V.S., Ahmed Hassan Awadallah, Sebastien Bubeck and Jianfeng Gao   11:00-12:30 (Pier 7&8)
Mixture-of-Expert (MoE) models have obtained state-of-the-art performance in Neural Machine Translation (NMT) tasks. Existing works in MoE mostly consider a homogeneous design where the same number of experts of the same size are placed uniformly throughout the network. Furthermore, existing MoE works do not consider computational constraints (e.g., FLOPs, latency) to guide their design. To this end, we develop AutoMoE – a framework for designing heterogeneous MoE's under computational constraints. AutoMoE leverages Neural Architecture Search (NAS) to obtain efficient sparse MoE sub-transformers with 4x inference speedup (CPU) and FLOPs reduction over manually designed Transformers, with parity in BLEU score over dense Transformer and within 1 BLEU point of MoE SwitchTransformer, on aggregate over benchmark datasets for NMT. Heterogeneous search space with dense and sparsely activated Transformer modules (e.g., how many experts? where to place them? what should be their sizes?) allows for adaptive compute – where different amounts of computations are used for different tokens in the input. Adaptivity comes naturally from routing decisions which send tokens to experts of different sizes. AutoMoE code, data, and trained models are available at https://aka.ms/AutoMoE.

Neural Architecture Search for Parameter-Efficient Fine-tuning of Large Pre-trained Language Models
Neal G Lawton, Anoop Kumar, Govind Thattai, Aram Galstyan and Greg Ver Steeg   11:00-12:30 (Pier 7&8)
Parameter-efficient tuning (PET) methods fit pre-trained language models (PLMs) to downstream tasks by either computing a small compressed update for a subset of model parameters, or appending and fine-tuning a small number of new model parameters to the pre-trained network. Hand-designed PET architectures from the literature perform well in practice, but have the potential to be improved via automated neural architecture search (NAS). We propose an efficient NAS method for learning PET architectures via structured and unstructured pruning. We present experiments on GLUE demonstrating the effectiveness of our algorithm and discuss how PET architectural design choices affect performance in practice.

Simplicity Bias in Transformers and their Ability to Learn Sparse Boolean Functions
Satwik Bhattamishra, Arkil Patel, Varun Kanade and Phil Blunsom   11:00-12:30 (Pier 7&8)
Despite the widespread success of Transformers on NLP tasks, recent works have found that they struggle to model several formal languages when compared to recurrent models. This raises the question of why Transformers perform well in practice and whether they have any properties that enable them to generalize better than recurrent models. In this work, we conduct an extensive empirical study on Boolean functions to demonstrate the following: (i) Random Transformers are relatively more biased towards functions of low sensitivity. (ii) When trained on Boolean functions, both Transformers and LSTMs prioritize learning functions of low sensitivity, with Transformers ultimately converging to functions of lower sensitivity. (iii) On sparse Boolean functions which have low sensitivity, we find that Transformers generalize near perfectly even in the presence of noisy labels whereas LSTMs overfit and achieve poor generalization accuracy. Overall, our results provide strong quantifiable evidence that suggests differences in the inductive biases of Transformers and recurrent models which may help explain Transformer's effective generalization performance despite relatively limited expressiveness.

Structured Mean-Field Variational Inference for Higher-Order Span-Based Semantic Role

Wei Liu, Songlin Yang and Kewei Tu   11:00-12:30 (Pier 7&8)
In this work, we enhance higher-order graph-based approaches for span-based semantic role labeling (SRL) by means of structured modeling. To decrease the complexity of higher-order modeling, we decompose the edge from predicate word to argument span into three different edges, predicate-to-head (P2H), predicate-to-tail (P2T), and head-to-tail (H2T), where head/tail means the first/last word of the semantic argument span. As such, we use a CRF-based higher-order dependency parser and leverage Mean-Field Variational Inference (MFVI) for higher-order inference. Moreover, since semantic arguments of predicates are often constituents within a constituency parse tree, we can leverage such nice structural property by defining a TreeCRF distribution over all H2T edges, using the idea of partial marginalization to define structural training loss. We further leverage structured MFVI to enhance inference. We experiment on span-based SRL benchmarks, showing the effectiveness of both higher-order and structured modeling and the combination thereof. In addition, we show superior performance of structured MFVI against vanilla MFVI.

Towards Distribution-shift Robust Text Classification of Emotional Content
Luana Bulla, Aldo Gangemi and Misael Mongiovi'   11:00-12:30 (Pier 7&8)
Supervised models based on Transformers have been shown to achieve impressive performances in many natural language processing tasks. However, besides requiring a large amount of costly manually annotated data, supervised models tend to adapt to the characteristics of the training dataset, which are usually created ad-hoc and whose data distribution often differs from the one in real applications, showing significant performance degradation in real-world scenarios. We perform an extensive assessment of the out-of-distribution performances of supervised models for classification in the emotion and hate-speech detection tasks and show that NLI-based zero-shot models often outperform them, making task-specific annotation useless when the characteristics of final-user data are not known in advance. To benefit from both supervised and zero-shot approaches, we propose to fine-tune an NLI-based model on the task-specific dataset. The resulting model often outperforms all available supervised models both in distribution and out of distribution, with only a few thousand training samples.

Boosting Text Augmentation via Hybrid Instance Filtering Framework
Heng Yang and Ke Li   11:00-12:30 (Pier 7&8)
Text augmentation is an effective technique for addressing the problem of insufficient data in natural language processing. However, existing text augmentation methods tend to focus on few-shot scenarios and usually perform poorly on large public datasets. Our research indicates that existing augmentation methods often generate instances with shifted feature spaces, which leads to a drop in performance on the augmented data (for example, EDA generally loses approximately 2% in aspect-based sentiment classification). To address this problem, we propose a hybrid instance-filtering framework (BoostAug) based on pre-trained language models that can maintain a similar feature space with natural datasets. BoostAug is transferable to existing text augmentation methods (such as synonym substitution and back translation) and significantly improves the augmentation performance by 2-3% in classification accuracy. Our experimental results on three classification tasks and nine public datasets show that BoostAug addresses the performance drop problem and outperforms state-of-the-art text augmentation methods. Additionally, we release the code to help improve existing augmentation methods on large datasets.

Code Execution with Pre-trained Language Models
Chenxiao Liu, Shuai Lu, Weizhu Chen, Daxin Jiang, Alexey Svyatkovskiy, Shengyu Fu, Neel Sundaresan and Nan Duan   11:00-12:30 (Pier 7&8)
Code execution is a fundamental aspect of programming language semantics that reflects the exact behavior of the code. However, most pre-trained models for code intelligence ignore the execution trace and only rely on source code and syntactic structures. In this paper, we investigate how well pre-trained models can understand and perform code execution. We develop a mutation-based data augmentation technique to create a large-scale and realistic Python dataset and task for code execution, which challenges existing models such as Codex. We then present CodeExecutor, a Transformer model that leverages code execution pre-training and curriculum learning to enhance its semantic comprehension. We evaluate CodeExecutor on code execution and show its promising performance and limitations. We also demonstrate its potential benefits for code intelligence tasks such as zero-shot code-to-code search and text-to-code generation. Our analysis provides insights into the learning and generalization abilities of pre-trained models for code execution.

HiFi: High-Information Attention Heads Hold for Parameter-Efficient Model Adaptation
Anchun Gui and Han Xiao   11:00-12:30 (Pier 7&8)
To fully leverage the advantages of large-scale pre-trained language models (PLMs) on downstream tasks, it has become a ubiquitous adaptation paradigm to fine-tune the entire parameters of PLMs. However, this paradigm poses issues of inefficient updating and resource over-consuming for fine-tuning in data-scarce and resource-limited scenarios, because of the large scale of parameters in PLMs. To alleviate these concerns, in this paper, we propose a parameter-efficient fine-tuning method HiFi, that is, only the highly informative and strongly correlated attention heads for the specific task are fine-tuned. To search for those significant attention heads, we develop a novel framework to analyze the effectiveness of heads. Specifically, we first model the relationship between heads into a graph from two perspectives of information richness and correlation, and then apply PageRank algorithm to determine the relative importance of each head. Extensive experiments on the GLUE benchmark demonstrate the effectiveness of our method, and show that HiFi obtains state-of-the-art performance over the prior baselines.

How does the task complexity of masked pretraining objectives affect downstream performance?
Atsuki Yamaguchi, Hiroaki Ozaki, Terufumi Morishita, Gaku Morio and Yasuhiro Sogawa   11:00-12:30 (Pier 7&8)
Masked language modeling (MLM) is a widely used self-supervised pretraining objective, where a model needs to predict an original token that is replaced with a mask given contexts. Although simpler and computationally efficient pretraining objectives, e.g., predicting the first character of a masked token, have recently shown comparable results to MLM, no objectives with a masking scheme actually outperform it in downstream tasks. Motivated by the assumption that their lack of complexity plays a vital role in the degradation, we validate whether more complex masked objectives can achieve better results and investigate how much complexity they should have to perform comparably to MLM. Our results using GLUE, SQuAD, and Universal Dependencies benchmarks demonstrate that more complicated objectives tend to show better downstream results with at least half of the MLM complexity needed to perform comparably to MLM. Finally, we discuss how we should pretrain a model using a masked objective from the task complexity perspective.

ThinkSum: Probabilistic reasoning over sets using large language models
Batu M Ozturkler, Nikolay Malkin, Zhen Wang and Nebojsa Jojic   11:00-12:30 (Pier 7&8)
Large language models (LLMs) have a substantial capacity for high-level analogical reasoning: reproducing patterns in linear text that occur in their training data (zero-shot evaluation) or in the provided context (few-shot in-context learning). However, recent studies show that even the more advanced LLMs fail in scenarios that require reasoning over multiple objects or facts and making sequences of logical deductions. We propose a two-stage probabilistic inference paradigm, ThinkSum, which reasons over sets of objects or facts in a structured manner. In the first stage (Think – retrieval of associations), a LLM is queried in parallel over a set of phrases extracted from the prompt or an auxiliary model call. In the second stage (Sum – probabilistic inference or reasoning), the results of these queries are aggregated to make the final prediction. We demonstrate the possibilities and advantages of ThinkSum on the BIG-bench suite of LLM evaluation tasks, achieving improvements over the state of the art using GPT-family models on thirteen difficult tasks, often with far smaller model variants. We also compare and contrast

ThinkSum with other proposed modifications to direct prompting of LLMs, such as variants of chain-of-thought prompting. Our results suggest that because the probabilistic inference in ThinkSum is performed outside of calls to the LLM, ThinkSum is less sensitive to prompt design, yields more interpretable predictions, and can be flexibly combined with latent variable models to extract structured knowledge from LLMs. Overall, our proposed paradigm represents a promising approach for enhancing the reasoning capabilities of LLMs.

Hierarchical Verbalizer for Few-Shot Hierarchical Text Classification
Ke Ji, Yixin Lian, Jingsheng Gao and Baoyuan Wang 11:00-12:30 (Pier 7&8)
Due to the complex label hierarchy and intensive labeling cost in practice, the hierarchical text classification (HTC) suffers a poor performance especially when low-resource or few-shot settings are considered. Recently, there is a growing trend of applying prompts on pre-trained language models (PLMs), which has exhibited effectiveness in the few-shot flat text classification tasks. However, limited work has studied the paradigm of prompt-based learning in the HTC problem when the training data is extremely scarce. In this work, we define a path-based few-shot setting and establish a strict path-based evaluation metric to further explore few-shot HTC tasks. To address the issue, we propose the hierarchical verbalizer ("HierVerb"), a multi-verbalizer framework treating HTC as a single- or multi-label classification problem at multiple layers and learning vectors as verbalizers constrained by hierarchical structure and hierarchical contrastive learning. In this manner, HierVerb fuses label hierarchy knowledge into verbalizers and remarkably outperforms those who inject hierarchy through graph encoders, maximizing the benefits of PLMs. Extensive experiments on three popular HTC datasets under the few-shot settings demonstrate that prompt with HierVerb significantly boosts the HTC performance, meanwhile indicating an elegant way to bridge the gap between the large pre-trained model and downstream hierarchical classification tasks.

Recyclable Tuning for Continual Pre-training
Yujia Qin, Cheng Qian, Xu Han, Yankai Lin, Huadong Wang, Ruobing Xie, Zhiyuan Liu, Maosong Sun and Jie Zhou 11:00-12:30 (Pier 7&8)
Continual pre-training is the paradigm where pre-trained language models (PLMs) continually acquire fresh knowledge from growing data and gradually get upgraded. Before an upgraded PLM is released, we may have tuned the original PLM for various tasks and stored the adapted weights. However, when tuning the upgraded PLM, these outdated adapted weights will typically be ignored and discarded, causing a potential waste of resources. We bring this issue to the forefront and contend that proper algorithms for recycling outdated adapted weights should be developed. To this end, we formulate the task of recyclable tuning for continual pre-training. In pilot studies, we find that after continual pre-training, the upgraded PLM remains compatible with the outdated adapted weights to some extent. Motivated by this finding, we analyze the connection between continually pre-trained PLMs from two novel aspects, i.e., mode connectivity, and functional similarity. Based on the corresponding findings, we propose both an initialization-based method and a distillation-based method for our task. We demonstrate their feasibility in improving the convergence and performance for tuning the upgraded PLM. We also show that both methods can be combined to achieve better performance.

Contrastive Novelty-Augmented Learning: Anticipating Outliers with Large Language Models
Albert Xu, Xiang Ren and Robin Jia 11:00-12:30 (Pier 7&8)
In many task settings, text classification models are likely to encounter examples from novel classes on which they cannot predict correctly. Selective prediction, in which models abstain on low-confidence examples, provides a possible solution, but existing models are often overly confident on unseen classes. To remedy this overconfidence, we introduce Contrastive Novelty-Augmented Learning (CoNAL), a two-step method that generates OOD examples representative of novel classes, then trains to decrease confidence on them. First, we generate OOD examples by prompting a large language model twice: we prompt it to enumerate relevant novel classes, then generate examples from each novel class matching the task format. Second, we train a classifier with a novel contrastive objective that encourages lower confidence on generated OOD examples than training examples. When trained with CoNAL, classifiers improve in their ability to detect and abstain on novel class examples over prior methods by an average of 2.3% in terms of accuracy under the accuracy-coverage curve (AUAC) and 5.5% AUROC across 4 NLP datasets, with no cost to in-distribution accuracy.

The Magic of IF: Investigating Causal Reasoning Abilities in Large Language Models of Code
Xiao Liu, Da Yin, Chen Zhang, Yansong Feng and Dongyan Zhao 11:00-12:30 (Pier 7&8)
Causal reasoning, the ability to identify cause-and-effect relationship, is crucial in human thinking. Although large language models (LLMs) succeed in many NLP tasks, it is still challenging for them to conduct complex causal reasoning like abductive reasoning and counterfactual reasoning. Given the fact that programming code may express causal relations more often and explicitly with conditional statements like "if", we want to explore whether Code-LLMs acquire better causal reasoning abilities. Our experiments show that compared to text-only LLMs, Code-LLMs with code prompts are better causal reasoners. We further intervene on the prompts from different aspects, and discover that the key point is the programming structure. Code and data are available at https://github.com/xxxiaol/magic-if.

Membership Inference Attacks against Language Models via Neighbourhood Comparison
Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan and Taylor Berg-Kirkpatrick 11:00-12:30 (Pier 7&8)
Membership Inference attacks (MIAs) aim to predict whether a data sample was present in the training data of a machine learning model or not, and are widely used for assessing the privacy risks of language models. Most existing attacks rely on the observation that models tend to assign higher probabilities to their training samples than non-training points. However, simple thresholding of the model score in isolation tends to lead to high false-positive rates as it does not account for the intrinsic complexity of a sample. Recent work has demonstrated that reference-based attacks which compare model scores to those obtained from a reference model trained on similar data can substantially improve the performance of MIAs. However, in order to train reference models, attacks of this kind make the strong and arguably unrealistic assumption that an adversary has access to samples closely resembling the original training data. Therefore, we investigate their performance in more realistic scenarios and find that they are highly fragile in relation to the data distribution used to train reference models. To investigate whether this fragility provides a layer of safety, we propose and evaluate neighbourhood attacks, which compare model scores for a given sample to scores of synthetically generated neighbour texts and therefore eliminate the need for access to the training data distribution. We show that, in addition to being competitive with reference-based attacks that have perfect knowledge about the training data distribution, our attack clearly outperforms existing reference-free attacks as well as reference-based attacks with imperfect knowledge, which demonstrates the need for a reevaluation of the threat model of adversarial attacks.

Complementary Explanations for Effective In-Context Learning
Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett and Ramakanth Pasunuru 11:00-12:30 (Pier 7&8)
Large language models (LLMs) have exhibited remarkable capabilities in learning from expla- nations in prompts, but there has been limited understanding of exactly how these explana- tions function or why they are effective. This work aims to better understand the mechanisms by which explanations are used for in-context learning. We first study the impact of two dif- ferent factors on the performance of prompts with explanations: the computation trace (the way the solution is decomposed) and the natural language used to express the prompt. By per- turbing explanations on three controlled tasks, we show that both factors contribute to the ef- fectiveness of explanations. We further study how to form maximally effective sets of expla- nations for solving a given test query. We find that LLMs can benefit from the complemen- tarity of the explanation set: diverse reasoning skills shown by different exemplars can lead to better performance. Therefore, we propose

a maximal marginal relevance-based exemplar selection approach for constructing exemplar sets that are both relevant as well as complementary, which successfully improves the in- context learning performance across three real- world tasks on multiple LLMs.

Aligning Instruction Tasks Unlocks Large Language Models as Zero-Shot Relation Extractors

Kai Zhang, Bernal Jimenez Gutierrez and Yu Su   11:00-12:30 (Pier 7&8)

Recent work has shown that fine-tuning large language models (LLMs) on large-scale instruction-following datasets substantially improves their performance on a wide range of NLP tasks, especially in the zero-shot setting. However, even advanced instruction-tuned LLMs still fail to outperform small LMs on relation extraction (RE), a fundamental information extraction task. We hypothesize that instruction-tuning has been unable to elicit strong RE capabilities in LLMs due to RE's low incidence in instruction-tuning datasets, making up less than 1% of all tasks (Wang et al. 2022). To address this limitation, we propose QA4RE, a framework that aligns RE with question answering (QA), a predominant task in instruction-tuning datasets. Comprehensive zero-shot RE experiments over four datasets with two series of instruction-tuned LLMs (six LLMs in total) demonstrate that our QA4RE framework consistently improves LLM performance, strongly verifying our hypothesis and enabling LLMs to outperform strong zero-shot baselines by a large margin. Additionally, we provide thorough experiments and discussions to show the robustness, few-shot effectiveness, and strong transferability of our QA4RE framework. This work illustrates a promising way of adapting LLMs to challenging and underrepresented tasks by aligning these tasks with more common instruction-tuning tasks like QA.

Model-Generated Pretraining Signals Improves Zero-Shot Generalization of Text-to-Text Transformers

Linyuan Gong, Chenyan Xiong, Xiaodong Liu, Payal Bajaj, Yiqing Xie, Alvin Cheung, Jianfeng Gao and Xia Song   11:00-12:30 (Pier 7&8)

This paper explores the effectiveness of model-generated signals in improving zero-shot generalization of text-to-text Transformers such as T5. We study various designs to pretrain T5 using an auxiliary model to construct more challenging token replacements for the main model to denoise. Key aspects under study include the decoding target, the location of the RTD head, and the masking pattern. Based on these studies, we develop a new model, METRO-T0, which is pretrained using the redesigned ELECTRA-Style pretraining strategies and then prompt-finetuned on a mixture of NLP tasks. METRO-T0 outperforms all similar-sized baselines on prompted NLP benchmarks, such as _T0 Eval_ and MMLU, and rivals the state-of-the-art T0-11B model with only **8%** of its parameters. Our analysis on model's neural activation and parameter sensitivity reveals that the effectiveness of METRO-T0 stems from more balanced contribution of parameters and better utilization of their capacity. The code and model checkpoints are available at https://github.com/gonglinyuan/metro_t0.

MultiInstruct: Improving Multi-Modal Zero-Shot Learning via Instruction Tuning

Zhiyang Xu, Ying Shen and Lifu Huang   11:00-12:30 (Pier 7&8)

Instruction tuning, a new learning paradigm that fine-tunes pre-trained language models on tasks specified through instructions, has shown promising zero-shot performance on various natural language processing tasks. However, it has yet to be explored for vision and multimodal tasks. In this work, we introduce MultiInstruct, the first multimodal instruction tuning benchmark dataset that consists of 62 diverse multimodal tasks in a unified seq-to-seq format covering 10 broad categories. The tasks are derived from 21 existing open-source datasets and each task is equipped with 5 expert-written instructions. We take OFA as the base pre-trained model for multimodal instruction tuning, and to further improve its zero-shot performance, we explore multiple transfer learning strategies to leverage the large-scale Natural Instructions dataset. Experimental results demonstrate strong zero-shot performance on various unseen multimodal tasks and the benefit of transfer learning from a text-only instruction dataset. We also design a new evaluation metric – Sensitivity, to evaluate how sensitive the model is to the variety of instructions. Our results indicate that fine-tuning the model on a diverse set of tasks and instructions leads to a reduced sensitivity to variations in instructions for each task.

Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor

Or Honovich, Thomas Scialom, Omer Levy and Timo Schick   11:00-12:30 (Pier 7&8)

Instruction tuning enables pretrained language models to perform new tasks from inference-time natural language descriptions. These approaches rely on vast amounts of human supervision in the form of crowdsourced datasets or user interactions. In this work, we introduce Unnatural Instructions: a large dataset of creative and diverse instructions, collected with virtually no human labor. We collect 64,000 examples by prompting a language model with three seed examples of instructions and eliciting a fourth. This set is then expanded by prompting the model to rephrase each instruction, creating a total of approximately 240,000 examples of instructions, inputs, and outputs. Experiments show that despite containing a fair amount of noise, training on Unnatural Instructions rivals the effectiveness of training on open-source manually-curated datasets, surpassing the performance of models such as T0++ and Tk-Instruct across various benchmarks. These results demonstrate the potential of model-generated data as a cost-effective alternative to crowdsourcing for dataset expansion and diversification.

Making Language Models Better Reasoners with Step-Aware Verifier

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang LOU and Weizhu Chen   11:00-12:30 (Pier 7&8)

Few-shot learning is a challenging task that requires language models to generalize from limited examples. Large language models like GPT-3 and PaLM have made impressive progress in this area, but they still face difficulties in reasoning tasks such as GSM8K, a benchmark for arithmetic problems. To improve their reasoning skills, previous work has proposed to guide the language model with prompts that elicit a series of reasoning steps before giving the final answer, achieving a significant improvement on GSM8K from 17.9% to 58.1% in problem-solving rate. In this paper, we present DiVeRSe (Diverse Verifier on Reasoning Step), a novel approach that further enhances the reasoning capability of language models. DiVeRSe has three main components: first, it generates diverse prompts to explore different reasoning paths for the same question; second, it uses a verifier to filter out incorrect answers based on a weighted voting scheme; and third, it verifies each reasoning step individually instead of the whole chain. We evaluate DiVeRSe on the latest language model code-davinci-002 and show that it achieves new state-of-the-art results on six of eight reasoning benchmarks (e.g., GSM8K 74.4% to 83.2%).

Black-box language model explanation by context length probing

Ondřej Cífka and Antoine Liutkus   11:00-12:30 (Pier 7&8)

The increasingly widespread adoption of large language models has highlighted the need for improving their explainability. We present *context length probing*, a novel explanation technique for causal language models, based on tracking the predictions of a model as a function of the length of available context, and allowing to assign *differential importance scores* to different contexts. The technique is model-agnostic and does not rely on access to model internals beyond computing token-level probabilities. We apply context length probing to large pre-trained language models and offer some initial analyses and insights, including the potential for studying long-range dependencies. The [source code](https://github.com/cifkao/context-probing/) and an [interactive demo](https://cifkao.github.io/context-probing/) of the method are available.

Revisiting Token Dropping Strategy in Efficient BERT Pretraining

Qihuang Zhong, Liang Ding, Juhua Liu, Xuebo Liu, Min Zhang, Bo Du and Dacheng Tao   11:00-12:30 (Pier 7&8)

Token dropping is a recently-proposed strategy to speed up the pretraining of masked language models, such as BERT, by skipping the computation of a subset of the input tokens at several middle layers. It can effectively reduce the training time without degrading much performance on downstream tasks. However, we empirically find that token dropping is prone to a semantic loss problem and falls short in handling

semantic-intense tasks. Motivated by this, we propose a simple yet effective semantic-consistent learning method (ScTD) to improve the token dropping. ScTD aims to encourage the model to learn how to preserve the semantic information in the representation space. Extensive experiments on 12 tasks show that, with the help of our ScTD, token dropping can achieve consistent and significant performance gains across all task types and model sizes. More encouragingly, ScTD saves up to 57% of pretraining time and brings up to +1.56% average improvement over the vanilla token dropping.

Attribute Controlled Dialogue Prompting
Runcheng Liu, Ahmad Rashid, Ivan Kobyzev, Mehdi Rezagholizadeh and Pascal Poupart 11:00-12:30 (Pier 7&8)
Prompt-tuning has become an increasingly popular parameter-efficient method for adapting large pretrained language models to downstream tasks. However, both discrete prompting and continuous prompting assume fixed prompts for all data samples within a task, neglecting the fact that inputs vary greatly in some tasks such as open-domain dialogue generation. In this paper, we present a novel, instance-specific prompt-tuning algorithm for dialogue generation. Specifically, we generate prompts based on instance-level control code, rather than the conversation history, to explore their impact on controlled dialogue generation. Experiments on popular open-domain dialogue datasets, evaluated on both automated metrics and human evaluation, demonstrate that our method is superior to prompting baselines and comparable to fine-tuning with only 5%-6% of total parameters.

Farewell to Aimless Large-scale Pretraining: Influential Subset Selection for Language Model
Xiao Wang, Weikang Zhou, Qi Zhang, Jie Zhou, SongYang Gao, Junzhe Wang, Menghan Zhang, Xiang Gao, Yun Wen Chen and Tao Gui
11:00-12:30 (Pier 7&8)
Pretrained language models have achieved remarkable success in various natural language processing tasks. However, pretraining has recently shifted toward larger models and larger data, which has resulted in significant computational and energy costs. In this paper, we propose Influence Subset Selection (ISS) for language model, which explicitly utilizes end-task knowledge to select a tiny subset of the pretraining corpus. Specifically, the ISS selects the samples that will provide the most positive influence on the performance of the end task. Furthermore, we design a gradient matching-based influence estimation method, which can drastically reduce the computation time of influence. With only 0.45% of the data and a three-orders-of-magnitude lower computational cost, ISS outperformed pretrained models (e.g., RoBERTa) on eight datasets covering four domains.

Recursion of Thought: A Divide-and-Conquer Approach to Multi-Context Reasoning with Language Models
Soochan Lee and Gunhee Kim 11:00-12:30 (Pier 7&8)
Generating intermediate steps, or Chain of Thought (CoT), is an effective way to significantly improve language models' (LM) multi-step reasoning capability. However, the CoT lengths can grow rapidly with the problem complexity, easily exceeding the maximum context size. Instead of increasing the context limit, which has already been heavily investigated, we explore an orthogonal direction: making LMs divide a problem into multiple contexts. We propose a new inference framework, called Recursion of Thought (RoT), which introduces several special tokens that the models can output to trigger context-related operations. Extensive experiments with multiple architectures including GPT-3 show that RoT dramatically improves LMs' inference capability to solve problems, whose solution consists of hundreds of thousands of tokens.

Learning Better Masking for Better Language Model Pre-training
Dongjie Yang, Zhuosheng Zhang and Hai Zhao 11:00-12:30 (Pier 7&8)
Masked Language Modeling (MLM) has been widely used as the denoising objective in pre-training language models (PrLMs). Existing PrLMs commonly adopt a Random-Token Masking strategy where a fixed masking ratio is applied and different contents are masked by an equal probability throughout the entire training. However, the model may receive complicated impact from pre-training status, which changes accordingly as training time goes on. In this paper, we show that such time-invariant MLM settings on masking ratio and masked content are unlikely to deliver an optimal outcome, which motivates us to explore the influence of time-variant MLM settings. We propose two scheduled masking approaches that adaptively tune the masking ratio and masked content in different training stages, which improves the pre-training efficiency and effectiveness verified on the downstream tasks. Our work is a pioneer study on time-variant masking strategy on ratio and content and gives a better understanding of how masking ratio and masked content influence the MLM pre-training.

UniLG: A Unified Structure-aware Framework for Lyrics Generation
Tao Qian, Fan Lou, Jiatong Shi, Yuning Wu, Shuai Guo, Xiang Yin and Qin Jin 11:00-12:30 (Pier 7&8)
As a special task of natural language generation, conditional lyrics generation needs to consider the structure of generated lyrics and the relationship between lyrics and music. Due to various forms of conditions, a lyrics generation system is expected to generate lyrics conditioned on different signals, such as music scores, music audio, or partially-finished lyrics, etc. However, most of the previous works have ignored the musical attributes hidden behind the lyrics and the structure of the lyrics. Additionally, most works only handle limited lyrics generation conditions, such as lyrics generation based on music score or partial lyrics, they can not be easily extended to other generation conditions with the same framework. In this paper, we propose a unified structure-aware lyrics generation framework named UniLG. Specifically, we design compound templates that incorporate textual and musical information to improve structure modeling and unify the different lyrics generation conditions. Extensive experiments demonstrate the effectiveness of our framework. Both objective and subjective evaluations show significant improvements in generating structural lyrics.

Critic-Guided Decoding for Controlled Text Generation
Minbeom Kim, Hwanhee Lee, Kang Min Yoo, Joonsuk Park, Hwaran Lee and Kyomin Jung 11:00-12:30 (Pier 7&8)
Steering language generation towards objectives or away from undesired content has been a long-standing goal in utilizing language models (LM). Recent work has demonstrated reinforcement learning and weighted decoding as effective approaches to achieve a higher level of language control and quality with pros and cons. In this work, we propose a novel critic decoding method for controlled language generation (CriticControl) that combines the strengths of reinforcement learning and weighted decoding. Specifically, we adopt the actor-critic framework and train an LM-steering critic from reward models. Similar to weighted decoding, our method freezes the language model and manipulates the output token distribution using a critic to improve training efficiency and stability. Evaluation of our method on three controlled generation tasks, topic control, sentiment control, and detoxification, shows that our approach generates more coherent and well-controlled texts than previous methods. In addition, CriticControl demonstrates superior generalization ability in zero-shot settings. Human evaluation studies also corroborate our findings.

Controllable Text Generation via Probability Density Estimation in the Latent Space
Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Lingyuan Zhang, Heng Gong, Weihong Zhong and Bing Qin 11:00-12:30 (Pier 7&8)
Previous work on controllable text generation has explored the idea of control from the latent space, such as optimizing a representation with attribute-specific classifiers or sampling one from relevant discrete samples. However, they cannot effectively model a complex space with diverse attributes, high dimensionality, and asymmetric structure, leaving subsequent controls unsatisfying. In this work, we propose a novel control framework using probability density estimation in the latent space. Our method utilizes an invertible transformation function, the Normalizing Flow, that maps the complex distributions in the latent space to simple Gaussian distributions in the prior space. Thus, we can

perform sophisticated and flexible controls in the prior space and feed the control effects back into the latent space owing to the bijection property of invertible transformations. Experiments on single-attribute and multi-attribute control reveal that our method outperforms several strong baselines on attribute relevance and text quality, achieving a new SOTA. Further analysis of control strength adjustment demonstrates the flexibility of our control strategy.

TAVT: Towards Transferable Audio-Visual Text Generation
Wang Lin, Tao Jin, Wenwen Pan, Linjun Li, Xize Cheng, Ye Wang and Zhou Zhao   11:00-12:30 (Pier 7&8)
Audio-visual text generation aims to understand multi-modality contents and translate them into texts. Although various transfer learning techniques of text generation have been proposed, they focused on uni-modal analysis (e.g. text-to-text, visual-to-text) and lack consideration of multi-modal content and cross-modal relation. Motivated by the fact that humans can recognize the timbre of the same low-level concepts (e.g., footstep, rainfall, and laughing), even in different visual conditions, we aim to mitigate the domain discrepancies by audio-visual correlation. In this paper, we propose a novel Transferable Audio-Visual Text Generation framework, named TAVT, which consists of two key components: Audio-Visual Meta-Mapper (AVMM) and Dual Counterfactual Contrastive Learning (DCCL). (1) AVMM first introduces a universal auditory semantic space and drifts the domain-invariant low-level concepts into visual prefixes. Then the reconstruct-based learning encourages the AVMM to learn "which pixels belong to the same sound" and achieve audio-enhanced visual prefix. The well-trained AVMM can be further applied to uni-modal setting. (2) Furthermore, DCCL leverages the destructive counterfactual transformations to provide cross-modal constraints for AVMM from the perspective of feature distribution and text generation. (3) The experimental results show that TAVT outperforms the state-of-the-art methods across multiple domains (cross-datasets, cross-categories) and various modal settings (uni-modal, multi-modal).

RARR: Researching and Revising What Language Models Say, Using Language Models
Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan and Kelvin Guu   11:00-12:30 (Pier 7&8)
Language models (LMs) now excel at many tasks such as question answering, reasoning, and dialog. However, they sometimes generate unsupported or misleading content. A user cannot easily determine whether their outputs are trustworthy or not, because most LMs do not have any built-in mechanism for attribution to external evidence. To enable attribution while still preserving all the powerful advantages of recent generation models, we propose RARR (Retrofit Attribution using Research and Revision), a system that 1) automatically finds attribution for the output of any text generation model, and 2) post-edits the output to fix unsupported content while preserving the original output as much as possible. When applied to the output of several state-of-the-art LMs on a diverse set of generation tasks, we find that RARR significantly improves attribution while otherwise preserving the original input to a much greater degree than previously explored edit models. Furthermore, the implementation of RARR requires only a handful of training examples, a large language model, and standard web search.

Revisiting Sentence Union Generation as a Testbed for Text Consolidation
Eran Hirsch, Valentina Pyatkin, Ruben Wolhandler, Avi Caciularu, Asi Shefer and Ido Dagan   11:00-12:30 (Pier 7&8)
Tasks involving text generation based on multiple input texts, such as multi-document summarization, long-form question answering and contemporary dialogue applications, challenge models for their ability to properly consolidate partly-overlapping multi-text information. However, these tasks entangle the consolidation phase with the often subjective and ill-defined content selection requirement, impeding proper assessment of models' consolidation capabilities. In this paper, we suggest revisiting the sentence union generation task as an effective well-defined testbed for assessing text consolidation capabilities, decoupling the consolidation challenge from subjective content selection. To support research on this task, we present refined annotation methodology and tools for crowdsourcing sentence union, create the largest union dataset to date and provide an analysis of its rich coverage of various consolidation aspects. We then propose a comprehensive evaluation protocol for union generation, including both human and automatic evaluation. Finally, as baselines, we evaluate state-of-the-art language models on the task, along with a detailed analysis of their capacity to address multi-text consolidation challenges and their limitations.

LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion
Dongfu Jiang, Xiang Ren and Bill Yuchen Lin   11:00-12:30 (Pier 7&8)
We present LLM-Blender, an ensembling framework designed to attain consistently superior performance by leveraging the diverse strengths of multiple open-source large language models (LLMs). Our framework consists of two modules: PairRanker and GenFuser, addressing the observation that optimal LLMs for different examples can significantly vary. PairRanker employs a specialized pairwise comparison method to distinguish subtle differences between candidate outputs. It jointly encodes the input text and a pair of candidates, using cross-attention encoders to determine the superior one. Our results demonstrate that PairRanker exhibits the highest correlation with ChatGPT-based ranking. Then, GenFuser aims to merge the top-ranked candidates, generating an improved output by capitalizing on their strengths and mitigating their weaknesses. To facilitate large-scale evaluation, we introduce a benchmark dataset, MixInstruct, which is a mixture of multiple instruction datasets featuring oracle pairwise comparisons. Our LLM-Blender significantly outperform individual LLMs and baseline methods across various metrics, establishing a substantial performance gap.

Explicit Syntactic Guidance for Neural Text Generation
Yafu Li, Leyang Cui, Jianhao Yan, Yongjing Yin, Wei Bi, Shuming Shi and Yue Zhang   11:00-12:30 (Pier 7&8)
Most existing text generation models follow the sequence-to-sequence paradigm. Generative Grammar suggests that humans generate natural language texts by learning language grammar. We propose a syntax-guided generation schema, which generates the sequence guided by a constituency parse tree in a top-down direction. The decoding process can be decomposed into two parts: (1) predicting the infilling texts for each constituent in the lexicalized syntax context given the source sentence; (2) mapping and expanding each constituent to construct the next-level syntax context. Accordingly, we propose a structural beam search method to find possible syntax structures hierarchically. Experiments on paraphrase generation and machine translation show that the proposed method outperforms autoregressive baselines, while also demonstrating effectiveness in terms of interpretability, controllability, and diversity.

Language Modeling with Latent Situations
Belinda Z. Li, Maxwell Nye and Jacob Andreas   11:00-12:30 (Pier 7&8)
Language models (LMs) often generate incoherent outputs: they refer to events and entity states that are incompatible with the state of the world described in inputs. We introduce SITUATIONSUPERVISION, a family of approaches for improving coherence in LMs by training them to construct and condition on explicit representations of entities and their states. SITUATIONSUPERVISION has two components: an *auxiliary situation modeling* task that trains models to predict entity state representations in context, and a *latent state inference* procedure that imputes these states from partially annotated training data. SITUATIONSUPERVISION can be applied via fine-tuning (by supervising LMs to encode state variables in their hidden representations) and prompting (by inducing LMs to interleave textual descriptions of entity states with output text). In both cases, it requires only a small number of state annotations to produce substantial coherence improvements (up to an 16% reduction in errors), showing that standard LMs can be efficiently adapted to explicitly model language and aspects of its meaning.

Evaluation of Question Generation Needs More References
Shinhyeok Oh, Hyojun Go, Hyeongdon Moon, Yunsung Lee, Myeongho Jeong, Hyun Seung Lee and Seungtaek Choi 11:00-12:30 (Pier 7&8)

Question generation (QG) is the task of generating a valid and fluent question based on a given context and the target answer. According to various purposes, even given the same context, instructors can ask questions about different concepts, and even the same concept can be written in different ways. However, the evaluation for QG usually depends on single reference-based similarity metrics, such as n-gram-based metric or learned metric, which is not sufficient to fully evaluate the potential of QG methods. To this end, we propose to paraphrase the reference question for a more robust QG evaluation. Using large language models such as GPT-3, we created semantically and syntactically diverse questions, then adopt the simple aggregation of the popular evaluation metrics as the final scores. Through our experiments, we found that using multiple (pseudo) references is more effective for QG evaluation while showing a higher correlation with human evaluations than evaluation with a single reference.

Open-ended Long Text Generation via Masked Language Modeling
Xiaobo Liang, Zecheng Tang, Juntao Li and Min Zhang   11:00-12:30 (Pier 7&8)
Pre-trained autoregressive (AR) language models such as BART and GPTs have dominated OPen-ended Long Text Generation (Open-LTG). However, the AR nature will decrease the inference efficiency along with the increase of generation length, which hinder their application in Open-LTG. To improve inference efficiency, we alternatively explore the potential of the pre-trained masked language models (MLMs) along with a representative iterative non-autoregressive (NAR) decoding strategy for Open-LTG. Our preliminary study shows that pre-trained MLMs can merely generate short text and will collapse for long text modeling. To enhance the long text generation capability of MLMs, we introduce two simple yet effective strategies for the iterative NAR model: dynamic sliding window attention (DSWA) and linear temperature decay (LTD). It can alleviate long-distance collapse problems and achieve longer text generation with a flexible trade-off between performance and inference speedup. Experiments on the storytelling and multi-paragraph opinionated article writing tasks show that pre-trained MLMs can achieve more than $3 \times \rightarrow 13 \times$ speedup with better performance than strong AR models.

Focused Prefix Tuning for Controllable Text Generation
Congda Ma, Tianyu Zhao, Makoto Shing, Kei Sawada and Manabu Okumura   11:00-12:30 (Pier 7&8)
In a controllable text generation dataset, there exist unannotated attributes that could provide irrelevant learning signals to models that use it for training and thus degrade their performance. We propose focused prefix tuning (FPT) to mitigate the problem and to enable the control to focus on the desired attribute. Experimental results show that FPT can achieve better control accuracy and text fluency than baseline models in single-attribute control tasks. In multi-attribute control tasks, FPT achieves comparable control accuracy with the state-of-the-art approach while keeping the flexibility to control new attributes without retraining existing models.

Controlled Text Generation with Hidden Representation Transformations
Vaibhav Kumar, Hana Koorehdavoudi, Masud Moshtaghi, Amita Misra, Ankit R Chadha and Emilio Ferrara   11:00-12:30 (Pier 7&8)
We propose CHRT (Control Hidden Representation Transformation) – a con- trolled language generation framework that steers large language models to generate text pertaining to certain attributes (such as toxicity). CHRT gains attribute control by modifying the hidden representation of the base model through learned transformations. We employ a contrastive-learning framework to learn these transformations that can be combined to gain multi-attribute control. The effectiveness of CHRT is experimentally shown by comparing it with seven baselines over three attributes. CHRT outperforms all the baselines in the task of detoxification, positive sentiment steering, and text simplification while minimizing the loss in linguistic qualities. Further, our approach has the lowest inference latency of only 0.01 seconds more than the base model, making it the most suitable for high-performance production environments. We open-source our code and release two novel datasets to further propel controlled language generation research

A New Dataset and Empirical Study for Sentence Simplification in Chinese
Shiping Yang, Renliang Sun and Xiaojun Wan   11:00-12:30 (Pier 7&8)
Sentence Simplification is a valuable technique that can benefit language learners and children a lot. However, current research focuses more on English sentence simplification. The development of Chinese sentence simplification is relatively slow due to the lack of data. To alleviate this limitation, this paper introduces CSS, a new dataset for assessing sentence simplification in Chinese. We collect manual simplifications from human annotators and perform data analysis to show the difference between English and Chinese sentence simplifications. Furthermore, we test several unsupervised and zero/few-shot learning methods on CSS and analyze the automatic evaluation and human evaluation results. In the end, we explore whether Large Language Models can serve as high-quality Chinese sentence simplification systems by evaluating them on CSS.

Unsupervised Summarization Re-ranking
Mathieu Ravaut, Shafiq Joty and Nancy Chen   11:00-12:30 (Pier 7&8)
With the rise of task-specific pre-training objectives, abstractive summarization models like PEGASUS offer appealing zero-shot performance on downstream summarization tasks. However, the performance of such unsupervised models still lags significantly behind their supervised counterparts. Similarly to the supervised setup, we notice a very high variance in quality among summary candidates from these models while only one candidate is kept as the summary output. In this paper, we propose to re-rank summary candidates in an unsupervised manner, aiming to close the performance gap between unsupervised and supervised models. Our approach improves the unsupervised PEGASUS by up to 7.27% and ChatGPT by up to 6.86% relative mean ROUGE across four widely-adopted summarization benchmarks ; and achieves relative gains of 7.51% (up to 23.73% from XSum to WikiHow) averaged over 30 zero-shot transfer setups (finetuning on a dataset, evaluating on another).

Interpretable Automatic Fine-grained Inconsistency Detection in Text Summarization
Hou Pong Chan, Qi Zeng and Heng Ji   11:00-12:30 (Pier 7&8)
Existing factual consistency evaluation approaches for text summarization provide binary predictions and limited insights into the weakness of summarization systems. Therefore, we propose the task of fine-grained inconsistency detection, the goal of which is to predict the fine-grained types of factual errors in a summary. Motivated by how humans inspect factual inconsistency in summaries, we propose an interpretable fine-grained inconsistency detection model, FineGrainFact, which explicitly represents the facts in the documents and summaries with semantic frames extracted by semantic role labeling, and highlights the related semantic frames to predict inconsistency. The highlighted semantic frames help verify predicted error types and correct inconsistent summaries. Experiment results demonstrate that our model outperforms strong baselines and provides evidence to support or refute the summary.

Language Models for German Text Simplification: Overcoming Parallel Data Scarcity through Style-specific Pre-training
Miriam Anschütz, Joshua Oehms, Thomas Wimmer, Bartłomiej Jezierski and Georg Groh   11:00-12:30 (Pier 7&8)
Automatic text simplification systems help to reduce textual information barriers on the internet. However, for languages other than English, only few parallel data to train these systems exists. We propose a two-step approach to overcome this data scarcity issue. First, we fine-tuned language models on a corpus of German Easy Language, a specific style of German. Then, we used these models as decoders in a sequence-to-sequence simplification task. We show that the language models adapt to the style characteristics of Easy Language and output more accessible texts. Moreover, with the style-specific pre-training, we reduced the number of trainable parameters in text simplification models. Hence, less parallel data is sufficient for training. Our results indicate that pre-training on unaligned data can reduce the required

parallel data while improving the performance on downstream tasks.

NonFactS: NonFactual Summary Generation for Factuality Evaluation in Document Summarization

Amir Soleimani, Christof Monz and marcel worring 11:00-12:30 (Pier 7&8)

Pre-trained abstractive summarization models can generate fluent summaries and achieve high ROUGE scores. Previous research has found that these models often generate summaries that are inconsistent with their context document and contain nonfactual information. To evaluate factuality in document summarization, a document-level Natural Language Inference (NLI) classifier can be used. However, training such a classifier requires large-scale high-quality factual and nonfactual samples. To that end, we introduce NonFactS, a data generation model, to synthesize nonfactual summaries given a context document and a human-annotated (reference) factual summary. Compared to previous methods, our nonfactual samples are more abstractive and more similar to their corresponding factual samples, resulting in state-of-the-art performance on two factuality evaluation benchmarks, FALSESUM and SUMMAC. Our experiments demonstrate that even without human-annotated summaries, NonFactS can use random sentences to generate nonfactual summaries and a classifier trained on these samples generalizes to out-of-domain documents.

Reference Matters: Benchmarking Factual Error Correction for Dialogue Summarization with Fine-grained Evaluation Framework

Mingqi Gao, Xiaojun Wan, Jia Su, Zhefeng Wang and baoxing Huai 11:00-12:30 (Pier 7&8)

Factuality is important to dialogue summarization. Factual error correction (FEC) of model-generated summaries is one way to improve factuality. Current FEC evaluation that relies on factuality metrics is not reliable and detailed enough. To address this problem, we are the first to manually annotate a FEC dataset for dialogue summarization containing 4000 items and propose FERRANTI, a fine-grained evaluation framework based on reference correction that automatically evaluates the performance of FEC models on different error categories. Using this evaluation framework, we conduct sufficient experiments with FEC approaches under a variety of settings and find the best training modes and significant differences in the performance of the existing approaches on different factual error categories.

Bridging the Domain Gaps in Context Representations for k-Nearest Neighbor Neural Machine Translation

Zhiwei Cao, Baosong Yang, Huan Lin, Suhang Wu, Xiangpeng Wei, Dayiheng Liu, Jun Xie, Min Zhang and Jinsong Su 11:00-12:30 (Pier 7&8)

k-Nearest neighbor machine translation (kNN-MT) has attracted increasing attention due to its ability to non-parametrically adapt to new translation domains. By using an upstream NMT model to traverse the downstream training corpus, it is equipped with a datastore containing vectorized key-value pairs, which are retrieved during inference to benefit translation. However, there often exists a significant gap between upstream and downstream domains, which hurts the datastore retrieval and the final translation quality. To deal with this issue, we propose a novel approach to boost the datastore retrieval of kNN-MT by reconstructing the original datastore. Concretely, we design a reviser to revise the key representations, making them better fit for the downstream domain. The reviser is trained using the collected semantically-related key-queries pairs, and optimized by two proposed losses: one is the key-queries semantic distance ensuring each revised key representation is semantically related to its corresponding queries, and the other is an L2-norm loss encouraging revised key representations to effectively retain the knowledge learned by the upstream NMT model. Extensive experiments on domain adaptation tasks demonstrate that our method can effectively boost the datastore retrieval and translation quality of kNN-MT.Our code is available at https://github.com/DeepLearnXMU/Revised-knn-mt.

Understanding and Improving the Robustness of Terminology Constraints in Neural Machine Translation

Huaao Zhang, Qiang Wang, Bo - Qin, Zelin Shi, haibo wang and MING CHEN 11:00-12:30 (Pier 7&8)

In this work, we study the robustness of two typical terminology translation methods: Placeholder (PH) and Code-Switch (CS), concerning (1) the number of constraints and (2) the target constraint length. We identify that existing terminology constraint test sets, such as IATE, Wiktionary, and TICO, are blind to this issue due to oversimplified constraint settings. To solve it, we create a new challenging test set of English-German, increasing the average constraint count per sentence from 1.1∼1.7 to 6.1 and the length per target constraint from 1.1∼1.2 words to 3.4 words. Then we find that PH and CS methods degrade as the number of constraints increases, but they have complementary strengths. Specifically, PH is better at retaining high constraint accuracy but lower translation quality as measured by BLEU and COMET scores. In contrast, CS has the opposite results. Based on these observations, we propose a simple but effective method combining the advantages of PH and CS. This approach involves training a model like PH to predict the term labels, and then during inference replacing those labels with target terminology text like CS, so that the subsequent generation is aware of the target term content. Extensive experimental results show that this approach can achieve high constraint accuracy and translation quality simultaneously, regardless of the number or length of constraints.

Target-Side Augmentation for Document-Level Machine Translation

Guangsheng Bao, ZHIYANG TENG and Yue Zhang 11:00-12:30 (Pier 7&8)

Document-level machine translation faces the challenge of data sparsity due to its long input length and a small amount of training data, increasing the risk of learning spurious patterns. To address this challenge, we propose a target-side augmentation method, introducing a data augmentation (DA) model to generate many potential translations for each source document. Learning on these wider range translations, an MT model can learn a smoothed distribution, thereby reducing the risk of data sparsity. We demonstrate that the DA model, which estimates the posterior distribution, largely improves the MT performance, outperforming the previous best system by 2.30 s-BLEU on News and achieving new state-of-the-art on News and Europarl benchmarks.

Towards Speech Dialogue Translation Mediating Speakers of Different Languages

Shuichiro Shimizu, Chenhui Chu, Sheng Li and Sadao Kurohashi 11:00-12:30 (Pier 7&8)

We present a new task, speech dialogue translation mediating speakers of different languages. We construct the SpeechBSD dataset for the task and conduct baseline experiments. Furthermore, we consider context to be an important aspect that needs to be addressed in this task and propose two ways of utilizing context, namely monolingual context and bilingual context. We conduct cascaded speech translation experiments using Whisper and mBART, and show that bilingual context performs better in our settings.

DUB: Discrete Unit Back-translation for Speech Translation

Dong Zhang, Rong Ye, Tom Ko, Mingxuan Wang and Yaqian Zhou 11:00-12:30 (Pier 7&8)

How can speech-to-text translation (ST) perform as well as machine translation (MT)? The key point is to bridge the modality gap between speech and text so that useful MT techniques can be applied to ST. Recently, the approach of representing speech with unsupervised discrete units yields a new way to ease the modality problem. This motivates us to propose Discrete Unit Back-translation(DUB) to answer two questions (1) Is it better to represent speech with discrete units than with continuous features in direct ST? (2) How much benefit can useful MT techniques bring to ST? With DUB, the back-translation technique can successfully be applied on direct ST and obtains an average boost of 5.5 BLEU on MuST-C En-De/Fr/Es. In the low-resource language scenario, our method achieves comparable performance to existing methods that rely on large-scale external data. Code and models are available at https://anonymous.4open.science/r/DUB/.

Synthetic Pre-Training Tasks for Neural Machine Translation

Zexue He, Graeme Blackwood, Rameswar Panda, Julian McAuley and Rogerio Feris   11:00-12:30 (Pier 7&8)

Pre-training models with large crawled corpora can lead to issues such as toxicity and bias, as well as copyright and privacy concerns. A promising way of alleviating such concerns is to conduct pre-training with synthetic tasks and data, since no real-world information is ingested by the model. Our goal in this paper is to understand the factors that contribute to the effectiveness of pre-training models when using synthetic resources, particularly in the context of neural machine translation. We propose several novel approaches to pre-training translation models that involve different levels of lexical and structural knowledge, including: 1) generating obfuscated data from a large parallel corpus 2) concatenating phrase pairs extracted from a small word-aligned corpus, and 3) generating synthetic parallel data without real human language corpora. Our experiments on multiple language pairs reveal that pre-training benefits can be realized even with high levels of obfuscation or purely synthetic parallel data. We hope the findings from our comprehensive empirical analysis will shed light on understanding what matters for NMT pre-training, as well as pave the way for the development of more efficient and less toxic models.

Do GPTs Produce Less Literal Translations?

Vikas Raunak, Arul Menezes, Matt Post and Hany Hassan   11:00-12:30 (Pier 7&8)

Large Language Models (LLMs) such as GPT-3 have emerged as general-purpose language models capable of addressing many natural language generation or understanding tasks. On the task of Machine Translation (MT), multiple works have investigated few-shot prompting mechanisms to elicit better translations from LLMs. However, there has been relatively little investigation on how such translations differ qualitatively from the translations generated by standard Neural Machine Translation (NMT) models. In this work, we investigate these differences in terms of the literalness of translations produced by the two systems. Using literalness measures involving word alignment and monotonicity, we find that translations out of English (E-X) from GPTs tend to be less literal, while exhibiting similar or better scores on MT quality metrics. We demonstrate that this finding is borne out in human evaluations as well. We then show that these differences are especially pronounced when translating sentences that contain idiomatic expressions.

Back Translation for Speech-to-text Translation Without Transcripts

Qingkai Fang and Yang Feng   11:00-12:30 (Pier 7&8)

The success of end-to-end speech-to-text translation (ST) is often achieved by utilizing source transcripts, e.g., by pre-training with automatic speech recognition (ASR) and machine translation (MT) tasks, or by introducing additional ASR and MT data. Unfortunately, transcripts are only sometimes available since numerous unwritten languages exist worldwide. In this paper, we aim to utilize large amounts of target-side monolingual data to enhance ST without transcripts. Motivated by the remarkable success of back translation in MT, we develop a back translation algorithm for ST (BT4ST) to synthesize pseudo ST data from monolingual target data. To ease the challenges posed by short-to-long generation and one-to-many mapping, we introduce self-supervised discrete units and achieve back translation by cascading a target-to-unit model and a unit-to-speech model. With our synthetic ST data, we achieve an average boost of 2.3 BLEU on MuST-C En-De, En-Fr, and En-Es datasets. More experiments show that our method is especially effective in low-resource scenarios.

Understanding and Bridging the Modality Gap for Speech Translation

Qingkai Fang and Yang Feng   11:00-12:30 (Pier 7&8)

How to achieve better end-to-end speech translation (ST) by leveraging (text) machine translation (MT) data? Among various existing techniques, multi-task learning is one of the effective ways to share knowledge between ST and MT in which additional MT data can help to learn source-to-target mapping. However, due to the differences between speech and text, there is always a gap between ST and MT. In this paper, we first aim to understand this modality gap from the target-side representation differences, and link the modality gap to another well-known problem in neural machine translation: exposure bias. We find that the modality gap is relatively small during training except for some difficult cases, but keeps increasing during inference due to the cascading effect. To address these problems, we propose the Cross-modal Regularization with Scheduled Sampling (Cress) method. Specifically, we regularize the output predictions of ST and MT, whose target-side contexts are derived by sampling between ground truth words and self-generated words with a varying probability. Furthermore, we introduce token-level adaptive training which assigns different training weights to target tokens to handle difficult cases with large modality gaps. Experiments and analysis show that our approach effectively bridges the modality gap, and achieves significant improvements over a strong baseline in all eight directions of the MuST-C dataset.

Rethinking the Word-level Quality Estimation for Machine Translation from Human Judgement

Zhen Yang, Fandong Meng, Yuanmeng Yan and Jie Zhou   11:00-12:30 (Pier 7&8)

Word-level Quality Estimation (QE) of Machine Translation (MT) aims to detect potential translation errors in the translated sentence without reference. Typically, conventional works on word-level QE are usually designed to predict the quality of translated words in terms of the post-editing effort, where the word labels in the dataset, i.e., OK or BAD, are automatically generated by comparing words between MT sentences and the post-edited sentences through a Translation Error Rate (TER) toolkit. While the post-editing effort can be used to measure the translation quality to some extent, we find it usually conflicts with human judgment on whether the word is well or poorly translated. To investigate this conflict, we first create a golden benchmark dataset, namely *HJQE* (Human Judgement on Quality Estimation), where the source and MT sentences are identical to the original TER-based dataset and the expert translators directly annotate the poorly translated words on their judgments. Based on our analysis, we further propose two tag-correcting strategies which can make the TER-based artificial QE corpus closer to *HJQE*. We conduct substantial experiments based on the publicly available WMT En-De and En-Zh corpora. The results not only show our proposed dataset is more consistent with human judgment but also confirm the effectiveness of the proposed tag-correcting strategies.For reviewers, the corpora and codes can be found in the attached files.

"A Little is Enough": Few-Shot Quality Estimation based Corpus Filtering improves Machine Translation

Akshay Batheja and Pushpak Bhattacharyya   11:00-12:30 (Pier 7&8)

Quality Estimation (QE) is the task of evaluating the quality of a translation when reference translation is not available. The goal of QE aligns with the task of corpus filtering, where we assign the quality score to the sentence pairs present in the pseudo-parallel corpus. We propose a Quality Estimation based Filtering approach to extract high-quality parallel data from the pseudo-parallel corpus. To the best of our knowledge, this is a novel adaptation of QE framework to extracting quality parallel corpus from the pseudo-parallel corpus.. By training with this filtered corpus, we observe an improvement in the Machine Translation (MT) system's performance by up to 1.8 BLEU points, for English-Marathi, Chinese-English, and Hindi-Bengali language pairs, over the baseline model. The baseline model is the one that is trained on the whole pseudo-parallel corpus. Our Few-shot QE model transfer learned from the English-Marathi QE model and fine-tuned on only 500 Hindi-Bengali training instances, shows an improvement of up to 0.6 BLEU points for Hindi-Bengali language pair, compared to the baseline model. This demonstrates the promise of transfer learning in the setting under discussion. QE systems typically require in the order of (7K-25K) of training data. Our Hindi-Bengali QE is trained on only 500 instances of training that is 1/40th of the normal requirement and achieves comparable performance. All the scripts and datasets utilized in this study will be publicly available.

A Holistic Approach to Reference-Free Evaluation of Machine Translation

Hanming Wu, Wenjuan Han, Hui Di, Yufeng Chen and Jinan Xu   11:00-12:30 (Pier 7&8)

Traditional machine translation evaluation relies on reference written by humans. While reference-free evaluation gets rid of the constraints of labor-intensive annotations, which can pivot easily to new domains and is more scalable. In this paper, we propose a reference-free evaluation

approach that characterizes evaluation as two aspects: (1) fluency: how well the translated text conforms to normal human language usage; (2) faithfulness: how well the translated text reflects the source data. We further split the faithfulness into word-level and sentence-level. Extensive experiments spanning WMT18/19/21 Metrics segment-level daRR and MQM datasets demonstrate that our proposed reference-free approach, ReFreeEval, outperforms SOTA reference-fee metrics like YiSi-2.

A Class-Rebalancing Self-Training Framework for Distantly-Supervised Named Entity Recognition
Qi Li, Tingyu Xie, Peng Peng, Hongwei Wang and Gaoang Wang   11:00-12:30 (Pier 7&8)
Distant supervision reduces the reliance on human annotation in the named entity recognition tasks. The class-level imbalanced distant annotation is a realistic and unexplored problem, and the popular method of self-training can not handle class-level imbalanced learning. More importantly, self-training is dominated by the high-performance class in selecting candidates, and deteriorates the low-performance class with the bias of generated pseudo label. To address the class-level imbalance performance, we propose a class-rebalancing self-training framework for improving the distantly-supervised named entity recognition. In candidate selection, a class-wise flexible threshold is designed to fully explore other classes besides the high-performance class. In label generation, injecting the distant label, a hybrid pseudo label is adopted to provide straight semantic information for the low-performance class. Experiments on five flat and two nested datasets show that our model achieves state-of-the-art results. We also conduct extensive research to analyze the effectiveness of the flexible threshold and the hybrid pseudo label.

Revisiting Event Argument Extraction: Can EAE Models Learn Better When Being Aware of Event Co-occurrences?
Yuxin He, Jingyue Hu and Buzhou Tang   11:00-12:30 (Pier 7&8)
Event co-occurrences have been proved effective for event extraction (EE) in previous studies, but have not been considered for event argument extraction (EAE) recently. In this paper, we try to fill this gap between EE research and EAE research, by highlighting the question that *"Can EAE models learn better when being aware of event co-occurrences?"*. To answer this question, we reformulate EAE as a problem of table generation and extend a SOTA prompt-based EAE model into a non-autoregressive generation framework, called TabEAE, which is able to extract the arguments of multiple events in parallel. Under this framework, we experiment with 3 different training-inference schemes on 4 datasets (ACE05, RAMS, WikiEvents and MLEE) and discover that via training the model to extract all events in parallel, it can better distinguish the semantic boundary of each event and its ability to extract single event gets substantially improved. Experimental results show that our method achieves new state-of-the-art performance on the 4 datasets. Our code is avilable at https://github.com/Stardust-hyx/TabEAE.

PromptRank: Unsupervised Keyphrase Extraction Using Prompt
Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun and Xiaoyan Bai   11:00-12:30 (Pier 7&8)
The keyphrase extraction task refers to the automatic selection of phrases from a given document to summarize its core content. State-of-the-art (SOTA) performance has recently been achieved by embedding-based algorithms, which rank candidates according to how similar their embeddings are to document embeddings. However, such solutions either struggle with the document and candidate length discrepancies or fail to fully utilize the pre-trained language model (PLM) without further fine-tuning. To this end, in this paper, we propose a simple yet effective unsupervised approach, PromptRank, based on the PLM with an encoder-decoder architecture. Specifically, PromptRank feeds the document into the encoder and calculates the probability of generating the candidate with a designed prompt by the decoder. We extensively evaluate the proposed PromptRank on six widely used benchmarks. PromptRank outperforms the SOTA approach MDERank, improving the F1 score relatively by 34.18%, 24.87%, and 17.57% for 5, 10, and 15 returned results, respectively. This demonstrates the great potential of using prompt for unsupervised keyphrase extraction. We release our code at https://github.com/HLT-NLP/PromptRank.

LayoutMask: Enhance Text-Layout Interaction in Multi-modal Pre-training for Document Understanding
Yi Tu, Ya Guo, Huan Chen and jinyang tang   11:00-12:30 (Pier 7&8)
Visually-rich Document Understanding (VrDU) has attracted much research attention over the past years. Pre-trained models on a large number of document images with transformer-based backbones have led to significant performance gains in this field. The major challenge is how to fusion the different modalities (text, layout, and image) of the documents in a unified model with different pre-training tasks. This paper focuses on improving text-layout interactions and proposes a novel multi-modal pre-training model, LayoutMask. LayoutMask uses local 1D position, instead of global 1D position, as layout input and has two pre-training objectives: (1) Masked Language Modeling: predicting masked tokens with two novel masking strategies; (2) Masked Position Modeling: predicting masked 2D positions to improve layout representation learning. LayoutMask can enhance the interactions between text and layout modalities in a unified model and produce adaptive and robust multi-modal representations for downstream tasks. Experimental results show that our proposed method can achieve state-of-the-art results on a wide variety of VrDU problems, including form understanding, receipt understanding, and document image classification.

Joint Constrained Learning with Boundary-adjusting for Emotion-Cause Pair Extraction
Huawen Feng, Junlong Liu, Junhao Zheng, Haibin Chen, Xichen Shang and Qianli Ma   11:00-12:30 (Pier 7&8)
Emotion-Cause Pair Extraction (ECPE) aims to identify the document's emotion clauses and corresponding cause clauses. Like other relation extraction tasks, ECPE is closely associated with the relationship between sentences. Recent methods based on Graph Convolutional Networks focus on how to model the multiplex relations between clauses by constructing different edges. However, the data of emotions, causes, and pairs are extremely unbalanced, and current methods get their representation using the same graph structure. In this paper, we propose a **J**oint **C**onstrained Learning framework with **B**oundary-adjusting for Emotion-Cause Pair Extraction (**JCB**). Specifically, through constrained learning, we summarize the prior rules existing in the data and force the model to take them into consideration in optimization, which helps the model learn a better representation from unbalanced data. Furthermore, we adjust the decision boundary of classifiers according to the relations between subtasks, which have always been ignored. No longer working independently as in the previous framework, the classifiers corresponding to three subtasks cooperate under the relation constraints. Experimental results show that **JCB** obtains competitive results compared with state-of-the-art methods and prove its robustness on unbalanced data.

UniEX: An Effective and Efficient Framework for Unified Information Extraction via a Span-extractive Perspective
yang ping, JunYu Lu, ruyi gan, Junjie Wang, Yuxiang Zhang, Pingjian Zhang and Jiaxing Zhang   11:00-12:30 (Pier 7&8)
We propose a new paradigm for universal information extraction (IE) that is compatible with any schema format and applicable to a list of IE tasks, such as named entity recognition, relation extraction, event extraction and sentiment analysis. Our approach converts the text-based IE tasks as the token-pair problem, which uniformly disassembles all extraction targets into joint span detection, classification and association problems with a unified extractive framework, namely UniEX. UniEX can synchronously encode schema-based prompt and textual information, and collaboratively learn the generalized knowledge from pre-defined information using the auto-encoder language models. We develop a traffine attention mechanism to integrate heterogeneous factors including tasks, labels and inside tokens, and obtain the extraction target via a scoring matrix. Experiment results show that UniEX can outperform generative universal IE models in terms of performance and inference-speed on 14 benchmarks IE datasets with the supervised setting. The state-of-the-art performance in low-resource scenarios also verifies the transferability and effectiveness of UniEX.

Enhancing Event Causality Identification with Event Causal Label and Event Pair Interaction Graph
Ruili Pu, Yang Li, Suge Wang, Deyu Li, Jianxing Zheng and Jian Liao   11:00-12:30 (Pier 7&8)

Most existing event causality identification (ECI) methods rarely consider the event causal label information and the interaction information between event pairs. In this paper, we propose a framework to enrich the representation of event pairs by introducing the event causal label information and the event pair interaction information. In particular, 1) we design an event-causal-label-aware module to model the event causal label information, in which we design the event causal label prediction task as an auxiliary task of ECI, aiming to predict which events are involved in the causal relationship (we call them causality-related events) by mining the dependencies between events. 2) We further design an event pair interaction graph module to model the interaction information between event pairs, in which we construct the interaction graph with event pairs as nodes and leverage graph attention mechanism to model the degree of dependency between event pairs. The experimental results show that our approach outperforms previous state-of-the-art methods on two benchmark datasets EventStoryLine and Causal-TimeBank.

A Diffusion Model for Event Skeleton Generation

Fangqi Zhu, Lin Zhang, Jun Gao, Bing Qin, Ruifeng Xu and Haiqin Yang   11:00-12:30 (Pier 7&8)
Event skeleton generation, aiming to induce an event schema skeleton graph with abstracted event nodes and their temporal relations from a set of event instance graphs, is a critical step in the temporal complex event schema induction task. Existing methods effectively address this task from a graph generation perspective but suffer from noise-sensitive and error accumulation, e.g., the inability to correct errors while generating schema. We, therefore, propose a novel Diffusion Event Graph Model (DEGM) to address these issues. Our DEGM is the first workable diffusion model for event skeleton generation, where the embedding and rounding techniques with a custom edge-based loss are introduced to transform a discrete event graph into learnable latent representations. Furthermore, we propose a denoising training process to maintain the model's robustness. Consequently, DEGM derives the final schema, where error correction is guaranteed by iteratively refining the latent representations during the schema generation process. Experimental results on three IED bombing datasets demonstrate that our DEGM achieves better results than other state-of-the-art baselines. Our code and data are available at https://github.com/zhufq00/EventSkeletonGeneration.

Towards Better Entity Linking with Multi-View Enhanced Distillation

*Yi Liu, Yuan Tian, Jianxun Lian, xinlong wang, Yanan Cao, Fang Fang, Wen Zhang, Haizhen Huang, Weiwei Deng and Qi Zhang*11:00-12:30 (Pier 7&8)
Dense retrieval is widely used for entity linking to retrieve entities from large-scale knowledge bases. Mainstream techniques are based on a dual-encoder framework, which encodes mentions and entities independently and calculates their relevances via rough interaction metrics, resulting in difficulty in explicitly modeling multiple mention-relevant parts within entities to match divergent mentions. Aiming at learning entity representations that can match divergent mentions, this paper proposes a Multi-View Enhanced Distillation (MVD) framework, which can effectively transfer knowledge of multiple fine-grained and mention-relevant parts within entities from cross-encoders to dual-encoders. Each entity is split into multiple views to avoid irrelevant information being over-squashed into the mention-relevant view. We further design cross-alignment and self-alignment mechanisms for this framework to facilitate fine-grained knowledge distillation from the teacher model to the student model. Meanwhile, we reserve a global-view that embeds the entity as a whole to prevent dispersal of uniform information. Experiments show our method achieves state-of-the-art performance on several entity linking benchmarks.

Benchmarking Diverse-Modal Entity Linking with Generative Models

Sijia Wang, Alexander Hanbo Li, Henghui Zhu, Sheng Zhang, Pramuditha Perera, Chung-Wei Hang, Jie Ma, William Yang Wang, Zhiguo Wang, Vittorio Castelli, Bing Xiang and Patrick Ng   11:00-12:30 (Pier 7&8)
Entities can be expressed in diverse formats, such as texts, images, or column names and cell values in tables. While existing entity linking (EL) models work well on per modality configuration, such as text-only EL, visual grounding or schema linking, it is more challenging to design a unified model for diverse modality configurations. To bring various modality configurations together, we constructed a benchmark for diverse-modal EL (DMEL) from existing EL datasets, covering all three modalities including text, image and table. To approach the DMEL task, we proposed a generative diverse-modal model (GDMM) following a multimodal-encoder-decoder paradigm. Pre-training GDMM with rich corpora builds a solid foundation for DMEL without storing the entire KB for inference. Fine-tuning GDMM builds a stronger DMEL baseline, outperforming state-of-the-art task-specific EL models by 8.51 F1 score on average. Additionally, extensive error analyses are conducted to highlight the challenge of DMEL, facilitating future researches on this task.

Concept2Box: Joint Geometric Embeddings for Learning Two-View Knowledge Graphs

Zijie Huang, Daheng Wang, Binxuan Huang, Chenwei Zhang, Jingbo Shang, Yan Liang, Zhengyang Wang, Xian Li, Christos Faloutsos, Yizhou Sun and Wei Wang   11:00-12:30 (Pier 7&8)
Knowledge graph embeddings (KGE) have been extensively studied to embed large-scale relational data for many real-world applications. Existing methods have long ignored the fact many KGs contain two fundamentally different views: high-level ontology-view concepts and fine-grained instance-view entities. They usually embed all nodes as vectors in one latent space. However, a single geometric representation fails to capture the structural differences between two views and lacks probabilistic semantics towards concepts' granularity. We propose Concept2Box, a novel approach that jointly embeds the two views of a KG using dual geometric representations. We model concepts with box embeddings, which learn the hierarchy structure and complex relations such as overlap and disjoint among them. Box volumes can be interpreted as concepts' granularity. Different from concepts, we model entities as vectors. To bridge the gap between concept box embeddings and entity vector embeddings, we propose a novel vector-to-box distance metric and learn both embeddings jointly. Experiments on both the public DBpedia KG and a newly-created industrial KG showed the effectiveness of Concept2Box.

Understanding Demonstration-based Learning from a Causal Perspective

Ruiyi Zhang and Tong Yu   11:00-12:30 (Pier 7&8)
Demonstration-based learning has shown impressive performance in exploiting pretrained language models under few-shot learning settings. It is interesting to see that demonstrations, even those composed of random tokens, can still improve performance. In this paper, we build a Structural Causal Model (SCM) to understand demonstration-based learning from causal perspectives and interpret random demonstrations as interventions on the demonstration variable within the causal model. We investigate the causal effects and find that the concurrence of specific words in the demonstration will induce bias, while randomly sampled tokens in the demonstration do not. Based on this finding, we further propose simple ways to construct random demonstrations, which even outperform hand-crafted, meaningful demonstrations on public sequence labeling benchmarks.

FSUIE: A Novel Fuzzy Span Mechanism for Universal Information Extraction

Tianshuo Peng, Zuchao Li, Lefei Zhang, Bo Du and Hai Zhao   11:00-12:30 (Pier 7&8)
Universal Information Extraction (UIE) has been introduced as a unified framework for various Information Extraction (IE) tasks and has achieved widespread success. Despite this, UIE models have limitations. For example, they rely heavily on span boundaries in the data during training, which does not reflect the reality of span annotation challenges. Slight adjustments to positions can also meet requirements. Additionally, UIE models lack attention to the limited span length feature in IE. To address these deficiencies, we propose the Fuzzy Span Universal Information Extraction (FSUIE) framework. Specifically, our contribution consists of two concepts: *fuzzy span loss* and *fuzzy span attention*. Our experimental results on a series of main IE tasks show significant improvement compared to the baseline, especially in terms of fast convergence and strong performance with small amounts of data and training epochs. These results demonstrate the effectiveness and

generalization of FSUIE in different tasks, settings, and scenarios.

Learning In-context Learning for Named Entity Recognition

Jiawei Chen, Yaojie Lu, Hongyu Lin, Jie Lou, Wei Jia, Dai Dai, Hua Wu, Boxi Cao, Xianpei Han and Le Sun   11:00-12:30 (Pier 7&8)
Named entity recognition in real-world applications suffers from the diversity of entity types, the emergence of new entity types, and the lack of high-quality annotations. To address the above problems, this paper proposes an in-context learning-based NER approach, which can effectively inject in-context NER ability into PLMs and recognize entities of novel types on-the-fly using only a few demonstrative instances. Specifically, we model PLMs as a meta-function and a new entity extractor can be implicitly constructed by applying new instruction and demonstrations to PLMs, i.e., (Lambda . M) (instruction, demonstrations) ->F where F will be a new entity extractor F: text -> entities. To inject the above in-context NER ability into PLMs, we propose a meta-function pre-training algorithm, which pre-trains PLMs by comparing the (instruction, demonstration)-initialized extractor with a surrogate golden extractor. Experimental results on 4 few-shot NER datasets show that our method can effectively inject in-context NER ability into PLMs and significantly outperforms the PLMs+fine-tuning counterparts.

Class-Adaptive Self-Training for Relation Extraction with Incompletely Annotated Training Data

Qingyu Tan, Lu Xu, Lidong Bing and Hwee Tou Ng   11:00-12:30 (Pier 7&8)
Relation extraction (RE) aims to extract relations from sentences and documents. Existing relation extraction models typically rely on supervised machine learning. However, recent studies showed that many RE datasets are incompletely annotated. This is known as the false negative problem in which valid relations are falsely annotated as 'no_relation'. Models trained with such data inevitably make similar mistakes during the inference stage. Self-training has been proven effective in alleviating the false negative problem. However, traditional self-training is vulnerable to confirmation bias and exhibits poor performance in minority classes. To overcome this limitation, we proposed a novel class-adaptive re-sampling self-training framework. Specifically, we re-sampled the pseudo-labels for each class by precision and recall scores. Our re-sampling strategy favored the pseudo-labels of classes with high precision and low recall, which improved the overall recall without significantly compromising precision. We conducted experiments on document-level and biomedical relation extraction datasets, and the results showed that our proposed self-training framework consistently outperforms existing competitive methods on the Re-DocRED and ChemDisgene datasets when the training data are incompletely annotated.

Double-Branch Multi-Attention based Graph Neural Network for Knowledge Graph Completion

Hongcai Xu, Junpeng Bao and Wenbo Liu   11:00-12:30 (Pier 7&8)
Graph neural networks (GNNs), which effectively use topological structures in the knowledge graphs (KG) to embed entities and relations in low-dimensional spaces, have shown great power in knowledge graph completion (KGC). KG has abundant global and local structural information, however, many GNN-based KGC models cannot capture these two types of information about the graph structure by designing complex aggregation schemes, and are not designed well to learn representations of seen entities with sparse neighborhoods in isolated subgraphs. In this paper, we find that a simple attention-based method can outperform a general GNN-based approach for KGC. We then propose a double-branch multi-attention based graph neural network (MA-GNN) to learn more expressive entity representations which contain rich global-local structural information. Specifically, we first explore the graph attention network-based local aggregator to learn entity representations. Furthermore, we propose a snowball local attention mechanism by leveraging the semantic similarity between two-hop neighbors to enrich the entity embedding. Finally, we use Transformer-based self-attention to learn long-range dependence between entities to obtain richer representations with the global graph structure and entity features. Experimental results on five benchmark datasets show that MA-GNN achieves significant improvements over strong baselines for inductive KGC.

Recall, Expand, and Multi-Candidate Cross-Encode: Fast and Accurate Ultra-Fine Entity Typing

Chengyue Jiang, Wenyang Hui, Yong Jiang, Xiaobin Wang, Pengjun Xie and Kewei Tu   11:00-12:30 (Pier 7&8)
Ultra-fine entity typing (UFET) predicts extremely free-formed types (e.g., *president, politician*) of a given entity mention (e.g., *Joe Biden*) in context. State-of-the-art (SOTA) methods use the cross-encoder (CE) based architecture. CE concatenates a mention (and its context) with each type and feeds the pair into a pretrained language model (PLM) to score their relevance. It brings deeper interaction between the mention and the type to reach better performance but has to perform N (the type set size) forward passes to infer all the types of a single mention. CE is therefore very slow in inference when the type set is large (e.g., $N = 10k$ for UFET). % Cross-encoder also ignores the correlation between different types. To this end, we propose to perform entity typing in a recall-expand-filter manner. The recall and expansion stages prune the large type set and generate K (typically much smaller than N) most relevant type candidates for each mention. At the filter stage, we use a novel model called {pasted macro 'NAME'} to concurrently encode and score all these K candidates in only one forward pass to obtain the final type prediction. We investigate different model options for each stage and conduct extensive experiments to compare each option. experiments show that our method reaches SOTA performance on UFET and is thousands of times faster than the CE-based architecture. We also found our method is very effective in fine-grained (130 types) and coarse-grained (9 types) entity typing. Our code is available at {pasted macro 'CODE'}.

CONE: An Efficient COarse-to-fiNE Alignment Framework for Long Video Temporal Grounding

Zhijian Hou, Wanjun Zhong, Lei Ji, DIFEI GAO, Kun Yan, W.K. Chan, Chong-Wah Ngo, Mike Zheng Shou and Nan Duan 11:00-12:30 (Pier 7&8)
This paper tackles an emerging and challenging problem of long video temporal grounding (VTG) that localizes video moments related to a natural language (NL) query. Compared with short videos, long videos are also highly demanded but less explored, which brings new challenges in higher inference computation cost and weaker multi-modal alignment. To address these challenges, we propose CONE, an efficient COarse-to-fiNE alignment framework. CONE is a plug-and-play framework on top of existing VTG models to handle long videos through a sliding window mechanism. Specifically, CONE (1) introduces a query-guided window selection strategy to speed up inference, and (2) proposes a coarse-to-fine mechanism via a novel incorporation of contrastive learning to enhance multi-modal alignment for long videos. Extensive experiments on two large-scale long VTG benchmarks consistently show both substantial performance gains (e.g., from 3.13 to 6.87% on MAD) and state-of-the-art results. Analyses also reveal higher efficiency as the query-guided window selection mechanism accelerates inference time by 2x on Ego4D-NLQ and 15x on MAD while keeping SOTA results. Codes have been released at https://github.com/houzhijian/CONE.

Cross-View Language Modeling: Towards Unified Cross-Lingual Cross-Modal Pre-training

Yan Zeng, Wangchunshu Zhou, Ao Luo, Ziming Cheng and Xinsong Zhang   11:00-12:30 (Pier 7&8)
In this paper, we introduce Cross-View Language Modeling, a simple and effective pre-training framework that unifies cross-lingual and cross-modal pre-training with shared architectures and objectives. Our approach is motivated by a key observation that cross-lingual and cross-modal pre-training share the same goal of aligning two different views of the same object into a common semantic space. To this end, the cross-view language modeling framework considers both multi-modal data (i.e., image-caption pairs) and multi-lingual data (i.e., parallel sentence pairs) as two different views of the same object, and trains the model to align the two views by maximizing the mutual information between them with conditional masked language modeling and contrastive learning. We pre-train CCLM, a Cross-lingual Cross-modal Language Model, with the cross-view language modeling framework. Empirical results on IGLUE, a multi-lingual multi-modal benchmark, and two multi-lingual image-text retrieval datasets show that while conceptually simpler, CCLM significantly outperforms the prior state-of-

the-art with an average absolute improvement of over 10%. Moreover, CCLM is the first multi-lingual multi-modal pre-trained model that surpasses the translate-test performance of representative English vision-language models by zero-shot cross-lingual transfer.

Multimedia Generative Script Learning for Task Planning
Qingyun Wang, Manling Li, Hou Pong Chan, Lifu Huang, Julia Hockenmaier, Girish Chowdhary and Heng Ji   11:00-12:30 (Pier 7&8)
Goal-oriented generative script learning aims to generate subsequent steps to reach a particular goal, which is an essential task to assist robots or humans in performing stereotypical activities. An important aspect of this process is the ability to capture historical states visually, which provides detailed information that is not covered by text and will guide subsequent steps. Therefore, we propose a new task, Multimedia Generative Script Learning, to generate subsequent steps by tracking historical states in both text and vision modalities, as well as presenting the first benchmark containing 5,652 tasks and 79,089 multimedia steps. This task is challenging in three aspects: the multimedia challenge of capturing the visual states in images, the induction challenge of performing unseen tasks, and the diversity challenge of covering different information in individual steps. We propose to encode visual state changes through a selective multimedia encoder to address the multimedia challenge, transfer knowledge from previously observed tasks using a retrieval-augmented decoder to overcome the induction challenge, and further present distinct information at each step by optimizing a diversity-oriented contrastive learning objective. We define metrics to evaluate both generation and inductive quality. Experiment results demonstrate that our approach significantly outperforms strong baselines.

Translation-Enhanced Multilingual Text-to-Image Generation
Yaoyiran Li, Ching-Yun Chang, Stephen Rawls, Ivan Vulić and Anna Korhonen   11:00-12:30 (Pier 7&8)
Research on text-to-image generation (TTI) still predominantly focuses on the English language due to the lack of annotated image-caption data in other languages; in the long run, this might widen inequitable access to TTI technology. In this work, we thus investigate multilingual TTI (termed mTTI) and the current potential of neural machine translation (NMT) to bootstrap mTTI systems. We provide two key contributions. 1) Relying on a multilingual multi-modal encoder, we provide a systematic empirical study of standard methods used in cross-lingual NLP when applied to mTTI: Translate Train, Translate Test, and Zero-Shot Transfer. 2) We propose Ensemble Adapter (EnsAd), a novel parameter-efficient approach that learns to weigh and consolidate the multilingual text knowledge within the mTTI framework, mitigating the language gap and thus improving mTTI performance. Our evaluations on standard mTTI datasets COCO-CN, Multi30K Task2, and LAION-5B demonstrate the potential of translation-enhanced mTTI systems and also validate the benefits of the proposed EnsAd which derives consistent gains across all datasets. Further investigations on model variants, ablation studies, and qualitative analyses provide additional insights on the inner workings of the proposed mTTI approaches.

PV2TEA: Patching Visual Modality to Textual-Established Information Extraction
Hejie Cui, Rongmei Lin, Nasser Zalmout, Chenwei Zhang, Jingbo Shang, Carl Yang and Xian Li   11:00-12:30 (Pier 7&8)
Information extraction, e.g., attribute value extraction, has been extensively studied and formulated based only on text. However, many attributes can benefit from image-based extraction, like color, shape, pattern, among others. The visual modality has long been underutilized, mainly due to multimodal annotation difficulty. In this paper, we aim to patch the visual modality to the textual-established attribute information extractor. The cross-modality integration faces several unique challenges: (C1) images and textual descriptions are loosely paired intra-sample and inter-samples; (C2) images usually contain rich backgrounds that can mislead the prediction; (C3) weakly supervised labels from textual-established ex- tractors are biased for multimodal training. We present PV2TEA, an encoder-decoder architecture equipped with three bias reduction schemes: (S1) Augmented label-smoothed contrast to improve the cross-modality alignment for loosely-paired image and text; (S2) Attention-pruning that adaptively distinguishes the visual foreground; (S3) Two-level neighborhood regularization that mitigates the label textual bias via reliability estimation. Empirical results on real-world e-Commerce datasets1 demonstrate up to 11.74% absolute (20.97% relatively) F1 increase over unimodal baselines.

A Neural Divide-and-Conquer Reasoning Framework for Image Retrieval from Linguistically Complex Text
Yunxin Li, Baotian Hu, Yuxin Ding, Lin Ma and Min Zhang   11:00-12:30 (Pier 7&8)
Pretrained Vision-Language Models (VLMs) have achieved remarkable performance in image retrieval from text. However, their performance drops drastically when confronted with linguistically complex texts that they struggle to comprehend. Inspired by the Divide-and-Conquer algorithm and dual-process theory, in this paper, we regard linguistically complex texts as compound proposition texts composed of multiple simple proposition sentences and propose an end-to-end Neural Divide-and-Conquer Reasoning framework, dubbed NDCR. It contains three main components: 1) Divide: a proposition generator divides the compound proposition text into simple proposition sentences and produces their corresponding representations, 2) Conquer: a pretrained VLMs-based visual-linguistic interactor achieves the interaction between decomposed proposition sentences and images, 3) Combine: a neural-symbolic reasoner combines the above reasoning states to obtain the final solution via a neural logic reasoning approach. According to the dual-process theory, the visual-linguistic interactor and neural-symbolic reasoner can be regarded as analogical reasoning System 1 and logical reasoning System 2. We conduct extensive experiments on a challenging image retrieval from contextual descriptions data set. Experimental results and analyses indicate NDCR significantly improves performance in the complex image-text reasoning problem.

Enhanced Chart Understanding via Visual Language Pre-training on Plot Table Pairs
Mingyang Zhou, Yi Fung, Long Chen, Christopher Thomas, Heng Ji and Shih-Fu Chang   11:00-12:30 (Pier 7&8)
Building cross-model intelligence that can understand charts and communicate the salient information hidden behind them is an appealing challenge in the vision and language (V+L) community. The capability to uncover the underlined table data of chart figures is a critical key to automatic chart understanding. We introduce ChartT5, a V+L model that learns how to interpret table information from chart images via cross-modal pre-training on plot table pairs. Specifically, we propose two novel pre-training objectives: Masked Header Prediction (MHP) and Masked Value Prediction (MVP) to facilitate the model with different skills to interpret the table information. We have conducted extensive experiments on chart question answering and chart summarization to verify the effectiveness of the proposed pre-training strategies. In particular, on the ChartQA benchmark, our ChartT5 outperforms the state-of-the-art non-pretraining methods by over 8% performance gains.

Towards Parameter-Efficient Integration of Pre-Trained Language Models In Temporal Video Grounding
Erica Kido Shimomoto, Edison Marrese-Taylor, Hiroya Takamura, Ichiro Kobayashi, Hideki Nakayama and Yusuke Miyao 11:00-12:30 (Pier 7&8)
This paper explores the task of Temporal Video Grounding (TVG) where, given an untrimmed video and a query sentence, the goal is to recognize and determine temporal boundaries of action instances in the video described by natural language queries. Recent works tackled this task by improving query inputs with large pre-trained language models (PLM), at the cost of more expensive training. However, the effects of this integration are unclear, as these works also propose improvements in the visual inputs. Therefore, this paper studies the role of query sentence representation with PLMs in TVG and assesses the applicability of parameter-efficient training with NLP adapters. We couple popular PLMs with a selection of existing approaches and test different adapters to reduce the impact of the additional parameters. Our results on three challenging datasets show that, with the same visual inputs, TVG models greatly benefited from the PLM integration and fine-tuning, stressing the importance of the text query representation in this task. Furthermore, adapters were an effective alternative to full fine-tuning, even though they are not tailored to our task, allowing PLM integration in larger TVG models and delivering results comparable to SOTA models. Finally, our results shed light on which adapters work best in different scenarios.

Evaluating pragmatic abilities of image captioners on A3DS

Polina Tsvilodub and Michael Franke   11:00-12:30 (Pier 7&8)

Evaluating grounded neural language model performance with respect to pragmatic qualities like the trade off between truthfulness, contrastivity and overinformativity of generated utterances remains a challenge in absence of data collected from humans. To enable such evaluation, we present a novel open source image-text dataset "Annotated 3D Shapes" (A3DS) comprising over nine million exhaustive natural language annotations and over 12 million variable-granularity captions for the 480,000 images provided by Burgess & Kim (2018). We showcase the evaluation of pragmatic abilities developed by a task-neutral image captioner fine-tuned in a multi-agent communication setting to produce contrastive captions. The evaluation is enabled by the dataset because the exhaustive annotations allow to quantify the presence of contrastive features in the model's generations. We show that the model develops human-like patterns (informativity, brevity, over-informativity for specific features (e.g., shape, color biases)).

Visual Coherence Loss for Coherent and Visually Grounded Story Generation

Xudong Hong, Vera Demberg, Asad Sayeed, Qiankun Zheng and Bernt Schiele   11:00-12:30 (Pier 7&8)

Local coherence is essential for long-form text generation models. We identify two important aspects of local coherence within the visual storytelling task: (1) the model needs to represent re-occurrences of characters within the image sequence in order to mention them correctly in the story; (2) character representations should enable us to find instances of the same characters and distinguish different characters. In this paper, we propose a loss function inspired by a linguistic theory of coherence for self-supervised learning for image sequence representations. We further propose combining features from an object and a face detector to construct stronger character features. To evaluate input-output relevance that current reference-based metrics don't measure, we propose a character matching metric to check whether the models generate referring expressions correctly for characters in input image sequences. Experiments on a visual story generation dataset show that our proposed features and loss function are effective for generating more coherent and visually grounded stories.

Define, Evaluate, and Improve Task-Oriented Cognitive Capabilities for Instruction Generation Models

Lingjun Zhao, Khanh Nguyen and Hal Daumé III   11:00-12:30 (Pier 7&8)

Recent work studies the cognitive capabilities of language models through psychological tests designed for humans. While these studies are helpful for understanding the general capabilities of these models, there is no guarantee that a model possessing sufficient capabilities to pass those tests would actually use those capabilities in performing real-life tasks. In this work, we formulate task-oriented cognitive capabilities, which are human-like cognitive capabilities that language models leverage to perform tasks. These capabilities are (i) the ability to quickly generate good candidate utterances (the search capability) (ii) the ability to predict how a listener interprets those utterances and choose the most appropriate one (the pragmatic capability). We design an evaluation scheme for comparing these capabilities of a language model with those of a human. Applying this scheme to examine various models in a navigation instruction generation problem, we find that their pragmatic capability is severely lacking. This insight leads us to augment them with better models of the listener and obtain a significant boost of 11% in success rate in guiding real humans. Our work advocates for having a principled procedure for aligning language models with humans that involves (i) formulating task-oriented capabilities, (ii) devising a method to quantify their deficiency, and (iii) iteratively improving them.

AltCLIP: Altering the Language Encoder in CLIP for Extended Language Capabilities

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Qinghong Yang and Ledell Y. Wu   11:00-12:30 (Pier 7&8)

CLIP (Contrastive Language–Image Pretraining) is an English multimodal representation model learned from a massive amount of English text-image pairs and has achieved great success in various downstream tasks, including image classification, text-to-image retrieval, and image generation. When extending CLIP to other languages, the major problem is the lack of good-quality text-image pairs. In this work, we present AltCLIP, a simple and low-resource method to build a strong multilingual multimodal representation model. Instead of training a model from scratch on multilingual text-image pairs, we take the original CLIP model trained on English text-image pairs and alter its text encoder with a pre-trained multilingual text encoder (XLM-R). We then align text and image representations by a two-stage training schema consisting of teacher learning and contrastive learning. Our method utilizes the existence of rich parallel text data and pre-trained multilingual language models. We present extensive experimental evaluations to demonstrate the effectiveness of our proposed method. Our model sets new state-of-the-art zero-shot performances on a wide range of tasks in multilingual multimodal benchmarks, including ImageNet-CN/IT/JA/KO serials, Flicker30k-CN, COCO-CN, Multi30k, and XTD. Further, our model outperforms the original CLIP model on zero-shot cross-modal retrieval, Image Classification in the Wild (ICinW) tasks, and CLIP Benchmark. We plan to open-source our code, pre-trained model weights, and evaluation toolkits of multilingual multimodal tasks, to facilitate research on multilingual multimodal representation learning.

Segment-Level and Category-Oriented Network for Knowledge-Based Referring Expression Comprehension

Yuqi Bu, Xin Wu, Liuwu Li, Yi Cai, Qiong Liu and Qingbao Huang   11:00-12:30 (Pier 7&8)

Knowledge-based referring expression comprehension (KB-REC) aims to identify visual objects referred to by expressions that incorporate knowledge. Existing methods employ sentence-level retrieval and fusion methods, which may lead to issues of similarity bias and interference from irrelevant information in unstructured knowledge sentences. To address these limitations, we propose a segment-level and category-oriented network (SLCO). Our approach includes a segment-level and prompt-based knowledge retrieval method to mitigate the similarity bias problem and a category-based grounding method to alleviate interference from irrelevant information in knowledge sentences. Experimental results show that our SLCO can eliminate interference and improve the overall performance of the KB-REC task.

Speaking the Language of Your Listener: Audience-Aware Adaptation via Plug-and-Play Theory of Mind

Ece Takmaz, Nicolo' Brandizzi, Mario Giulianelli, Sandro Pezzelle and Raquel Fernandez   11:00-12:30 (Pier 7&8)

Dialogue participants may have varying levels of knowledge about the topic under discussion. In such cases, it is essential for speakers to adapt their utterances by taking their audience into account. Yet, it is an open question how such adaptation can be modelled in computational agents. In this paper, we model a visually grounded referential game between a knowledgeable speaker and a listener with more limited visual and linguistic experience. Inspired by psycholinguistic theories, we endow our speaker with the ability to adapt its referring expressions via a simulation module that monitors the effectiveness of planned utterances from the listener's perspective. We propose an adaptation mechanism building on plug-and-play approaches to controlled language generation, where utterance generation is steered on the fly by the simulator without finetuning the speaker's underlying language model. Our results and analyses show that our approach is effective: the speaker's utterances become closer to the listener's domain of expertise, which leads to higher communicative success.

FastDiff 2: Revisiting and Incorporating GANs and Diffusion Models in High-Fidelity Speech Synthesis

Rongjie Huang, Yi Ren, Ziyue Jiang, Chenye Cui, Jinglin Liu and Zhou Zhao   11:00-12:30 (Pier 7&8)

Generative adversarial networks (GANs) and denoising diffusion probabilistic models (DDPMs) have recently achieved impressive performances in image and audio synthesis. After revisiting their success in conditional speech synthesis, we find that 1) GANs sacrifice sample diversity for quality and speed, 2) diffusion models exhibit outperformed sample quality and diversity at a high computational cost, where achieving high-quality, fast, and diverse speech synthesis challenges all neural synthesizers. In this work, we propose to converge advantages from GANs and diffusion models by incorporating both classes, introducing dual-empowered modeling perspectives: 1) FastDiff 2 (Diff-GAN), a diffusion model whose denoising process is parametrized by conditional GANs, and the non-Gaussian denoising distribution makes

it much more stable to implement the reverse process with large steps sizes; and 2) FastDiff 2 (GANDiff), a generative adversarial network whose forward process is constructed by multiple denoising diffusion iterations, which exhibits better sample diversity than traditional GANs. Experimental results show that both variants enjoy an efficient 4-step sampling process and demonstrate superior sample quality and diversity. Audio samples are available at https://RevisitSpeech.github.io/

Werewolf Among Us: Multimodal Resources for Modeling Persuasion Behaviors in Social Deduction Games
Bolin Lai, Hongxin Zhang, Miao Liu, Aryan J Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James M Rehg and Diyi Yang 11:00-12:30 (Pier 7&8)
Persuasion modeling is a key building block for conversational agents. Existing works in this direction are limited to analyzing textual dialogue corpus. We argue that visual signals also play an important role in understanding human persuasive behaviors. In this paper, we introduce the first multimodal dataset for modeling persuasion behaviors. Our dataset includes 199 dialogue transcriptions and videos captured in a multi-player social deduction game setting, 26,647 utterance level annotations of persuasion strategy, and game level annotations of deduction game outcomes. We provide extensive experiments to show how dialogue context and visual signals benefit persuasion strategy prediction. We also explore the generalization ability of language models for persuasion modeling and the role of persuasion strategies in predicting social deduction game outcomes. Our dataset can be found at https://persuasion-deductiongame. socialai-data.org. The codes and models are available at https://github.com/ SALT-NLP/PersuationGames.

Multi-modal Action Chain Abductive Reasoning
Mengze Li, Tianbao Wang, Jiahe Xu, Kairong Han, Shengyu Zhang, Zhou Zhao, Jiaxu Miao, wenqiao zhang, Shiliang Pu and Fei Wu 11:00-12:30 (Pier 7&8)
Abductive Reasoning, has long been considered to be at the core ability of humans, which enables us to infer the most plausible explanation of incomplete known phenomena in daily life. However, such critical reasoning capability is rarely investigated for contemporary AI systems under such limited observations. To facilitate this research community, this paper sheds new light on *Abductive Reasoning* by studying a new vision-language task, **M**ulti-modal **A**ction chain abductive **R**easoning (**MAR**), together with a large-scale *Abductive Reasoning* dataset: Given an incomplete set of language described events, MAR aims to imagine the most plausible event by spatio-temporal grounding in past video and then infer the hypothesis of subsequent action chain that can best explain the language premise. To solve this task, we propose a strong baseline model that realizes MAR from two perspectives: (i) We first introduce the transformer, which learns to encode the observation to imagine the plausible event with explicitly interpretable event grounding in the video based on the commonsense knowledge recognition ability. (ii) To complete the assumption of a follow-up action chain, we design a novel symbolic module that can complete strict derivation of the progressive action chain layer by layer. We conducted extensive experiments on the proposed dataset, and the experimental study shows that the proposed model significantly outperforms existing video-language models in terms of effectiveness on our newly created MAR dataset.

Visually-Enhanced Phrase Understanding
Tsu-Yuan Hsu, Chen-An Li, Chao-Wei Huang and Yun-Nung Chen 11:00-12:30 (Pier 7&8)
Large-scale vision-language pre-training has exhibited strong performance in various visual and textual understanding tasks. Recently, the textual encoders of multi-modal pre-trained models have been shown to generate high-quality textual representations, which often outperform models that are purely text-based, such as BERT. In this study, our objective is to utilize both textual and visual encoders of multi-modal pre-trained models to enhance language understanding tasks. We achieve this by generating an image associated with a textual prompt, thus enriching the representation of a phrase for downstream tasks. Results from experiments conducted on four benchmark datasets demonstrate that our proposed method, which leverages visually-enhanced text representations, significantly improves performance in the entity clustering task.

Simple and Effective Unsupervised Speech Translation
Changhan Wang, Hirofumi Inaguma, Peng-Jen Chen, Ilia Kulikov, Yun Tang, Wei-Ning Hsu, Michael Auli and Juan Pino 11:00-12:30 (Pier 7&8)
The amount of labeled data to train models for speech tasks is limited for most languages, however, the data scarcity is exacerbated for speech translation which requires labeled data covering two different languages. To address this issue, we study a simple and effective approach to build speech translation systems without labeled data by leveraging recent advances in unsupervised speech recognition, machine translation and speech synthesis, either in a pipeline approach, or to generate pseudo-labels for training end-to-end speech translation models. Furthermore, we present an unsupervised domain adaptation technique for pre-trained speech models which improves the performance of downstream unsupervised speech recognition, especially for low-resource settings. Experiments show that unsupervised speech-to-text translation outperforms the previous unsupervised state of the art by 3.2 BLEU on the Libri-Trans benchmark, on CoVoST 2, our best systems outperform the best supervised end-to-end models (without pre-training) from only two years ago by an average of 5.0 BLEU over five X-En directions. We also report competitive results on MuST-C and CVSS benchmarks.

TableVLM: Multi-modal Pre-training for Table Structure Recognition
Leiyuan Chen, Chengsong Huang, Xiaoqing Zheng, Jinshu Lin and Xuanjing Huang 11:00-12:30 (Pier 7&8)
Tables are widely used in research and business, which are suitable for human consumption, but not easily machine-processable, particularly when tables are present in images. One of the main challenges to extracting data from images of tables is accurately recognizing table structures, especially for complex tables with cross rows and columns. In this study, we propose a novel multi-modal pre-training model for table structure recognition, named TableVLM. With a two-stream multi-modal transformer-based encoder-decoder architecture, TableVLM learns to capture rich table structure-related features by multiple carefully-designed unsupervised objectives inspired by the notion of masked visual-language modeling. To pre-train this model, we also created a dataset, called ComplexTable, which consists of 1,000K samples to be released publicly. Experiment results show that the model built on pre-trained TableVLM can improve the performance up to 1.97% in tree-editing-distance-score on ComplexTable.

Exploring Speaker-Related Information in Spoken Language Understanding for Better Speaker Diarization
Luyao Cheng, Siqi Zheng, Zhang Qinglin, Hui Wang, Yafeng Chen and Qian Chen 11:00-12:30 (Pier 7&8)
Speaker diarization is a classic task in speech processing and is crucial in multi-party scenarios such as meetings and conversations. Current mainstream speaker diarization approaches consider acoustic information only, which result in performance degradation when encountering adverse acoustic environment. In this paper, we propose methods to extract speaker-related information from semantic content in multi-party meetings, which, as we will show, can further benefit speaker diarization. We introduce two sub-tasks, Dialogue Detection and Speaker-Turn Detection, in which we effectively extract speaker information from conversational semantics. We also propose a simple yet effective algorithm to jointly model acoustic and semantic information and obtain speaker-identified texts. Experiments on both AISHELL-4 and AliMeeting datasets show that our method achieves consistent improvements over acoustic-only speaker diarization systems.

Deeply Coupled Cross-Modal Prompt Learning
Xuejing Liu, Wei Tang, Jinghui Lu, Rui Zhao, Zhaojun Guo and Fei Tan 11:00-12:30 (Pier 7&8)
Recent advancements in multimodal foundation models (e.g., CLIP) have excelled in zero-shot generalization. Prompt tuning involved in the

knowledge transfer from foundation models to downstream tasks has gained significant attention recently. Existing prompt-tuning methods in cross-modal learning, however, either solely focus on language branch, or learn vision-language interaction in a shallow mechanism. In this context, we propose a Deeply coupled Cross-modal Prompt learning (DCP) method based on CLIP. DCP flexibly accommodates the interplay between vision and language with a Cross-Modal Prompt Attention (CMPA) mechanism, which enables the mutual exchange of respective representation through a well-connected multi-head attention progressively and strongly. We then conduct comprehensive few-shot learning experiments on 11 image classification datasets and analyze the robustness to domain shift as well. Thorough experimental analysis evidently demonstrates the superb few-shot generalization and compelling domain adaption capacity of a well-executed DCP.

Unsupervised Mapping of Arguments of Deverbal Nouns to Their Corresponding Verbal Labels
Aviv Weinstein and Yoav Goldberg   11:00-12:30 (Pier 7&8)
Deverbal nouns are nominal forms of verbs commonly used in written English texts to describe events or actions, as well as their arguments. However, many NLP systems, and in particular pattern-based ones, neglect to handle such nominalized constructions. The solutions that do exist for handling arguments of nominalized constructions are based on semantic annotation and require semantic ontologies, making their applications restricted to a small set of nouns. We propose to adopt instead a more syntactic approach, which maps the arguments of deverbal nouns to the universal-dependency relations of the corresponding verbal construction. We present an unsupervised mechanism—based on contextualized word representations—which allows to enrich universal-dependency trees with dependency arcs denoting arguments of deverbal nouns, using the same labels as the corresponding verbal cases. By sharing the same label set as in the verbal case, patterns that were developed for verbs can be applied without modification but with high accuracy also to the nominal constructions.

Probabilistic Transformer: A Probabilistic Dependency Model for Contextual Word Representation
Haoyi Wu and Kewei Tu   11:00-12:30 (Pier 7&8)
Syntactic structures used to play a vital role in natural language processing (NLP), but since the deep learning revolution, NLP has been gradually dominated by neural models that do not consider syntactic structures in their design. One vastly successful class of neural models is transformers. When used as an encoder, a transformer produces contextual representation of words in the input sentence. In this work, we propose a new model of contextual word representation, not from a neural perspective, but from a purely syntactic and probabilistic perspective. Specifically, we design a conditional random field that models discrete latent representations of all words in a sentence as well as dependency arcs between them; and we use mean field variational inference for approximate inference. Strikingly, we find that the computation graph of our model resembles transformers, with correspondences between dependencies and self-attention and between distributions over latent representations and contextual embeddings of words. Experiments show that our model performs competitively to transformers on small to medium sized datasets. We hope that our work could help bridge the gap between traditional syntactic and probabilistic approaches and cutting-edge neural approaches to NLP, and inspire more linguistically-principled neural approaches in the future.

Convergence and Diversity in the Control Hierarchy
Alexandra Cristina Butoi, Ryan Cotterell and David Chiang   11:00-12:30 (Pier 7&8)
Weir has defined a hierarchy of language classes whose second member (L2) is generated by tree-adjoining grammars (TAG), linear indexed grammars (LIG), combinatory categorial grammars, and head grammars. The hierarchy is obtained using the mechanism of control, and L2 is obtained using a context-free grammar (CFG) whose derivations are controlled by another CFG. We adapt Weir's definition of a controllable CFG (called a labeled distinguished CFG) to give a definition of controllable pushdown automata (PDAs), called labeled distinguished PDAs. This yields three new characterizations of L2 as the class of languages generated by PDAs controlling PDAs, PDAs controlling CFGs, and CFGs controlling PDAs. We show that these four formalisms are not only weakly equivalent but equivalent in a stricter sense that we call d-weak equivalence. Furthermore, using an even stricter notion of equivalence called d-strong equivalence, we make precise the intuition that a CFG controlling a CFG is a TAG, a PDA controlling a PDA is an embedded PDA, and a PDA controlling a CFG is a LIG. The fourth member of this family, a CFG controlling a PDA, does not correspond to any kind of automaton we know of, so we invent one and call it a Pushdown Adjoining Automaton (PAA).

Improving Grammar-based Sequence-to-Sequence Modeling with Decomposition and Constraints
Chao Lou and Kewei Tu   11:00-12:30 (Pier 7&8)
Neural QCFG is a grammar-based sequence-to-sequence model with strong inductive biases on hierarchical structures. It excels in interpretability and generalization but suffers from expensive inference. In this paper, we study two low-rank variants of Neural QCFG for faster inference with different trade-offs between efficiency and expressiveness. Furthermore, utilizing the symbolic interface provided by the grammar, we introduce two soft constraints over tree hierarchy and source coverage. We experiment with various datasets and find that our models outperform vanilla Neural QCFG in most settings.

Another Dead End for Morphological Tags? Perturbed Inputs and Parsing
Alberto Muñoz-Ortiz and David Vilares   11:00-12:30 (Pier 7&8)
The usefulness of part-of-speech tags for parsing has been heavily questioned due to the success of word-contextualized parsers. Yet, most studies are limited to coarse-grained tags and high quality written content; while we know little about their influence when it comes to models in production that face lexical errors. We expand these setups and design an adversarial attack to verify if the use of morphological information by parsers: (i) contributes to error propagation or (ii) if on the other hand it can play a role to correct mistakes that word-only neural parsers make. The results on 14 diverse UD treebanks show that under such attacks, for transition- and graph-based models their use contributes to degrade the performance even faster, while for the (lower-performing) sequence labeling parsers they are helpful. We also show that if morphological tags were utopically robust against lexical perturbations, they would be able to correct parsing mistakes.

ParaLS: Lexical Substitution via Pretrained Paraphraser
Jipeng Qiang, Kang Liu, Yun Li, Yunhao Yuan and Yi Zhu   11:00-12:30 (Pier 7&8)
Lexical substitution (LS) aims at finding appropriate substitutes for a target word in a sentence. Recently, LS methods based on pretrained language models have made remarkable progress, generating potential substitutes for a target word through analysis of its contextual surroundings. However, these methods tend to overlook the preservation of the sentence's meaning when generating the substitutes. This study explores how to generate the substitute candidates from a paraphraser, as the generated paraphrases from a paraphraser contain variations in word choice and preserve the sentence's meaning. Since we cannot directly generate the substitutes via commonly used decoding strategies, we propose two simple decoding strategies that focus on the variations of the target word during decoding. Experimental results show that our methods outperform state-of-the-art LS methods based on pre-trained language models on three benchmarks.

Unsupervised Paraphrasing of Multiword Expressions
Takashi Wada, Yuji Matsumoto, Timothy Baldwin and Jey Han Lau   11:00-12:30 (Pier 7&8)
We propose an unsupervised approach to paraphrasing multiword expressions (MWEs) in context. Our model employs only monolingual corpus data and pre-trained language models (without fine-tuning), and does not make use of any external resources such as dictionaries. We evaluate our method on the SemEval 2022 idiomatic semantic text similarity task, and show that it outperforms all unsupervised systems and rivals supervised systems.

Solving Cosine Similarity Underestimation between High Frequency Words by ℓ_2 Norm Discounting
Saeth Wannasuphoprasit, Yi Zhou and Danushka Bollegala   11:00-12:30 (Pier 7&8)
Cosine similarity between two words, computed using their contextualised token embeddings obtained from masked language models (MLMs) such as BERT has shown to underestimate the actual similarity between those words CITATION. This similarity underestimation problem is particularly severe for high frequent words. Although this problem has been noted in prior work, no solution has been proposed thus far. We observe that the ℓ_2 norm of contextualised embeddings of a word correlates with its log-frequency in the pretraining corpus. Consequently, the larger ℓ_2 norms associated with the high frequent words reduce the cosine similarity values measured between them, thus underestimating the similarity scores. To solve this issue, we propose a method to *discount* the ℓ_2 norm of a contextualised word embedding by the frequency of that word in a corpus when measuring the cosine similarities between words. We show that the so called *stop* words behave differently from the rest of the words, which require special consideration during their discounting process. Experimental results on a contextualised word similarity dataset show that our proposed discounting method accurately solves the similarity underestimation problem.An anonymized version of the source code of our proposed method is submitted to the reviewing system.

Ambiguity Meets Uncertainty: Investigating Uncertainty Estimation for Word Sense Disambiguation
Zhu Liu and Ying Liu   11:00-12:30 (Pier 7&8)
Word sense disambiguation (WSD), which aims to determine an appropriate sense for a target word given its context, is crucial for natural language understanding. Existing supervised methods treat WSD as a classification task and have achieved remarkable performance. However, they ignore uncertainty estimation (UE) in the real-world setting, where the data is always noisy and out of distribution. This paper extensively studies UE on the benchmark designed for WSD. Specifically, we first compare four uncertainty scores for a state-of-the-art WSD model and verify that the conventional predictive probabilities obtained at the end of the model are inadequate to quantify uncertainty. Then, we examine the capability of capturing data and model uncertainties by the model with the selected UE score on well-designed test scenarios and discover that the model reflects data uncertainty satisfactorily but underestimates model uncertainty. Furthermore, we explore numerous lexical properties that intrinsically affect data uncertainty and provide a detailed analysis of four critical aspects: the syntactic category, morphology, sense granularity, and semantic relations.

Multi-Level Knowledge Distillation for Out-of-Distribution Detection in Text
Qianhui Wu, Huiqiang Jiang, Haonan Yin, Börje F. Karlsson and Chin-Yew Lin   11:00-12:30 (Pier 7&8)
Self-supervised representation learning has proved to be a valuable component for out-of-distribution (OoD) detection with only the texts of in-distribution (ID) examples. These approaches either train a language model from scratch or fine-tune a pre-trained language model using ID examples, and then take the perplexity output by the language model as OoD scores. In this paper, we analyze the complementary characteristic of both methods and propose a multi-level knowledge distillation approach that integrates their strengths while mitigating their limitations. Specifically, we use a fine-tuned model as the teacher to teach a randomly initialized student model on the ID examples. Besides the prediction layer distillation, we present a similarity-based intermediate layer distillation method to thoroughly explore the representation space of the teacher model. In this way, the learned student can better represent the ID data manifold while gaining a stronger ability to map OoD examples outside the ID data manifold with the regularization inherited from pre-training. Besides, the student model sees only ID examples during parameter training, further promoting more distinguishable features for OoD detection. We conduct extensive experiments over multiple benchmark datasets, i.e., CLINC150, SST, ROSTD, 20 NewsGroups, and AG News; showing that the proposed method yields new state-of-the-art performance. We also explore its application as an AIGC detector to distinguish answers generated by ChatGPT and human experts. It is observed that our model exceeds human evaluators in the pair-expert task on the Human ChatGPT Comparison Corpus.

On Isotropy, Contextualization and Learning Dynamics of Contrastive-based Sentence Representation Learning
Chenghao Xiao, Yang Long and Noura Al Moubayed   11:00-12:30 (Pier 7&8)
Incorporating contrastive learning objectives in sentence representation learning (SRL) has yielded significant improvements on many sentence-level NLP tasks. However, it is not well understood why contrastive learning works for learning sentence-level semantics. In this paper, we aim to help guide future designs of sentence representation learning methods by taking a closer look at contrastive SRL through the lens of isotropy, contextualization and learning dynamics. We interpret its successes through the geometry of the representation shifts and show that contrastive learning brings isotropy, and drives high intra-sentence similarity: when in the same sentence, tokens converge to similar positions in the semantic space. We also find that what we formalize as "spurious contextualization" is mitigated for semantically meaningful tokens, while augmented for functional ones. We find that the embedding space is directed towards the origin during training, with more areas now better defined. We ablate these findings by observing the learning dynamics with different training temperatures, batch sizes and pooling methods.

Exploring Non-Verbal Predicates in Semantic Role Labeling: Challenges and Opportunities
Riccardo Orlando, Simone Conia and Roberto Navigli   11:00-12:30 (Pier 7&8)
Although we have witnessed impressive progress in Semantic Role Labeling (SRL), most of the research in the area is carried out assuming that the majority of predicates are verbs. Conversely, predicates can also be expressed using other parts of speech, e.g., nouns and adjectives. However, non-verbal predicates appear in the benchmarks we commonly use to measure progress in SRL less frequently than in some real-world settings – newspaper headlines, dialogues, and tweets, among others. In this paper, we put forward a new PropBank dataset which boasts wide coverage of multiple predicate types. Thanks to it, we demonstrate empirically that standard benchmarks do not provide an accurate picture of the current situation in SRL and that state-of-the-art systems are still incapable of transferring knowledge across different predicate types. Having observed these issues, we also present a novel, manually-annotated challenge set designed to give equal importance to verbal, nominal, and adjectival predicate-argument structures. We use such dataset to investigate whether we can leverage different linguistic resources to promote knowledge transfer. In conclusion, we claim that SRL is far from "solved", and its integration with other semantic tasks might enable significant improvements in the future, especially for the long tail of non-verbal predicates, thereby facilitating further research on SRL for non-verbal predicates. We release our software and datasets at https://github.com/sapienzanlp/exploring-srl.

Incorporating Graph Information in Transformer-based AMR Parsing
Pavlo Vasylenko, Pere Lluís Huguet Cabot, Abelardo Carlos Martínez Lorenzo and Roberto Navigli   11:00-12:30 (Pier 7&8)
Abstract Meaning Representation (AMR) is a Semantic Parsing formalism that aims at providing a semantic graph abstraction representing a given text. Current approaches are based on autoregressive language models such as BART or T5, fine-tuned through Teacher Forcing to obtain a linearized version of the AMR graph from a sentence. In this paper, we present LeakDistill, a model and method that explores a modification to the Transformer architecture, using structural adapters to explicitly incorporate graph information into the learned representations and improve AMR parsing performance. Our experiments show how, by employing word-to-node alignment to embed graph structural information into the encoder at training time, we can obtain state-of-the-art AMR parsing through self-knowledge distillation, even without the use of additional data. We release the code at http://www.github.com/sapienzanlp/LeakDistill.

Entailment as Robust Self-Learner
Jiaxin Ge, Hongyin Luo, Yoon Kim and James Glass   11:00-12:30 (Pier 7&8)

Entailment has been recognized as an important metric for evaluating natural language understanding (NLU) models, and recent studies have found that entailment pretraining benefits weakly supervised fine-tuning. In this work, we design a prompting strategy that formulates a number of different NLU tasks as contextual entailment. This approach improves the zero-shot adaptation of pretrained entailment models. Secondly, we notice that self-training entailment-based models with unlabeled data can significantly improve the adaptation performance on downstream tasks. To achieve more stable improvement, we propose the Simple Pseudo-Label Editing (SimPLE) algorithm for better pseudo-labeling quality in self-training. We also found that both pretrained entailment-based models and the self-trained models are robust against adversarial evaluation data. Experiments on binary and multi-class classification tasks show that SimPLE leads to more robust self-training results, indicating that the self-trained entailment models are more efficient and trustworthy than large language models on language understanding tasks.

Composition-contrastive Learning for Sentence Embeddings
Sachin J. Chanchani and Ruihong Huang   11:00-12:30 (Pier 7&8)
Vector representations of natural language are ubiquitous in search applications. Recently, various methods based on contrastive learning have been proposed to learn textual representations from unlabelled data; by maximizing alignment between minimally-perturbed embeddings of the same text, and encouraging a uniform distribution of embeddings across a broader corpus. Differently, we propose maximizing alignment between texts and a composition of their phrasal constituents. We consider several realizations of this objective and elaborate the impact on representations in each case. Experimental results on semantic textual similarity tasks show improvements over baselines that are comparable with state-of-the-art approaches. Moreover, this work is the first to do so without incurring costs in auxiliary training objectives or additional network parameters.

The Best of Both Worlds: Combining Human and Machine Translations for Multilingual Semantic Parsing with Active Learning
Zhuang Li, Lizhen Qu, Philip Cohen, Raj V Tumuluri and Gholamreza Haffari   11:00-12:30 (Pier 7&8)
Multilingual semantic parsing aims to leverage the knowledge from the high-resource languages to improve low-resource semantic parsing, yet commonly suffers from the data imbalance problem. Prior works propose to utilize the translations by either humans or machines to alleviate such issues. However, human translations are expensive, while machine translations are cheap but prone to error and bias. In this work, we propose an active learning approach that exploits the strengths of both human and machine translations by iteratively adding small batches of human translations into the machine-translated training set. Besides, we propose novel aggregated acquisition criteria that help our active learning method select utterances to be manually translated. Our experiments demonstrate that an ideal utterance selection can significantly reduce the error and bias in the translated data, resulting in higher parser accuracies than the parsers merely trained on the machine-translated data.

Cross-lingual AMR Aligner: Paying Attention to Cross-Attention
Abelardo Carlos Martínez Lorenzo, Pere Lluís Huguet Cabot and Roberto Navigli   11:00-12:30 (Pier 7&8)
This paper introduces a novel aligner for Abstract Meaning Representation (AMR) graphs that can scale cross-lingual, and is thus capable of aligning units and spans in sentences of different languages. Our approach leverages modern Transformer-based parsers, which inherently encode alignment information in their cross-attention weights, allowing us to extract this information during parsing. This eliminates the need for English-specific rules or the Expectation Maximization (EM) algorithm that have been used in previous approaches. In addition, we propose a guided supervised method using alignment to further enhance the performance of our aligner. We achieve state-of-the-art results in the benchmarks for AMR alignment and demonstrate our aligner's ability to obtain them across multiple languages. Our code will be available at https://www.github.com/babelscape/AMR-alignment.

Alleviating Over-smoothing for Unsupervised Sentence Representation
Nuo Chen, Linjun Shou, Jian Pei, Ming Gong, Bowen Cao, Jianhui Chang, Jia Li and Daxin Jiang   11:00-12:30 (Pier 7&8)
Currently, learning better unsupervised sentence representations is the pursuit of many natural language processing communities. Lots of approaches based on pre-trained language models (PLMs) and contrastive learning have achieved promising results on this task. Experimentally, we observe that the over-smoothing problem reduces the capacity of these powerful PLMs, leading to sub-optimal sentence representations. In this paper, we present a Simple method named Self-Contrastive Learning (SSCL) to alleviate this issue, which samples negatives from PLMs intermediate layers, improving the quality of the sentence representation. Our proposed method is quite simple and can be easily extended to various state-of-the-art models for performance boosting, which can be seen as a plug-and-play contrastive framework for learning unsupervised sentence representation. Extensive results prove that SSCL brings the superior performance improvements of different strong baselines (e.g., BERT and SimCSE) on Semantic Textual Similarity and Transfer datasets

Enhancing Language Representation with Constructional Information for Natural Language Understanding
Lvxiaowei Xu, Jianwang Wu, Jiawei Peng, Zhilin Gong, Ming Cai and Tianxiang Wang   11:00-12:30 (Pier 7&8)
Natural language understanding (NLU) is an essential branch of natural language processing, which relies on representations generated by pre-trained language models (PLMs). However, PLMs primarily focus on acquiring lexico-semantic information, while they may be unable to adequately handle the meaning of constructions. To address this issue, we introduce construction grammar (CxG), which highlights the pairings of form and meaning, to enrich language representation. We adopt usage-based construction grammar as the basis of our work, which is highly compatible with statistical models such as PLMs. Then a HyCxG framework is proposed to enhance language representation through a three-stage solution. First, all constructions are extracted from sentences via a slot-constraints approach. As constructions can overlap with each other, bringing redundancy and imbalance, we formulate the conditional max coverage problem for selecting the discriminative constructions. Finally, we propose a relational hypergraph attention network to acquire representation from constructional information by capturing high-order word interactions among constructions. Extensive experiments demonstrate the superiority of the proposed model on a variety of NLU tasks.

SETI: Systematicity Evaluation of Textual Inference
Xiyan Fu and Anette Frank   11:00-12:30 (Pier 7&8)
We propose SETI (Systematicity Evaluation of Textual Inference), a novel and comprehensive benchmark designed for evaluating pre-trained language models (PLMs) for their systematicity capabilities in the domain of textual inference. Specifically, SETI offers three different NLI tasks and corresponding datasets to evaluate various types of systematicity in reasoning processes. In order to solve these tasks, models are required to perform compositional inference based on known primitive constituents. We conduct experiments of SETI on six widely used PLMs. Results show that various PLMs are able to solve unseen compositional inferences when having encountered the knowledge of how to combine primitives, with good performance. However, they are considerably limited when this knowledge is unknown to the model (40-100 % points decrease). Furthermore, we find that PLMs are able to improve dramatically once exposed to crucial compositional knowledge in minimalistic shots. These findings position SETI as the first benchmark for measuring the future progress of PLMs in achieving systematicity generalization in the textual inference.

Predicting Numerals in Text Using Nearest Neighbor Language Models
Taku Sakamoto and Akiko Aizawa   11:00-12:30 (Pier 7&8)

Commonsense about quantitative properties is essential for a deep understanding of texts containing numerals. However, naive language models (LMs) treat numerals as string tokens; therefore, they lack an understanding of the magnitudes of numerals, resulting in a difficulty in acquiring the commonsense. In this study, we apply the k-nearest neighbor LM (kNN-LM) to the masked numeral prediction (MNP) task, which measures the quantitative commonsense of numerals. kNN-LM extends pre-trained neural LMs with the k-nearest neighbor (kNN) search. Since it can utilize patterns that appear in the datastore for prediction, we expect an improvement in numeral prediction accuracy, which is associated with a high rate of occurrence of out-of-vocabulary (OOV) words. Through experiments, we verified that the retrieval-based method is effective for fine-grained predictions of numerals from context, especially for the OOV numerals. We also compared two different context spans for context representations to improve the accuracy of kNN search by using only the words that are closely related to the masked numeral: the mask and its surrounding words, and the mask and its subsequent words. Our results reveal that using only the embeddings of mask tokens for numerals in kNN search is the most effective approach for realizing MNP tasks.

Investigating Transformer-Guided Chaining for Interpretable Natural Logic Reasoning
Kanagasabai Rajaraman, Saravanan Rajamanickam and Wei Shi   11:00-12:30 (Pier 7&8)
Natural logic reasoning has received increasing attention lately, with several datasets and neural models proposed, though with limited success. More recently, a new class of works have emerged adopting a Neuro-Symbolic approach, called transformer guided chaining, whereby the idea is to iteratively perform 1-step neural inferences and chain together the results to generate a multi-step reasoning trace. Several works have adapted variants of this central idea and reported significantly high accuracies compared to vanilla LLM's. In this paper, we perform a critical empirical investigation of the chaining approach on a multi-hop First-Order Logic (FOL) reasoning benchmark. In particular, we develop a reference implementation, called Chainformer, and conduct several experiments to analyze the accuracy, generalization, interpretability, and performance over FOLs. Our findings highlight key strengths and possible current limitations and suggest potential areas for future research in logic reasoning.

Categorial grammar induction from raw data
Christian Clark and William Schuler   11:00-12:30 (Pier 7&8)
Grammar induction, the task of learning a set of grammatical rules from raw or minimally labeled text data, can provide clues about what kinds of syntactic structures are learnable without prior knowledge. Recent work (e.g., Kim et al., 2019; Zhu et al., 2020; Jin et al., 2021a) has achieved advances in unsupervised induction of probabilistic context-free grammars (PCFGs). However, categorial grammar induction has received less recent attention, despite allowing inducers to support a larger set of syntactic categories—due to restrictions on how categories can combine—and providing a transparent interface with compositional semantics, opening up possibilities for models that jointly learn form and meaning. Motivated by this, we propose a new model for inducing a basic (Ajdukiewicz, 1935; Bar-Hillel, 1953) categorial grammar. In contrast to earlier categorial grammar induction systems (e.g., Bisk and Hockenmaier, 2012), our model learns from raw data without any part-of-speech information. Experiments on child-directed speech show that our model attains a recall-homogeneity of 0.33 on average, which dramatically increases to 0.59 when a bias toward forward function application is added to the model.

How does the brain process syntactic structure while listening?
SUBBA REDDY OOTA, Mounika Marreddy, Manish Gupta and Raju Surampudi Bapi   11:00-12:30 (Pier 7&8)
Syntactic parsing is the task of assigning a syntactic structure to a sentence. There are two popular syntactic parsing methods: constituency and dependency parsing. Recent works have used syntactic embeddings based on constituency trees, incremental top-down parsing, and other word syntactic features for brain activity prediction given the text stimuli to study how the syntax structure is represented in the brain's language network. However, the effectiveness of dependency parse trees or the relative predictive power of the various syntax parsers across brain areas, especially for the listening task, is yet unexplored. In this study, we investigate the predictive power of the brain encoding models in three settings: (i) individual performance of the constituency and dependency syntactic parsing based embedding methods, (ii) efficacy of these syntactic parsing based embedding methods when controlling for basic syntactic signals, (iii) relative effectiveness of each of the syntactic embedding methods when controlling for the other. Further, we explore the relative importance of syntactic information (from these syntactic embedding methods) versus semantic information using BERT embeddings. We find that constituency parsers help explain activations in the temporal lobe and middle-frontal gyrus, while dependency parsers better encode syntactic structure in the angular gyrus and posterior cingulate cortex. Although semantic signals from BERT are more effective compared to any of the syntactic features or embedding methods, syntactic embedding methods explain additional variance for a few brain regions.

How Well Do Large Language Models Perform on Faux Pas Tests?
Natalie Shapira, Guy Zwirn and Yoav Goldberg   11:00-12:30 (Pier 7&8)
Motivated by the question of the extent to which large language models "understand" social intelligence, we investigate the ability of such models to generate correct responses to questions involving descriptions of faux pas situations. The faux pas test is a test used in clinical psychology, which is known to be more challenging for children than individual tests of theory-of-mind or social intelligence. Our results demonstrate that, while the models seem to sometimes offer correct responses, they in fact struggle with this task, and that many of the seemingly correct responses can be attributed to over-interpretation by the human reader ("the ELIZA effect"). An additional phenomenon observed is the failure of most models to generate a correct response to presupposition questions. Finally, in an experiment in which the models are tasked with generating original faux pas stories, we find that while some models are capable of generating novel faux pas stories, the stories are all explicit, as the models are limited in their abilities to describe situations in an implicit manner.

PragmatiCQA: A Dataset for Pragmatic Question Answering in Conversations
Peng Qi, Nina Du, Christopher D. Manning and Jing Huang   11:00-12:30 (Pier 7&8)
Pragmatic reasoning about another speaker's unspoken intent and state of mind is crucial to efficient and effective human communication. It is virtually omnipresent in conversations between humans, e.g., when someone asks "do you have a minute?", instead of interpreting it literally as a query about your schedule, you understand that the speaker might have requests that take time, and respond accordingly. In this paper, we present PragmatiCQA, the first large-scale open-domain question answering (QA) dataset featuring 6873 QA pairs that explores pragmatic reasoning in conversations over a diverse set of topics. We designed innovative crowdsourcing mechanisms for interest-based and task-driven data collection to address the common issue of incentive misalignment between crowdworkers and potential users. To compare computational models' capability at pragmatic reasoning, we also propose several quantitative metrics to evaluate question answering systems on PragmatiCQA. We find that state-of-the-art systems still struggle to perform human-like pragmatic reasoning, and highlight their limitations for future research.

A Match Made in Heaven: A Multi-task Framework for Hyperbole and Metaphor Detection
Naveen Badathala, Abisek Rajakumar Kalarani, Tejpalsingh Siledar and Pushpak Bhattacharyya   11:00-12:30 (Pier 7&8)
Hyperbole and metaphor are common in day-to-day communication (e.g., "I am in deep trouble": how does trouble have depth?), which makes their detection important, especially in a conversational AI setting. Existing approaches to automatically detect metaphor and hyperbole have studied these language phenomena independently, but their relationship has hardly, if ever, been explored computationally. In this paper, we propose a multi-task deep learning framework to detect hyperbole and metaphor simultaneously. We hypothesize that metaphors help in hyperbole detection, and vice-versa. To test this hypothesis, we annotate two hyperbole datasets- HYPO and HYPO-L- with metaphor

labels. Simultaneously, we annotate two metaphor datasets- TroFi and LCC- with hyperbole labels. Experiments using these datasets give an improvement of the state of the art of hyperbole detection by 12%. Additionally, our multi-task learning (MTL) approach shows an improvement of up to 17% over single-task learning (STL) for both hyperbole and metaphor detection, supporting our hypothesis. To the best of our knowledge, ours is the first demonstration of computational leveraging of linguistic intimacy between metaphor and hyperbole, leading to showing the superiority of MTL over STL for hyperbole and metaphor detection.

The Coreference under Transformation Labeling Dataset: Entity Tracking in Procedural Texts Using Event Models

Kyeongmin Rim, Jingxuan Tu, Bingyang Ye, Marc Verhagen, Eben Holderness and James Pustejovsky   11:00-12:30 (Pier 7&8)

We demonstrate that coreference resolution in procedural texts is significantly improved when performing transformation-based entity linking prior to coreference relation identification. When events in the text introduce changes to the state of participating entities, it is often impossible to accurately link entities in anaphoric and coreference relations without an understanding of the transformations those entities undergo. We show how adding event semantics helps to better model entity coreference. We argue that all transformation predicates, not just creation verbs, introduce a new entity into the discourse, as a kind of generalized Result Role, which is typically not textually mentioned. This allows us to model procedural texts as process graphs and to compute the coreference type for any two entities in the recipe. We present our annotation methodology and the corpus generated as well as describe experiments on coreference resolution of entity mentions under a process-oriented model of events.

Learning Event-aware Measures for Event Coreference Resolution

Yao Yao, Zuchao Li and Hai Zhao   11:00-12:30 (Pier 7&8)

Researchers are witnessing knowledge-inspired natural language processing shifts the focus from entity-level to event-level, whereas event coreference resolution is one of the core challenges. This paper proposes a novel model for within-document event coreference resolution. On the basis of event but not entity as before, our model learns and integrates multiple representations from both event alone and event pair. For the former, we introduce multiple linguistics-motivated event alone features for more discriminative event representations. For the latter, we consider multiple similarity measures to capture the distinction of event pair. Our proposed model achieves new state-of-the-art on the ACE 2005 benchmark, demonstrating the effectiveness of our proposed framework.

$2 * n$ is better than n^2: Decomposing Event Coreference Resolution into Two Tractable Problems

Shafiuddin Rehan Ahmed, Abhijnan Nath, James H. Martin and Nikhil Krishnaswamy   11:00-12:30 (Pier 7&8)

Event Coreference Resolution (ECR) is the task of linking mentions of the same event either within or across documents. Most mention pairs are not coreferent, yet many that are coreferent can be identified through simple techniques such as lemma matching of the event triggers or the sentences in which they appear. Existing methods for training coreference systems sample from a largely skewed distribution, making it difficult for the algorithm to learn coreference beyond surface matching. Additionally, these methods are intractable because of the quadratic operations needed. To address these challenges, we break the problem of ECR into two parts: a) a heuristic to efficiently filter out a large number of non-coreferent pairs, and b) a training approach on a balanced set of coreferent and non-coreferent mention pairs. By following this approach, we show that we get comparable results to the state of the art on two popular ECR datasets while significantly reducing compute requirements. We also analyze the mention pairs that are "hard" to accurately classify as coreferent or non-coreferentcode repo: github.com/ahmeshaf/lemma_ce_coref.

Automatic Identification of Code-Switching Functions in Speech Transcripts

Ritu Madhura Belani and Jeffrey Flanigan   11:00-12:30 (Pier 7&8)

Code-switching, or switching between languages, occurs for many reasons and has important linguistic, sociological, and cultural implications. Multilingual speakers code-switch for a variety of communicative functions, such as expressing emotions, borrowing terms, making jokes, introducing a new topic, etc. The function of code-switching may be quite useful for the analysis of linguists, cognitive scientists, speech therapists, and others, but is not readily apparent. To remedy this situation, we annotate and release a new dataset of functions of code-switching in Spanish-English. We build the first system (to our knowledge) to automatically identify a wide range of functions for which speakers code-switch in everyday speech, achieving an accuracy of 75% across all functions.

Exploring the Relationship between Alignment and Cross-lingual Transfer in Multilingual Transformers

Felix Gaschi, Patricio Cerda, Parisa RASTIN and Yannick Toussaint   11:00-12:30 (Pier 7&8)

Without any explicit cross-lingual training data, multilingual language models can achieve cross-lingual transfer. One common way to improve this transfer is to perform realignment steps before fine-tuning, i.e., to train the model to build similar representations for pairs of words from translated sentences. But such realignment methods were found to not always improve results across languages and tasks, which raises the question of whether aligned representations are truly beneficial to cross-lingual transfer. We provide evidence that alignment is actually significantly correlated with cross-lingual transfer across languages, models and random seeds. We show that fine-tuning can have a significant impact on alignment, depending mainly on the downstream task and the model. Finally, we show that realignment can, in some instances, improve cross-lingual transfer, and we identify conditions in which realignment methods provide significant improvements. Namely, we find that realignment works better on tasks for which alignment is correlated with cross-lingual transfer when generalizing to a distant language and with smaller models, as well as when using a bilingual dictionary rather than FastAlign to extract realignment pairs. For example, for POS-tagging, between English and Arabic, realignment can bring a +15.8 accuracy improvement on distilmBERT, even outperforming XLM-R Large by 1.7. We thus advocate for further research on realignment methods for smaller multilingual models as an alternative to scaling.

Typology Guided Multilingual Position Representations: Case on Dependency Parsing

Tao Ji, Yuanbin Wu and xiaoling Wang   11:00-12:30 (Pier 7&8)

Recent multilingual models benefit from strong unified semantic representation models. However, due to conflict linguistic regularities, ignoring language-specific features during multilingual learning may suffer from negative transfer. In this work, we analyze the relation between a language's position space and its typological characterization, and suggest deploying different position spaces for different languages. We develop a position generation network which combines prior knowledge from typology features and existing position vectors. Experiments on the multilingual dependency parsing task show that the learned position vectors exhibit meaningful hidden structures, and they can help achieving the best multilingual parsing results.

Speaking Multiple Languages Affects the Moral Bias of Language Models

Katharina Haemmerl, Bjoern Deiseroth, Patrick Schramowski, Jindřich Libovický, Constantin A Rothkopf, Alexander Fraser and Kristian Kersting   11:00-12:30 (Pier 7&8)

Pre-trained multilingual language models (PMLMs) are commonly used when dealing with data from multiple languages and cross-lingual transfer. However, PMLMs are trained on varying amounts of data for each language. In practice this means their performance is often much better on English than many other languages. We explore to what extent this also applies to moral norms. Do the models capture moral norms from English and impose them on other languages? Do the models exhibit random and thus potentially harmful beliefs in certain languages? Both these issues could negatively impact cross-lingual transfer and potentially lead to harmful outcomes. In this paper, we (1) apply the MORALDIRECTION framework to multilingual models, comparing results in German, Czech, Arabic, Chinese, and English, (2)

analyse model behaviour on filtered parallel subtitles corpora, and (3) apply the models to a Moral Foundations Questionnaire, comparing with human responses from different countries. Our experiments demonstrate that, indeed, PMLMs encode differing moral biases, but these do not necessarily correspond to cultural differences or commonalities in human opinions. We release our code and models.

Dual-Alignment Pre-training for Cross-lingual Sentence Embedding
Ziheng Li, Shaohan Huang, Zihan Zhang, Zhi-Hong Deng, Qiang Lou, Haizhen Huang, Jian Jiao, Furu Wei, Weiwei Deng and Qi Zhang
11:00-12:30 (Pier 7&8)
Recent studies have shown that dual encoder models trained with the sentence-level translation ranking task are effective methods for cross-lingual sentence embedding. However, our research indicates that token-level alignment is also crucial in multilingual scenarios, which has not been fully explored previously. Based on our findings, we propose a dual-alignment pre-training (DAP) framework for cross-lingual sentence embedding that incorporates both sentence-level and token-level alignment. To achieve this, we introduce a novel representation translation learning (RTL) task, where the model learns to use one-side contextualized token representation to reconstruct its translation counterpart. This reconstruction objective encourages the model to embed translation information into the token representation. Compared to other token-level alignment methods such as translation language modeling, RTL is more suitable for dual encoder architectures and is computationally efficient. Extensive experiments on three sentence-level cross-lingual benchmarks demonstrate that our approach can significantly improve sentence embedding. Our code is available at https://github.com/ChillingDream/DAP.

TADA : Task Agnostic Dialect Adapters for English
William Held, Caleb Ziems and Diyi Yang   11:00-12:30 (Pier 7&8)
Large Language Models, the dominant starting point for Natural Language Processing (NLP) applications, fail at a higher rate for speakers of English dialects other than Standard American English (SAE). Prior work addresses this using task specific data or synthetic data augmentation, both of which require intervention for each dialect and task pair. This poses a scalability issue that prevents the broad adoption of robust dialectal English NLP. We introduce a simple yet effective method for task-agnostic dialect adaptation by aligning non-SAE dialects using adapters and composing them with task-specific adapters from SAE. Task-Agnostic Dialect Adapters (TADA) improve dialectal robustness on 4 dialectal variants of the GLUE benchmark without task-specific supervision.

Inducing Character-level Structure in Subword-based Language Models with Type-level Interchange Intervention Training
Jing Huang, Zhengxuan Wu, Kyle Mahowald and Christopher Potts   11:00-12:30 (Pier 7&8)
Language tasks involving character-level manipulations (e.g., spelling corrections, arithmetic operations, word games) are challenging for models operating on subword units. To address this, we develop a causal intervention framework to learn robust and interpretable character representations inside subword-based language models. Our method treats each character as a typed variable in a causal model and learns such causal structures by adapting the interchange intervention training method of Geiger et al. (2021). We additionally introduce a suite of character-level tasks that systematically vary in their dependence on meaning and sequence-level context. While character-level models still perform best on purely form-based tasks like string reversal, our method outperforms character-level models on more complex tasks that blend form, meaning, and context, such as spelling correction in context and word search games. Compared with standard subword-based models, our approach also significantly improves robustness on unseen token sequences and leads to human-interpretable internal representations of characters.

Limitations of Language Models in Arithmetic and Symbolic Induction
Jing Qian, Hong Wang, Zekun Li, Shiyang Li and Xifeng Yan   11:00-12:30 (Pier 7&8)
Recent work has shown that large pretrained Language Models (LMs) can not only perform remarkably well on a range of Natural Language Processing (NLP) tasks but also start improving on reasoning tasks such as arithmetic induction, symbolic manipulation, and commonsense reasoning with increasing size of models. However, it is still unclear what the underlying capabilities of these LMs are. Surprisingly, we find that these models have limitations on certain basic symbolic manipulation tasks such as copy, reverse, and addition. When the total number of symbols or repeating symbols increases, the model performance drops quickly. We investigate the potential causes behind this phenomenon and examine a set of possible methods, including explicit positional markers, fine-grained computation steps, and LMs with callable programs. Experimental results show that none of these techniques can solve the simplest addition induction problem completely. In the end, we introduce LMs with tutor, which demonstrates every single step of teaching. LMs with tutor is able to deliver 100% accuracy in situations of OOD and repeating symbols, shedding new insights on the boundary of large LMs in induction.

Explanation Regeneration via Information Bottleneck
Qintong Li, Zhiyong Wu, Lingpeng Kong and Wei Bi   11:00-12:30 (Pier 7&8)
Explaining the black-box predictions of NLP models naturally and accurately is an important open problem in natural language generation. These free-text explanations are expected to contain sufficient and carefully-selected evidence to form supportive arguments for predictions. Thanks to the superior generative capacity of large pretrained language models (PLM), recent work built on prompt engineering enables explanations generated without specific training. However, explanations generated through single-pass prompting often lack sufficiency and conciseness, due to the prompt complexity and hallucination issues. To discard the dross and take the essence of current PLM's results, we propose to produce sufficient and concise explanations via the information bottleneck (EIB) theory. EIB regenerates explanations by polishing the single-pass output of PLM but retaining the information that supports the contents being explained by balancing two information bottleneck objectives. Experiments on two different tasks verify the effectiveness of EIB through automatic evaluation and thoroughly-conducted human evaluation.

Characterizing the Impacts of Instances on Robustness
*Rui Zheng, Zhiheng Xi, Qin Liu, Wenbin Lai, Tao Gui, Qi Zhang, Xuanjing Huang, Jin Ma, Ying Shan and Weifeng Ge*11:00-12:30 (Pier 7&8)
Building robust deep neural networks (DNNs) against adversarial attacks is an important but challenging task. Previous defense approaches mainly focus on developing new model structures or training algorithms, but they do little to tap the potential of training instances, especially instances with robust patterns carring innate robustness. In this paper, we show that robust and non-robust instances in the training dataset, though are both important for test performance, have contrary impacts on robustness, which makes it possible to build a highly robust model by leveraging the training dataset in a more effective way. We propose a new method that can distinguish between robust instances from non-robust ones according to the model's sensitivity to perturbations on individual instances during training. Surprisingly, we find that the model under standard training easily overfits the robust instances by relying on their simple patterns before the model completely learns their robust features. Finally, we propose a new mitigation algorithm to further release the potential of robust instances. Experimental results show that proper use of robust instances in the original dataset is a new line to achieve highly robust models.

COCKATIEL: COntinuous Concept ranKed ATtribution with Interpretable ELements for explaining neural net classifiers on NLP
Fanny Jourdan, Agustin Martin Picard, Thomas Fel, Laurent Risser, Jean-Michel Loubes and Nicholas Asher   11:00-12:30 (Pier 7&8)
Transformer architectures are complex and their use in NLP, while it has engendered many successes, makes their interpretability or explainability challenging. Recent debates have shown that attention maps and attribution methods are unreliable (Pruthi et al., 2019; Brunner et al., 2019). In this paper, we present some of their limitations and introduce COCKATIEL, which successfully addresses some of them.

COCKATIEL is a novel, post-hoc, concept-based, model-agnostic XAI technique that generates meaningful explanations from the last layer of a neural net model trained on an NLP classification task by using Non-Negative Matrix Factorization (NMF) to discover the concepts the model leverages to make predictions and by exploiting a Sensitivity Analysis to estimate accurately the importance of each of these concepts for the model. It does so without compromising the accuracy of the underlying model or requiring a new one to be trained.

We conduct experiments in single and multi-aspect sentiment analysis tasks and we show COCKATIEL's superior ability to discover concepts that align with humans' on Transformer models without any supervision, we objectively verify the faithfulness of its explanations through fidelity metrics, and we showcase its ability to provide meaningful explanations in two different datasets.

Our code is freely available: https://github.com/fanny-jourdan/cockatiel

Interpreting Positional Information in Perspective of Word Order
Zhang Xilong, Liu Ruochen, Liu Jin and Liang Xuefeng 11:00-12:30 (Pier 7&8)
The attention mechanism is a powerful and effective method utilized in natural language processing. However, it has been observed that this method is insensitive to positional information. Although several studies have attempted to improve positional encoding and investigate the influence of word order perturbation, it remains unclear how positional encoding impacts NLP models from the perspective of word order. In this paper, we aim to shed light on this problem by analyzing the working mechanism of the attention module and investigating the root cause of its inability to encode positional information. Our hypothesis is that the insensitivity can be attributed to the weight sum operation utilized in the attention module. To verify this hypothesis, we propose a novel weight concatenation operation and evaluate its efficacy in neural machine translation tasks. Our enhanced experimental results not only reveal that the proposed operation can effectively encode positional information but also confirm our hypothesis.

Model Interpretability and Rationale Extraction by Input Mask Optimization
Marc Felix Brinner and Sina Zarrieß 11:00-12:30 (Pier 7&8)
Concurrent with the rapid progress in neural network-based models in NLP, the need for creating explanations for the predictions of these black-box models has risen steadily. Yet, especially for complex inputs like texts or images, existing interpretability methods still struggle with deriving easily interpretable explanations that also accurately represent the basis for the model's decision. To this end, we propose a new, model-agnostic method to generate extractive explanations for predictions made by neural networks, that is based on masking parts of the input which the model does not consider to be indicative of the respective class. The masking is done using gradient-based optimization combined with a new regularization scheme that enforces sufficiency, comprehensiveness, and compactness of the generated explanation. Our method achieves state-of-the-art results in a challenging paragraph-level rationale extraction task, showing that this task can be performed without training a specialized model. We further apply our method to image inputs and obtain high-quality explanations for image classifications, which indicates that the objectives for optimizing explanation masks in text generalize to inputs of other modalities.

Nonlinear Structural Equation Model Guided Gaussian Mixture Hierarchical Topic Modeling
HeGang Chen, Pengbo Mao, Yuyin Lu and Yanghui Rao 11:00-12:30 (Pier 7&8)
Hierarchical topic models, which can extract semantically meaningful topics from a text corpus in an unsupervised manner and automatically organise them into a topic hierarchy, have been widely used to discover the underlying semantic structure of documents. However, the existing models often assume in the prior that the topic hierarchy is a tree structure, ignoring symmetrical dependencies between topics at the same level. Moreover, the sparsity of text data often complicate the analysis. To address these issues, we propose NSEM-GMHTM as a deep topic model, with a Gaussian mixture prior distribution to improve the model's ability to adapt to sparse data, which explicitly models hierarchical and symmetric relations between topics through the dependency matrices and nonlinear structural equations. Experiments on widely used datasets show that our NSEM-GMHTM generates more coherent topics and a more rational topic structure when compared to state-of-theart baselines. Our code is available at https: //github.com/nbnbhwyy/NSEM-GMHTM.

Layerwise universal adversarial attack on NLP models
Olga Tsymboi, Danil Malaev, Andrei Petrovskii and Ivan Oseledets 11:00-12:30 (Pier 7&8)
In this work, we examine the vulnerability of language models to universal adversarial triggers (UATs). We propose a new white-box approach to the construction of layerwise UATs (LUATs), which searches the triggers by perturbing hidden layers of a network. On the example of three transformer models and three datasets from the GLUE benchmark, we demonstrate that our method provides better transferability in a model-to-model setting with an average gain of 9.3% in the fooling rate over the baseline. Moreover, we investigate triggers transferability in the task-to-task setting. Using small subsets from the datasets similar to the target tasks for choosing a perturbed layer, we show that LUATs are more efficient than vanilla UATs by 7.1% in the fooling rate.

Robust Natural Language Understanding with Residual Attention Debiasing
Fei Wang, James Y. Huang, Tianyi Yan, Wenxuan Zhou and Muhao Chen 11:00-12:30 (Pier 7&8)
Natural language understanding (NLU) models often suffer from unintended dataset biases. Among bias mitigation methods, ensemble-based debiasing methods, especially product-of-experts (PoE), have stood out for their impressive empirical success. However, previous ensemble-based debiasing methods typically apply debiasing on top-level logits without directly addressing biased attention patterns. Attention serves as the main media of feature interaction and aggregation in PLMs and plays a crucial role in providing robust prediction. In this paper, we propose REsidual Attention Debiasing (READ), an end-to-end debiasing method that mitigates unintended biases from attention. Experiments on three NLU benchmarks show that READ significantly improves the OOD performance of BERT-based models, including +12.9% accuracy on HANS, +11.0% accuracy on FEVER-Symmetric, and +2.7% F1 on PAWS. Detailed analyses demonstrate the crucial role of unbiased attention in robust NLU models and that READ effectively mitigates biases in attention.

Deep Model Compression Also Helps Models Capture Ambiguity
Hancheol Park and Jong Park 11:00-12:30 (Pier 7&8)
Natural language understanding (NLU) tasks face a non-trivial amount of ambiguous samples where veracity of their labels is debatable among annotators. NLU models should thus account for such ambiguity, but they approximate the human opinion distributions quite poorly and tend to produce over-confident predictions. To address this problem, we must consider how to exactly capture the degree of relationship between each sample and its candidate classes. In this work, we propose a novel method with deep model compression and show how such relationship can be accounted for. We see that more reasonably represented relationships can be discovered in the lower layers and that validation accuracies are converging at these layers, which naturally leads to layer pruning. We also see that distilling the relationship knowledge from a lower layer helps models produce better distribution. Experimental results demonstrate that our method makes substantial improvement on quantifying ambiguity without gold distribution labels. As positive side-effects, our method is found to reduce the model size significantly and improve latency, both attractive aspects of NLU products.

On Prefix-tuning for Lightweight Out-of-distribution Detection
Yawen Ouyang, Yongchang Cao, Yuan Gao, Zhen Wu, Jianbing Zhang and Xinyu Dai 11:00-12:30 (Pier 7&8)
Out-of-distribution (OOD) detection, a fundamental task vexing real-world applications, has attracted growing attention in the NLP community. Recently fine-tuning based methods have made promising progress. However, it could be costly to store fine-tuned models for each

scenario. In this paper, we depart from the classic fine-tuning based OOD detection toward a parameter-efficient alternative, and propose an unsupervised prefix-tuning based OOD detection framework termed PTO. Additionally, to take advantage of optional training data labels and targeted OOD data, two practical extensions of PTO are further proposed. Overall, PTO and its extensions offer several key advantages of being lightweight, easy-to-reproduce, and theoretically justified. Experimental results show that our methods perform comparably to, even better than, existing fine-tuning based OOD detection approaches under a wide range of metrics, detection settings, and OOD types.

Towards Stable Natural Language Understanding via Information Entropy Guided Debiasing
Li Du, Xiao Ding, Zhouhao Sun, Ting Liu, Bing Qin and Jingshuo Liu   11:00-12:30 (Pier 7&8)
Although achieving promising performance, current Natural Language Understanding models tend to utilize dataset biases instead of learning the intended task, which always leads to performance degradation on out-of-distribution (OOD) samples. To increase the performance stability, previous debiasing methods *empirically* capture bias features from data to prevent the model from corresponding biases. However, our analyses show that the empirical debiasing methods may fail to capture part of the potential dataset biases and mistake semantic information of input text as biases, which limits the effectiveness of debiasing. To address these issues, we propose a debiasing framework IEGDB that comprehensively detects the dataset biases to induce a set of biased features, and then purifies the biased features with the guidance of information entropy. Experimental results show that IEGDB can consistently improve the stability of performance on OOD datasets for a set of widely adopted NLU models.

Transformer Language Models Handle Word Frequency in Prediction Head
Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi and Kentaro Inui   11:00-12:30 (Pier 7&8)
Prediction head is a crucial component of Transformer language models. Despite its direct impact on prediction, this component has often been overlooked in analyzing Transformers. In this study, we investigate the inner workings of the prediction head, specifically focusing on bias parameters. Our experiments with BERT and GPT-2 models reveal that the biases in their word prediction heads play a significant role in the models' ability to reflect word frequency in a corpus, aligning with the logit adjustment method commonly used in long-tailed learning. We also quantify the effect of controlling the biases in practical auto-regressive text generation scenarios; under a particular setting, more diverse text can be generated without compromising text quality.

Is Continuous Prompt a Combination of Discrete Prompts? Towards a Novel View for Interpreting Continuous Prompts
Tianjie Ju, Yubin Zheng, Hanyi Wang, Haodong Zhao and Gongshen Liu   11:00-12:30 (Pier 7&8)
The broad adoption of continuous prompts has brought state-of-the-art results on a diverse array of downstream natural language processing (NLP) tasks. Nonetheless, little attention has been paid to the interpretability and transferability of continuous prompts. Faced with the challenges, we investigate the feasibility of interpreting continuous prompts as the weighting of discrete prompts by jointly optimizing prompt fidelity and downstream fidelity. Our experiments show that: (1) one can always find a combination of discrete prompts as the replacement of continuous prompts that performs well on downstream tasks; (2) our interpretable framework faithfully reflects the reasoning process of source prompts; (3) our interpretations provide effective readability and plausibility, which is helpful to understand the decision-making of continuous prompts and discover potential shortcuts. Moreover, through the bridge constructed between continuous prompts and discrete prompts using our interpretations, it is promising to implement the cross-model transfer of continuous prompts without extra training signals. We hope this work will lead to a novel perspective on the interpretations of continuous prompts.

Causal interventions expose implicit situation models for commonsense language understanding
Takateru Yamakoshi, James L McClelland, Adele Goldberg and Robert Hawkins   11:00-12:30 (Pier 7&8)
Accounts of human language processing have long appealed to implicit "situation models" that enrich comprehension with relevant but unstated world knowledge. Here, we apply causal intervention techniques to recent transformer models to analyze performance on the Winograd Schema Challenge (WSC), where a single context cue shifts interpretation of an ambiguous pronoun. We identify a relatively small circuit of attention heads that are responsible for propagating information from the context word that guides which of the candidate noun phrases the pronoun ultimately attends to. We then compare how this circuit behaves in a closely matched "syntactic" control where the situation model is not strictly necessary. These analyses suggest a distinct pathway through which implicit situation models may be constructed to guide pronoun resolution

Constructing Word-Context-Coupled Space Aligned with Associative Knowledge Relations for Interpretable Language Modeling
Fanyu Wang and Zhenping Xie   11:00-12:30 (Pier 7&8)
As the foundation of current natural language processing methods, pre-trained language model has achieved excellent performance. However, the black-box structure of the deep neural network in pre-trained language models seriously limits the interpretability of the language modeling process. After revisiting the coupled requirement of deep neural representation and semantics logic of language modeling, a Word-Context-Coupled Space (W2CSpace) is proposed by introducing the alignment processing between uninterpretable neural representation and interpretable statistical logic. Moreover, a clustering process is also designed to connect the word- and context-level semantics. Specifically, an associative knowledge network (AKN), considered interpretable statistical logic, is introduced in the alignment process for word-level semantics. Furthermore, the context-relative distance is employed as the semantic feature for the downstream classifier, which is greatly different from the current uninterpretable semantic representations of pre-trained models. Our experiments for performance evaluation and interpretable analysis are executed on several types of datasets, including SIGHAN, Weibo, and ChnSenti. Wherein a novel evaluation strategy for the interpretability of machine learning is first proposed. According to the experimental results, our language model can achieve better performance and highly credible interpretable ability compared to related state-of-the-art methods.

Easy to Decide, Hard to Agree: Reducing Disagreements Between Saliency Methods
Josip Jukić, Martin Tutek and Jan Snajder   11:00-12:30 (Pier 7&8)
A popular approach to unveiling the black box of neural NLP models is to leverage saliency methods, which assign scalar importance scores to each input component. A common practice for evaluating whether an interpretability method is faithful has been to use evaluation-by-agreement – if multiple methods agree on an explanation, its credibility increases. However, recent work has found that saliency methods exhibit weak rank correlations even when applied to the same model instance and advocated for alternative diagnostic methods. In our work, we demonstrate that rank correlation is not a good fit for evaluating agreement and argue that Pearson-r is a better-suited alternative. We further show that regularization techniques that increase faithfulness of attention explanations also increase agreement between saliency methods. By connecting our findings to instance categories based on training dynamics, we show that the agreement of saliency method explanations is very low for easy-to-learn instances. Finally, we connect the improvement in agreement across instance categories to local representation space statistics of instances, paving the way for work on analyzing which intrinsic model properties improve their predisposition to interpretability methods.

HyHTM: Hyperbolic Geometry-based Hierarchical Topic Model
Simra Shahid, Tanay Anand, Nikitha Srikanth, Sumit Bhatia, Balaji Krishnamurthy and nikaash puri   11:00-12:30 (Pier 7&8)
Hierarchical Topic Models (HTMs) are useful for discovering topic hierarchies in a collection of documents. However, traditional HTMs often produce hierarchies where lower-level topics are unrelated and not specific enough to their higher-level topics. Additionally, these meth-

ods can be computationally expensive. We present HyHTM - a Hyperbolic geometry-based Hierarchical Topic Model - that addresses these limitations by incorporating hierarchical information from hyperbolic geometry to explicitly model hierarchies in topic models. Experimental results with four baselines show that HyHTM can better attend to parent-child relationships among topics. HyHTM produces coherent topic hierarchies that specialize in granularity from generic higher-level topics to specific lower-level topics. Further, our model is significantly faster and leaves a much smaller memory footprint than our best-performing baseline. We have made the source code for our algorithm publicly accessible.

Do PLMs Know and Understand Ontological Knowledge?

Weiqi Wu, Chengyue Jiang, Yong Jiang, Pengjun Xie and Kewei Tu   11:00-12:30 (Pier 7&8)
Ontological knowledge, which comprises classes and properties and their relationships, is integral to world knowledge. It is significant to explore whether Pretrained Language Models (PLMs) know and understand such knowledge. However, existing PLM-probing studies focus mainly on factual knowledge, lacking a system- atic probing of ontological knowledge. In this paper, we focus on probing whether PLMs store ontological knowledge and have a semantic un- derstanding of the knowledge rather than rote memorization of the surface form. To probe whether PLMs know ontological knowledge, we investigate how well PLMs memorize: (1) types of entities; (2) hierarchical relationships among classes and properties, e.g., Person is a subclass of Animal and Member of Sports Team is a subproperty of Member of ; (3) domain and range constraints of properties, e.g., the subject of Member of Sports Team should be a Person and the object should be a Sports Team. To further probe whether PLMs truly understand ontological knowledge beyond memorization, we comprehensively study whether they can reliably perform logical reasoning with given knowledge according to ontological entailment rules. Our probing results show that PLMs can memorize certain ontological knowledge and utilize implicit knowledge in reasoning. How- ever, both the memorizing and reasoning per- formances are less than perfect, indicating in- complete knowledge and understanding.

Bias Beyond English: Counterfactual Tests for Bias in Sentiment Analysis in Four Languages

Seraphina Goldfarb-Tarrant, Adam Lopez, Roi Blanco and Diego Marcheggiani   11:00-12:30 (Pier 7&8)
Sentiment analysis (SA) systems are used in many products and hundreds of languages. Gender and racial biases are well-studied in English SA systems, but understudied in other languages, with few resources for such studies. To remedy this, we build a counterfactual evaluation corpus for gender and racial/migrant bias in four languages. We demonstrate its usefulness by answering a simple but important question that an engineer might need to answer when deploying a system: What biases do systems import from pre-trained models when compared to a baseline with no pre-training? Our evaluation corpus, by virtue of being counterfactual, not only reveals which models have less bias, but also pinpoints changes in model bias behaviour, which enables more targeted mitigation strategies. We release our code and evaluation corpora to facilitate future research.

A Multi-dimensional study on Bias in Vision-Language models

Gabriele Ruggeri and Debora Nozza   11:00-12:30 (Pier 7&8)
In recent years, joint Vision-Language (VL) models have increased in popularity and capability. Very few studies have attempted to investigate bias in VL models, even though it is a well-known issue in both individual modalities. This paper presents the first multi-dimensional analysis of bias in English VL models, focusing on gender, ethnicity, and age as dimensions. When subjects are input as images, pre-trained VL models complete a neutral template with a hurtful word 5% of the time, with higher percentages for female and young subjects. Bias presence in downstream models has been tested on Visual Question Answering. We developed a novel bias metric called the Vision-Language Association Test based on questions designed to elicit biased associations between stereotypical concepts and targets. Our findings demonstrate that pre-trained VL models contain biases that are perpetuated in downstream tasks.

D-CALM: A Dynamic Clustering-based Active Learning Approach for Mitigating Bias

Sabit Hassan and Malihe Alikhani   11:00-12:30 (Pier 7&8)
Despite recent advancements, NLP models continue to be vulnerable to bias. This bias often originates from the uneven distribution of real-world data and can propagate through the annotation process. Escalated integration of these models in our lives calls for methods to mitigate bias without overbearing annotation costs. While active learning (AL) has shown promise in training models with a small amount of annotated data, AL's reliance on the model's behavior for selective sampling can lead to an accumulation of unwanted bias rather than bias mitigation. However, infusing clustering with AL can overcome the bias issue of both AL and traditional annotation methods while exploiting AL's annotation efficiency. In this paper, we propose a novel adaptive clustering-based active learning algorithm, D-CALM, that dynamically adjusts clustering and annotation efforts in response to an estimated classifier error-rate. Experiments on eight datasets for a diverse set of text classification tasks, including emotion, hatespeech, dialog act, and book type detection, demonstrate that our proposed algorithm significantly outperforms baseline AL approaches with both pretrained transformers and traditional Support Vector Machines. D-CALM showcases robustness against different measures of information gain and, as evident from our analysis of label and error distribution, can significantly reduce unwanted model bias.

Causal-Debias: Unifying Debiasing in Pretrained Language Models and Fine-tuning via Causal Invariant Learning

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang and Ting Zhong   11:00-12:30 (Pier 7&8)
Demographic biases and social stereotypes are common in pretrained language models (PLMs), and a burgeoning body of literature focuses on removing the unwanted stereotypical associations from PLMs. However, when fine-tuning these bias-mitigated PLMs in downstream natural language processing (NLP) applications, such as sentiment classification, the unwanted stereotypical associations resurface or even get amplified. Since pretrain&fine-tune is a major paradigm in NLP applications, separating the debiasing procedure of PLMs from fine-tuning would eventually harm the actual downstream utility. In this paper, we propose a unified debiasing framework Causal-Debias to remove unwanted stereotypical associations in PLMs during fine-tuning. Specifically, CausalDebias mitigates bias from a causal invariant perspective by leveraging the specific downstream task to identify bias-relevant and labelrelevant factors. We propose that bias-relevant factors are non-causal as they should have little impact on downstream tasks, while labelrelevant factors are causal. We perform interventions on non-causal factors in different demographic groups and design an invariant risk minimization loss to mitigate bias while maintaining task performance. Experimental results on three downstream tasks show that our proposed method can remarkably reduce unwanted stereotypical associations after PLMs are finetuned, while simultaneously minimizing the impact on PLMs and downstream applications.

Uncovering and Categorizing Social Biases in Text-to-SQL

Yan Liu, Yan Gao, Zhe Su, Xiaokang Chen, Elliott Ash and Jian-Guang LOU   11:00-12:30 (Pier 7&8)
Large pre-trained language models are acknowledged to carry social bias towards different demographics, which can further amplify existing stereotypes in our society and cause even more harm. Text-to-SQL is an important task, models of which are mainly adopted by administrative industries, where unfair decisions may lead to catastrophic consequences. However, existing Text-to-SQL models are trained on clean, neutral datasets, such as Spider and WikiSQL. This, to some extent, cover up social bias in models under ideal conditions, which nevertheless may emerge in real application scenarios. In this work, we aim to uncover and mitigate social bias in Text-to-SQL models. We summarize the categories of social bias that may occur in structural data for Text-to-SQL models. We build test benchmarks and reveal that models with similar task accuracy can contain social bias at very different rates. We show how to take advantage of our methodology to assess and mitigate social bias in the downstream Text-to-SQL task.

117

Nichelle and Nancy: The Influence of Demographic Attributes and Tokenization Length on First Name Biases
Haozhe An and Rachel Rudinger 11:00-12:30 (Pier 7&8)
Through the use of first name substitution experiments, prior research has demonstrated the tendency of social commonsense reasoning models to systematically exhibit social biases along the dimensions of race, ethnicity, and gender (An et al., 2023). Demographic attributes of first names, however, are strongly correlated with corpus frequency and tokenization length, which may influence model behavior independent of or in addition to demographic factors. In this paper, we conduct a new series of first name substitution experiments that measures the influence of these factors while controlling for the others. We find that demographic attributes of a name (race, ethnicity, and gender) and name tokenization length are both factors that systematically affect the behavior of social commonsense reasoning models.

With Prejudice to None: A Few-Shot, Multilingual Transfer Learning Approach to Detect Social Bias in Low Resource Languages
Nihar Ranjan Sahoo, Niteesh Kumar Reddy Mallela and Pushpak Bhattacharyya 11:00-12:30 (Pier 7&8)
In this paper, we describe our work on social bias detection in a low-resource multilingual setting in which the languages are from two very divergent families- Indo-European (English, Hindi, and Italian) and Altaic (Korean). Currently, the majority of the social bias datasets available are in English and this inhibits progress on social bias detection in low-resource languages. To address this problem, we introduce a new dataset for social bias detection in Hindi and investigate multilingual transfer learning using publicly available English, Italian, and Korean datasets. The Hindi dataset contains 9k social media posts annotated for (i) binary bias labels (bias/neutral), (ii) binary labels for sentiment (positive/negative), (iii) target groups for each bias category, and (iv) rationale for annotated bias labels (a short piece of text). We benchmark our Hindi dataset using different multilingual models, with XLM-R achieving the best performance of 80.8 macro-F1 score. Our results show that the detection of social biases in resource-constrained languages such as Hindi and Korean may be improved with the use of a similar dataset in English. We also show that translating all datasets into English does not work effectively for detecting social bias, since the nuances of source language are lost in translation. All the scripts and datasets utilized in this study will be publicly available.

Large Language Models Meet NL2Code: A Survey
Daoguang Zan, Bei Chen, Fengji Zhang, Dianjie Lu, Bingchao Wu, Bei Guan, Wang Yongji and Jian-Guang LOU 11:00-12:30 (Pier 7&8)
The task of generating code from a natural language description, or NL2Code, is considered a pressing and significant challenge in code intelligence. Thanks to the rapid development of pre-training techniques, surging large language models are being proposed for code, sparking the advances in NL2Code. To facilitate further research and applications in this field, in this paper, we present a comprehensive survey of 27 existing large language models for NL2Code, and also review benchmarks and metrics. We provide an intuitive comparison of all existing models on the HumanEval benchmark. Through in-depth observation and analysis, we provide some insights and conclude that the key factors contributing to the success of large language models for NL2Code are "Large Size, Premium Data, Expert Tuning". In addition, we discuss challenges and opportunities regarding the gap between models and humans. We also create a website https://nl2code.github.io to track the latest progress through crowd-sourcing. To the best of our knowledge, this is the first survey of large language models for NL2Code, and we believe it will contribute to the ongoing development of the field.

An Exploratory Study on Model Compression for Text-to-SQL
Shuo Sun, Yuze Gao, Yuchen Zhang, Jian Su, Bin Chen, Yingzhan Lin and Shuqi Sun 11:00-12:30 (Pier 7&8)
Text-to-SQL translates user queries into SQL statements that can retrieve relevant answers from relational databases. Recent approaches to Text-to-SQL rely on pre-trained language models that are computationally expensive and technically challenging to deploy in real-world applications that require real-time or on-device processing capabilities. In this paper, we perform a focused study on the feasibility of applying recent model compression techniques to sketch-based and sequence-to-sequence Text-to-SQL models. Our results reveal that sketch-based Text-to-SQL models generally have higher inference efficiency and respond better to model compression than sequence-to-sequence models, making them ideal for real-world deployments, especially in use cases with simple SQL statements.

Follow the leader(board) with confidence: Estimating p-values from a single test set with item and response variance
Shira Wein, Christopher Homan, Lora Aroyo and Chris Welty 11:00-12:30 (Pier 7&8)
Among the problems with leaderboard culture in NLP has been the widespread lack of confidence estimation in reported results. In this work, we present a framework and simulator for estimating p-values for comparisons between the results of two systems, in order to understand the confidence that one is actually better (i.e. ranked higher) than the other. What has made this difficult in the past is that each system must itself be evaluated by comparison to a gold standard. We define a null hypothesis that each system's metric scores are drawn from the same distribution, using variance found naturally (though rarely reported) in test set items and individual labels on an item (responses) to produce the metric distributions. We create a test set that evenly mixes the responses of the two systems under the assumption the null hypothesis is true. Exploring how to best estimate the true p-value from a single test set under different metrics, tests, and sampling methods, we find that the presence of response variance (from multiple raters or multiple model versions) has a profound impact on p-value estimates for model comparison, and that choice of metric and sampling method is critical to providing statistical guarantees on model comparisons.

Can Language Models Be Specific? How?
Jie Huang, Kevin Chen-Chuan Chang, Jinjun Xiong and Wen-mei Hwu 11:00-12:30 (Pier 7&8)
"He is a person", "Paris is located on the earth". Both statements are correct but meaningless - due to lack of specificity. In this paper, we propose to measure how specific the language of pre-trained language models (PLMs) is. To achieve this, we introduce a novel approach to build a benchmark for specificity testing by forming masked token prediction tasks with prompts. For instance, given "Toronto is located in [MASK].", we want to test whether a more specific answer will be better filled in by PLMs, e.g., Ontario instead of Canada. From our evaluations, we show that existing PLMs have only a slight preference for more specific answers. We identify underlying factors affecting the specificity and design two prompt-based methods to improve the specificity. Results show that the specificity of the models can be improved by the proposed methods without additional training. We hope this work can bring to awareness the notion of specificity of language models and encourage the research community to further explore this important but understudied problem.

Are Layout-Infused Language Models Robust to Layout Distribution Shifts? A Case Study with Scientific Documents
Catherine Chen, Zejiang Shen, Dan Klein, Gabriel Stanovsky, Doug Downey and Kyle Lo 11:00-12:30 (Pier 7&8)
Recent work has shown that infusing layout features into language models (LMs) improves processing of visually-rich documents such as scientific papers. Layout-infused LMs are often evaluated on documents with familiar layout features (e.g., papers from the same publisher), but in practice models encounter documents with unfamiliar distributions of layout features, such as new combinations of text sizes and styles, or new spatial configurations of textual elements. In this work we test whether layout-infused LMs are robust to layout distribution shifts. As a case study we use the task of scientific document structure recovery, segmenting a scientific paper into its structural categories (e.g., "title", "caption", "reference"). To emulate distribution shifts that occur in practice we re-partition the GROTOAP2 dataset. We find that under layout distribution shifts model performance degrades by up to 20 F1. Simple training strategies, such as increasing training diversity, can reduce this degradation by over 35% relative F1; however, models fail to reach in-distribution performance in any tested out-of-distribution conditions. This work highlights the need to consider layout distribution shifts during model evaluation, and presents a methodology for conducting such evaluations.

A Comparative Analysis of the Effectiveness of Rare Tokens on Creative Expression using ramBERT

Youbin Lee, Deokgi Kim, Byung-Won On and Ingyu Lee   11:00-12:30 (Pier 7&8)

Until now, few studies have been explored on Automated Creative Essay Scoring (ACES), in which a pre-trained model automatically labels an essay as a creative or a non-creative. Since the creativity evaluation of essays is very subjective, each evaluator often has his or her own criteria for creativity. For this reason, quantifying creativity in essays is very challenging. In this work, as one of preliminary studies in developing a novel model for ACES, we deeply investigate the correlation between creative essays and expressiveness. Specifically, we explore how rare tokens affect the evaluation of creativity for essays. For such a journey, we present five distinct methods to extract rare tokens, and conduct a comparative study on the correlation between rare tokens and creative essay evaluation results using BERT. Our experimental results showed clear correlation between rare tokens and creative essays. In all test sets, accuracies of our rare token masking-based BERT (ramBERT) model were improved over the existing BERT model up to 14%.

A Survey on Zero Pronoun Translation

Longyue Wang, Siyou Liu, Mingzhou Xu, Linfeng Song, Shuming Shi and Zhaopeng Tu   11:00-12:30 (Pier 7&8)

Zero pronouns (ZPs) are frequently omitted in pro-drop languages (e.g. Chinese, Hungarian, and Hindi), but should be recalled in non-pro-drop languages (e.g. English). This phenomenon has been studied extensively in machine translation (MT), as it poses a significant challenge for MT systems due to the difficulty in determining the correct antecedent for the pronoun. This survey paper highlights the major works that have been undertaken in zero pronoun translation (ZPT) after the neural revolution so that researchers can recognize the current state and future directions of this field. We provide an organization of the literature based on evolution, dataset, method, and evaluation. In addition, we compare and analyze competing models and evaluation metrics on different benchmarks. We uncover a number of insightful findings such as: 1) ZPT is in line with the development trend of large language model; 2) data limitation causes learning bias in languages and domains; 3) performance improvements are often reported on single benchmarks, but advanced methods are still far from real-world use; 4) general-purpose metrics are not reliable on nuances and complexities of ZPT, emphasizing the necessity of targeted metrics; 5) apart from commonly-cited errors, ZPs will cause risks of gender bias.

A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty and Jimmy Xiangji Huang 11:00-12:30 (Pier 7&8)

The development of large language models (LLMs) such as ChatGPT has brought a lot of attention recently. However, their evaluation in the benchmark academic datasets remains under-explored due to the difficulty of evaluating the generative outputs produced by this model against the ground truth. In this paper, we aim to present a thorough evaluation of ChatGPT's performance on diverse academic datasets, covering tasks like question-answering, text summarization, code generation, commonsense reasoning, mathematical problem-solving, machine translation, bias detection, and ethical considerations. Specifically, we evaluate ChatGPT across 140 tasks and analyze 255K responses it generates in these datasets. This makes our work the largest evaluation of ChatGPT in NLP benchmarks. In short, our study aims to validate the strengths and weaknesses of ChatGPT in various tasks and provide insights for future research using LLMs. We also report a new emergent ability to follow multi-query instructions that we mostly found in ChatGPT and other instruction-tuned models. Our extensive evaluation shows that even though ChatGPT is capable of performing a wide variety of tasks, and may obtain impressive performance in several benchmark datasets, it is still far from achieving the ability to reliably solve many challenging tasks. By providing a thorough assessment of ChatGPT's performance across diverse NLP tasks, this paper sets the stage for a targeted deployment of ChatGPT-like LLMs in real-world applications.

Pulling Out All The Full Stops: Punctuation Sensitivity in Neural Machine Translation and Evaluation

Prathyusha Jwalapuram   11:00-12:30 (Pier 7&8)

Much of the work testing machine translation systems for robustness and sensitivity has been adversarial or tended towards testing noisy input such as spelling errors, or non-standard input such as dialects. In this work, we take a step back to investigate a sensitivity problem that can seem trivial and is often overlooked: punctuation. We perform basic sentence-final insertion and deletion perturbation tests with full stops, exclamation and questions marks across source languages and demonstrate a concerning finding: commercial, production-level machine translation systems are vulnerable to mere single punctuation insertion or deletion, resulting in unreliable translations. Moreover, we demonstrate that both string-based and model-based evaluation metrics also suffer from this vulnerability, producing significantly different scores when translations only differ in a single punctuation, with model-based metrics penalizing each punctuation differently. Our work calls into question the reliability of machine translation systems and their evaluation metrics, particularly for real-world use cases, where inconsistent punctuation is often the most common and the least disruptive noise.

Session 2 - 14:00-15:30

Theme: Reality Check

14:00-15:30 (Metropolitan East)

Credible without Credit: Domain Experts Assess Generative Language Models

Denis Peskoff and Brandon Stewart   14:00-14:15 (Metropolitan East)

Language models have recently broken into the public consciousness with the release of the wildly popular ChatGPT. Commentators have argued that language models could replace search engines, make college essays obsolete, or even write academic research papers. All of these tasks rely on accuracy of specialized information which can be difficult to assess for non-experts. Using 10 domain experts across science and culture, we provide an initial assessment of the coherence, conciseness, accuracy, and sourcing of two language models across 100 expert-written questions. While we find the results are consistently cohesive and concise, we find that they are mixed in their accuracy. These results raise questions of the role language models should play in general-purpose and expert knowledge seeking.

What's the Meaning of Superhuman Performance in Today's NLU?

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard H Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova and Roberto Navigli   14:15-14:30 (Metropolitan East)

In the last five years, there has been a significant focus in Natural Language Processing (NLP) on developing larger Pretrained Language Models (PLMs) and introducing benchmarks such as SuperGLUE and SQuAD to measure their abilities in language understanding, reasoning, and reading comprehension. These PLMs have achieved impressive results on these benchmarks, even surpassing human performance in some cases. This has led to claims of superhuman capabilities and the provocative idea that certain tasks have been solved. In this position

paper, we take a critical look at these claims and ask whether PLMs truly have superhuman abilities and what the current benchmarks are really evaluating. We show that these benchmarks have serious limitations affecting the comparison between humans and PLMs and provide recommendations for fairer and more transparent benchmarks.

Why Aren't We NER Yet? Artifacts of ASR Errors in Named Entity Recognition in Spontaneous Speech Transcripts
Piotr Szymański, Lukasz Augustyniak, Mikolaj Morzy, Adrian Szymczak, Krzysztof Surdyk and Piotr Żelasko 14:30-14:45 (Metropolitan East)
Transcripts of spontaneous human speech present a significant obstacle for traditional NER models. The lack of grammatical structure of spoken utterances and word errors introduced by the ASR make downstream NLP tasks challenging. In this paper, we examine in detail the complex relationship between ASR and NER errors which limit the ability of NER models to recover entity mentions from spontaneous speech transcripts. Using publicly available benchmark datasets (SWNE, Earnings-21, OntoNotes), we present the full taxonomy of ASR-NER errors and measure their true impact on entity recognition. We find that NER models fail spectacularly even if no word errors are introduced by the ASR. We also show why the F1 score is inadequate to evaluate NER models on conversational transcripts.

Mind the Gap between the Application Track and the Real World
Ananya Ganesh, Jie Cao, E. Margaret Perkoff, Rosy Southwell, Martha Palmer and Katharina Kann 14:45-15:00 (Metropolitan East)
Recent advances in NLP have led to a rise in inter-disciplinary and application-oriented research. While this demonstrates the growing real-world impact of the field, research papers frequently feature experiments that do not account for the complexities of realistic data and environments. To explore the extent of this gap, we investigate the relationship between the real-world motivations described in NLP papers and the models and evaluation which comprise the proposed solution. We first survey papers from the NLP Applications track from ACL 2020 and EMNLP 2020, asking which papers have differences between their stated motivation and their experimental setting, and if so, mention them. We find that many papers fall short of considering real-world input and output conditions due to adopting simplified modeling or evaluation settings. As a case study, we then empirically show that the performance of an educational dialog understanding system deteriorates when used in a realistic classroom environment.

Weaker Than You Think: A Critical Look at Weakly Supervised Learning
Dawei Zhu, Xiaoyu Shen, Marius Mosbach, Andreas Joseph Stephan and Dietrich Klakow 15:00-15:15 (Metropolitan East)
Weakly supervised learning is a popular approach for training machine learning models in low-resource settings. Instead of requesting high-quality yet costly human annotations, it allows training models with noisy annotations obtained from various weak sources. Recently, many sophisticated approaches have been proposed for robust training under label noise, reporting impressive results. In this paper, we revisit the setup of these approaches and find that the benefits brought by these approaches are significantly overestimated. Specifically, we find that the success of existing weakly supervised learning approaches heavily relies on the availability of clean validation samples which, as we show, can be leveraged much more efficiently by simply training on them. After using these clean labels in training, the advantages of using these sophisticated approaches are mostly wiped out. This remains true even when reducing the size of the available clean data to just five samples per class, making these approaches impractical. To understand the true value of weakly supervised learning, we thoroughly analyze diverse NLP datasets and tasks to ascertain when and why weakly supervised approaches work. Based on our findings, we provide recommendations for future research.

On "Scientific Debt" in NLP: A Case for More Rigour in Language Model Pre-Training Research
Made Nindyatama Nityasya, Haryo Akbarianto Wibowo, Alham Fikri Aji, Genta Indra Winata, Radityo Eko Prasojo, Phil Blunsom and Adhiguna Kuncoro 15:15-15:30 (Metropolitan East)
This evidence-based position paper critiques current research practices within the language model pre-training literature. Despite rapid recent progress afforded by increasingly better pre-trained language models (PLMs), current PLM research practices often conflate different possible sources of model improvement, without conducting proper ablation studies and principled comparisons between different models under comparable conditions. These practices (i) leave us ill-equipped to understand which pre-training approaches should be used under what circumstances; (ii) impede reproducibility and credit assignment; and (iii) render it difficult to understand: "How exactly does each factor contribute to the progress that we have today?" We provide a case in point by revisiting the success of BERT over its baselines, ELMo and GPT-1, and demonstrate how — under comparable conditions where the baselines are tuned to a similar extent — these baselines (and even-simpler variants thereof) can, in fact, achieve competitive or better performance than BERT. These findings demonstrate how disentangling different factors of model improvements can lead to valuable new insights. We conclude with recommendations for how to encourage and incentivize this line of work, and accelerate progress towards a better and more systematic understanding of what factors drive the progress of our foundation models today.

Machine Learning for NLP

14:00-15:30 (Metropolitan Centre)

Bridging the Gap between Decision and Logits in Decision-based Knowledge Distillation for Pre-trained Language Models
Qinhong Zhou, Zonghan Yang, Peng Li and Yang Liu 14:00-14:15 (Metropolitan Centre)
Conventional knowledge distillation (KD) methods require access to the internal information of teachers, e.g., logits. However, such information may not always be accessible for large pre-trained language models (PLMs). In this work, we focus on decision-based KD for PLMs, where only teacher decisions (i.e., top-1 labels) are accessible. Considering the information gap between logits and decisions, we propose a novel method to estimate logits from the decision distributions. Specifically, decision distributions can be both derived as a function of logits theoretically and estimated with test-time data augmentation empirically. By combining the theoretical and empirical estimations of the decision distributions together, the estimation of logits can be successfully reduced to a simple root-finding problem. Extensive experiments show that our method significantly outperforms strong baselines on both natural language understanding and machine reading comprehension datasets.

f-Divergence Minimization for Sequence-Level Knowledge Distillation
Yuqiao Wen, Zichao Li, Wenyu Du and Lili Mou 14:15-14:30 (Metropolitan Centre)
Knowledge distillation (KD) is the process of transferring knowledge from a large model to a small one. It has gained increasing attention in the natural language processing community, driven by the demands of compressing ever-growing language models. In this work, we propose an FDISTILL framework, which formulates sequence-level knowledge distillation as minimizing a generalized f-divergence function. We propose four distilling variants under our framework and show that existing SeqKD and ENGINE approaches are approximations of our FDISTILL methods. We further derive step-wise decomposition for our FDISTILL, reducing intractable sequence-level divergence to word-level losses that can be computed in a tractable manner. Experiments across four datasets show that our methods outperform existing KD approaches, and that our symmetric distilling losses can better force the student to learn from the teacher distribution.

Patton: Language Model Pretraining on Text-Rich Networks

Bowen Jin, Wentao Zhang, Yu Zhang, Yu Meng, Xinyang Zhang, Qi Zhu and Jiawei Han   14:30-14:45 (Metropolitan Centre)

A real-world text corpus sometimes comprises not only text documents, but also semantic links between them (e.g., academic papers in a bibliographic network are linked by citations and co-authorships). Text documents and semantic connections form a text-rich network, which empowers a wide range of downstream tasks such as classification and retrieval. However, pretraining methods for such structures are still lacking, making it difficult to build one generic model that can be adapted to various tasks on text-rich networks. Current pretraining objectives, such as masked language modeling, purely model texts and do not take inter-document structure information into consideration. To this end, we propose our PretrAining on TexT-Rich NetwOrk framework Patton. Patton includes two pretraining strategies: network-contextualized masked language modeling and masked node prediction, to capture the inherent dependency between textual attributes and network structure. We conduct experiments on four downstream tasks in five datasets from both academic and e-commerce domains, where Patton outperforms baselines significantly and consistently.

Binary and Ternary Natural Language Generation

Zechun Liu, Barlas Oguz, Aasish Pappu, Yangyang Shi and Raghuraman Krishnamoorthi   14:45-15:00 (Metropolitan Centre)

Ternary and binary neural networks enable multiplication-free computation and promise multiple orders of magnitude efficiency gains over full-precision networks if implemented on specialized hardware. However, since both the parameter and the output space are highly discretized, such networks have proven very difficult to optimize. The difficulties are compounded for the class of transformer text generation models due to the sensitivity of the attention operation to quantization and the noise-compounding effects of autoregressive decoding in the high-cardinality output space. We approach the problem with a mix of statistics-based quantization for the weights and elastic quantization of the activations and demonstrate the first ternary and binary transformer models on the downstream tasks of summarization and machine translation. Our ternary BART base achieves an R1 score of 41 on the CNN/DailyMail benchmark, which is merely 3.9 points behind the full model while being 16x more efficient. Our binary model, while less accurate, achieves a highly non-trivial score of 35.6. For machine translation, we achieved BLEU scores of 21.7 and 17.6 on the WMT16 En-Ro benchmark, compared with a full precision mBART model score of 26.8. We also compare our approach in the 8-bit activation setting, where our ternary and even binary weight models can match or outperform the best existing 8-bit weight models in the literature. Our code and models are available at: https://github.com/facebookresearch/Ternary_Binary_Transformer.

Pruning Pre-trained Language Models Without Fine-Tuning

Ting Jiang, deqing wang, Fuzhen Zhuang, Ruobing Xie and Feng Xia   15:00-15:15 (Metropolitan Centre)

To overcome the overparameterized problem in Pre-trained Language Models (PLMs), pruning is widely used as a simple and straightforward compression method by directly removing unimportant weights. Previous first-order methods successfully compress PLMs to extremely high sparsity with little performance drop. These methods, such as movement pruning, use first-order information to prune PLMs while fine-tuning the remaining weights. In this work, we argue fine-tuning is redundant for first-order pruning, since first-order pruning is sufficient to converge PLMs to downstream tasks without fine-tuning. Under this motivation, we propose Static Model Pruning (SMP), which only uses first-order pruning to adapt PLMs to downstream tasks while achieving the target sparsity level. In addition, we also design a new masking function and training objective to further improve SMP. Extensive experiments at various sparsity levels show SMP has significant improvements over first-order and zero-order methods. Unlike previous first-order methods, SMP is also applicable to low sparsity and outperforms zero-order methods. Meanwhile, SMP is more parameter efficient than other methods due to it does not require fine-tuning.

Small Pre-trained Language Models Can be Fine-tuned as Large Models via Over-Parameterization

Ze-Feng Gao, Kun Zhou, Peiyu Liu, Wayne Xin Zhao and Ji-Rong Wen   15:15-15:30 (Metropolitan Centre)

By scaling the model size, large pre-trained language models (PLMs) have shown remarkable performance in various natural language processing tasks, mostly outperforming small PLMs by a large margin. However, due to the high computational cost, the huge number of parameters also restricts the applicability of large PLMs in real-world systems. In this paper, we focus on scaling up the parameters of PLMs *only during* fine-tuning, to benefit from the over-parameterization, while without increasing *the inference latency*. Given a relatively small PLM, we over-parameterize it by employing a matrix product operator, an efficient and almost lossless decomposition method to factorize its contained parameter matrices into a set of higher-dimensional tensors. Considering the efficiency, we further propose both static and dynamic strategies to select the most important parameter matrices for over-parameterization. Extensive experiments have demonstrated that our approach can significantly boost the fine-tuning performance of small PLMs and even help small PLMs outperform $3\times$ parameterized larger ones. Our code is publicly available at https://github.com/zfgao66/OPF.

Machine Translation

14:00-15:30 (Metropolitan West)

Extrinsic Evaluation of Machine Translation Metrics

Nikita Moghe, Tom Sherborne, Mark Steedman and Alexandra Birch   14:00-14:15 (Metropolitan West)

Automatic machine translation (MT) metrics are widely used to distinguish the quality of machine translation systems across relatively large test sets (system-level evaluation). However, it is unclear if automatic metrics are reliable at distinguishing good translations from bad translations at the sentence level (segment-level evaluation). In this paper, we investigate how useful MT metrics are at detecting the segment-level quality by correlating metrics with how useful the translations are for downstream task. We evaluate the segment-level performance of the most widely used MT metrics (chrF, COMET, BERTScore, etc.) on three downstream cross-lingual tasks (dialogue state tracking, question answering, and semantic parsing). For each task, we only have access to a monolingual task-specific model and a translation model. We calculate the correlation between the metric's ability to predict a good/bad translation with the success/failure on the final task for the machine translated test sentences. Our experiments demonstrate that all metrics exhibit negligible correlation with the extrinsic evaluation of the downstream outcomes. We also find that the scores provided by neural metrics are not interpretable, in large part due to having undefined ranges. We synthesise our analysis into recommendations for future MT metrics to produce labels rather than scores for more informative interaction between machine translation and multilingual language understanding.

[TACL] FRMT: A Benchmark for Few-Shot Region-Aware Machine Translation

Parker Riley, Timothy Dozat, Jan Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat and Noah Constant   14:15-14:30 (Metropolitan West)

We present FRMT, a new dataset and evaluation benchmark for Few-shot Region-aware Machine Translation, a type of style-targeted translation. The dataset consists of professional translations from English into two regional variants each of Portuguese and Mandarin Chinese. Source documents are selected to enable detailed analysis of phenomena of interest, including lexically distinct terms and distractor

terms. We explore automatic evaluation metrics for FRMT and validate their correlation with expert human evaluation across both region-matched and mismatched rating scenarios. Finally, we present a number of baseline models for this task, and offer guidelines for how researchers can train, evaluate, and compare their own models.Our dataset and evaluation code are publicly available: https://anonymized.

Knowledge Transfer in Incremental Learning for Multilingual Neural Machine Translation
Kaiyu Huang, Peng Li, Jin Ma, Ting Yao and Yang Liu   14:30-14:45 (Metropolitan West)
In the real-world scenario, a longstanding goal of multilingual neural machine translation (MNMT) is that a single model can incrementally adapt to new language pairs without accessing previous training data. In this scenario, previous studies concentrate on overcoming catastrophic forgetting while lacking encouragement to learn new knowledge from incremental language pairs, especially when the incremental language is not related to the set of original languages. To better acquire new knowledge, we propose a knowledge transfer method that can efficiently adapt original MNMT models to diverse incremental language pairs. The method flexibly introduces the knowledge from an external model into original models, which encourages the models to learn new language pairs, completing the procedure of knowledge transfer. Moreover, all original parameters are frozen to ensure that translation qualities on original language pairs are not degraded. Experimental results show that our method can learn new knowledge from diverse language pairs incrementally meanwhile maintaining performance on original language pairs, outperforming various strong baselines in incremental learning for MNMT.

Discourse-Centric Evaluation of Document-level Machine Translation with a New Densely Annotated Parallel Corpus of Novels
Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan and Ryan Cotterell   14:45-15:00 (Metropolitan West)
Several recent papers claim to have achieved human parity at sentence-level machine translation (MT)—especially between high-resource language pairs. In response, the MT community has, in part, shifted its focus to document-level translation. Translating documents requires a deeper understanding of the structure and meaning of text, which is often captured by various kinds of discourse phenomena such as consistency, coherence, and cohesion. However, this renders conventional sentence-level MT evaluation benchmarks inadequate for evaluating the performance of context-aware MT systems. This paper presents a new dataset with rich discourse annotations, built upon the large-scale parallel corpus BWB introduced in Jiang et al. (2022a). The new BWB annotation introduces four extra evaluation aspects, i.e., entity, terminology, coreference, and quotation, covering 15,095 entity mentions in both languages. Using these annotations, we systematically investigate the similarities and differences between the discourse structures of source and target languages, and the challenges they pose to MT. We discover that MT outputs differ fundamentally from human translations in terms of their latent discourse structures. This gives us a new perspective on the challenges and opportunities in document-level MT. We make our resource publicly available to spur future research in document-level MT and its generalization to other language translation tasks.

BLEURT Has Universal Translations: An Analysis of Automatic Metrics by Minimum Risk Training
Yiming Yan, Tao Wang, Chengqi Zhao, Shujian Huang, Jiajun CHEN and Mingxuan Wang   15:00-15:15 (Metropolitan West)
Automatic metrics play a crucial role in machine translation. Despite the widespread use of n-gram-based metrics, there has been a recent surge in the development of pre-trained model-based metrics that focus on measuring sentence semantics. However, these neural metrics, while achieving higher correlations with human evaluations, are often considered to be black boxes with potential biases that are difficult to detect. In this study, we systematically analyze and compare various mainstream and cutting-edge automatic metrics from the perspective of their guidance for training machine translation systems. Through Minimum Risk Training (MRT), we find that certain metrics exhibit robustness defects, such as the presence of universal adversarial translations in BLEURT and BARTScore. In-depth analysis suggests two main causes of these robustness deficits: distribution biases in the training datasets, and the tendency of the metric paradigm. By incorporating token-level constraints, we enhance the robustness of evaluation metrics, which in turn leads to an improvement in the performance of machine translation systems. Codes are available at https://github.com/powerpuffpomelo/fairseq_mrt.

xSIM++: An Improved Proxy to Bitext Mining Performance for Low-Resource Languages
Mingda Chen, Kevin Heffernan, Onur Çelebi, Alexandre Mourachko and Holger Schwenk   15:15-15:30 (Metropolitan West)
We introduce a new proxy score for evaluating bitext mining based on similarity in a multilingual embedding space: xsim++. In comparison to xsim, this improved proxy leverages rule-based approaches to extend English sentences in an evaluation set with synthetic, hard-to-distinguish examples which more closely mirror the scenarios we encounter during large-scale mining. We validate this proxy by running a significant number of bitext mining experiments for a set of low-resource languages, and subsequently train NMT systems on the mined data. In comparison to xsim, we show that xsim++ is better correlated with the downstream BLEU scores of translation systems trained on mined bitexts, providing a reliable proxy of bitext mining performance without needing to run expensive bitext mining pipelines. xsim++ also reports performance for different error types, offering more fine-grained feedbacks for model development.

Posters

14:00-15:30 (Frontenac Ballroom and Queen's Quay)

[TACL] Aggretriever: A Simple Approach to Aggregate Textual Representations for Robust Dense Passage Retrieval
Sheng-Chieh Lin, Minghan Li and Jimmy Lin   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Pre-trained language models have been successful in many knowledge-intensive NLP tasks. However, recent work has shown that models such as BERT are not "structurally ready" to aggregate textual information into a [CLS] vector for dense passage retrieval (DPR). This "lack of readiness" results from the gap between language model pre-training and DPR fine-tuning. Previous solutions call for computationally expensive techniques such as hard negative mining, cross-encoder distillation, and further pre-training to learn a robust DPR model. In this work, we instead propose to fully exploit knowledge in a pre-trained language model for DPR by aggregating the contextualized token embeddings into a dense vector, which we call agg*. By concatenating vectors from the [CLS] token and agg*, our Aggretriever model substantially improves the effectiveness of dense retrieval models on both in-domain and zero-shot evaluations without introducing substantial training overhead. Code is available at https://github.com/castorini/dhr

[TACL] Efficient Methods for Natural Language Processing: A Survey
Marcos Treviso, Ji-Ung Lee, Tianchu Ji, Betty van Aken, Qingqing Cao, Manuel Ciosici, Michael Hassid, Kenneth Heafield, Sara Hooker, Pedro Martins, André Martins, Peter Milder, Colin Raffel, Jessica Forde, Emma Strubell, Edwin Simpson, Noam Slonim, Jesse Dodge, Iryna Gurevych, Niranjan Balasubramanian, Leon Derczynski and Roy Schwartz   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Recent work in natural language processing (NLP) has yielded appealing results from scaling; however, using only scale to improve performance means that resource consumption also scales. Resources include data, time, storage, or energy, all of which are naturally limited and unevenly distributed. This motivates research into efficient methods that require fewer resources to achieve similar results. This survey synthesizes and relates current methods and findings in efficient NLP. We aim to provide both guidance for conducting NLP under limited resources, and point towards promising research directions for developing more efficient methods.

[Demo] The D-WISE Tool Suite: Multi-Modal Machine-Learning-Powered Tools Supporting and Enhancing Digital Discourse Analysis
Chris Biemann, Gertraud Koch, Isabel Eiser, Fynn Petersen-Frey, Tim Fischer and Florian Schneider 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
This work introduces the D-WISE Tool Suite (DWTS), a novel working environment for digital qualitative discourse analysis in the Digital Humanities (DH). The DWTS addresses limitations of current DH tools induced by the ever-increasing amount of heterogeneous, unstructured, and multi-modal data in which the discourses of contemporary societies are encoded. To provide meaningful insights from such data, our system leverages and combines state-of-the-art machine learning technologies from Natural Language Processing and Computer Vision. Further, the DWTS is conceived and developed by an interdisciplinary team of cultural anthropologists and computer scientists to ensure the tool's usability for modern DH research. Central features of the DWTS are: a) import of multi-modal data like text, image, audio, and video b) preprocessing pipelines for automatic annotations c) lexical and semantic search of documents d) manual span, bounding box, time-span, and frame annotations e) documentation of the research process.

[Demo] Zshot: An Open-source Framework for Zero-Shot Named Entity Recognition and Relation Extraction
Thanh Lam Hoang, Vanessa Lopez, Leopold Fuchs, Alberto Purpura, Marcos Martinez Galindo and Gabriele Picco 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
The Zero-Shot Learning (ZSL) task pertains to the identification of entities or relations in texts that were not seen during training. ZSL has emerged as a critical research area due to the scarcity of labeled data in specific domains, and its applications have grown significantly in recent years. With the advent of large pretrained language models, several novel methods have been proposed, resulting in substantial improvements in ZSL performance. There is a growing demand, both in the research community and industry, for a comprehensive ZSL framework that facilitates the development and accessibility of the latest methods and pretrained models. In this study, we propose a novel ZSL framework called Zshot that aims to address the aforementioned challenges. Our primary objective is to provide a platform that allows researchers to compare different state-of-the-art ZSL methods with standard benchmark datasets. Additionally, we have designed our framework to support the industry with readily available APIs for production under the standard SpaCy NLP pipeline. Our API is extendible and evaluable, moreover, we include numerous enhancements such as boosting the accuracy with pipeline ensembling and visualization utilities available as a SpaCy extension.

[Demo] TabGenie: A Toolkit for Table-to-Text Generation
Ondrej Dusek, Ondrej Platek, Ekaterina Garanina and Zdeněk Kasner 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Heterogenity of data-to-text generation datasets limits the research on data-to-text generation systems. We present TabGenie – a toolkit which enables researchers to explore, preprocess, and analyze a variety of data-to-text generation datasets through the unified framework of table-to-text generation. In TabGenie, all inputs are represented as tables with associated metadata. The tables can be explored through a web interface, which also provides an interactive mode for debugging table-to-text generation, facilitates side-by-side comparison of generated system outputs, and allows easy exports for manual analysis. Furthermore, TabGenie is equipped with command line processing tools and Python bindings for unified dataset loading and processing. We release TabGenie as a PyPI package and provide its open-source code and a live demo at https://github.com/kasnerz/tabgenie.

[Demo] An Efficient Conversational Smart Compose System
Ning Ruan, Jindong Chen, Maria Wang, Lijuan Liu, Xinying Song, Bowen Tan, Lei Shu, Xiayu Chen and Yun Zhu 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Online conversation is a ubiquitous way to share information and connect everyone but repetitive idiomatic text typing takes users a lot of time. This paper demonstrates a simple yet effective cloud based smart compose system to improve human-to-human conversation efficiency. Heuristics from different perspectives are designed to achieve the best trade-off between quality and latency. From the modeling side, the decoder-only model exploited the previous turns of conversational history in a computation lightweight manner. Besides, a novel phrase tokenizer is proposed to reduce latency without losing the composing quality further. Additionally, the caching mechanism is applied to the serving framework. The demo video of the system is available at https://youtu.be/U1KXkaqr60g. We open-sourced our phrase tokenizer in https://github.com/tensorflow/text.

[Demo] KWJA: A Unified Japanese Analyzer Based on Foundation Models
Sadao Kurohashi, Daisuke Kawahara, Yugo Murawaki, Hirokazu Kiyomaru, Takashi Kodama, Kazumasa Omura and Nobuhiro Ueda 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
We present KWJA, a high-performance unified Japanese text analyzer based on foundation models. KWJA supports a wide range of tasks, including typo correction, word segmentation, word normalization, morphological analysis, named entity recognition, linguistic feature tagging, dependency parsing, PAS analysis, bridging reference resolution, coreference resolution, and discourse relation analysis, making it the most versatile among existing Japanese text analyzers. KWJA solves these tasks in a multi-task manner but still achieves competitive or better performance compared to existing analyzers specialized for each task. KWJA is publicly available under the MIT license at https://github.com/ku-nlp/kwja.

[Demo] DeepPavlov Dream: Platform for Building Generative AI Assistants
Mikhail Burtsev, Dmitry Kosenko, Yana Shishkina, Dmitry Karpov, Veronika Smilga, Ksenya Petukhova, Dmitry Evseev, Maxim Talimanchuk, Fedor Ignatov, Daniel Kornev and Diliara Zharikova 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
An open-source DeepPavlov Dream Platform is specifically tailored for development of complex dialog systems like Generative AI Assistants. The stack prioritizes efficiency, modularity, scalability, and extensibility with the goal to make it easier to develop complex dialog systems from scratch. It supports modular approach to implementation of conversational agents enabling their development through the choice of NLP components and conversational skills from a rich library organized into the distributions of ready-for-use multi-skill AI assistant systems. In DeepPavlov Dream, multi-skill Generative AI Assistant consists of NLP components that extract features from user utterances, conversational skills that generate or retrieve a response, skill and response selectors that facilitate choice of relevant skills and the best response, as well as a conversational orchestrator that enables creation of multi-skill Generative AI Assistants scalable up to industrial grade AI assistants. The platform allows to integrate large language models into dialog pipeline, customize with prompt engineering, handle multiple prompts during the same dialog session and create simple multimodal assistants.

A Survey of Deep Learning for Mathematical Reasoning
Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck and Kai-Wei Chang 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Mathematical reasoning is a fundamental aspect of human intelligence and is applicable in various fields, including science, engineering, finance, and everyday life. The development of artificial intelligence (AI) systems capable of solving math problems and proving theorems in language has garnered significant interest in the fields of machine learning and natural language processing. For example, mathematics serves as a testbed for aspects of reasoning that are challenging for powerful deep learning models, driving new algorithmic and modeling advances. On the other hand, recent advances in large-scale neural language models have opened up new benchmarks and opportunities to use deep

learning for mathematical reasoning. In this survey paper, we review the key tasks, datasets, and methods at the intersection of mathematical reasoning and deep learning over the past decade. We also evaluate existing benchmarks and methods, and discuss future research directions in this domain.

IDRISI-RA: The First Arabic Location Mention Recognition Dataset of Disaster Tweets
Reem Suwaileh, Muhammad Imran and Tamer Elsayed   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Extracting geolocation information from social media data enables effective disaster management, as it helps response authorities; for example, in locating incidents for planning rescue activities, and affected people for evacuation. Nevertheless, geolocation extraction is greatly understudied for the low resource languages such as Arabic. To fill this gap, we introduce IDRISI-RA, the first publicly-available Arabic Location Mention Recognition (LMR) dataset that provides human- and automatically-labeled versions in order of thousands and millions of tweets, respectively. It contains both location mentions and their types (e.g., district, city). Our extensive analysis shows the decent geographical, domain, location granularity, temporal, and dialectical coverage of IDRISI-RA. Furthermore, we establish baselines using the standard Arabic NER models and build two simple, yet effective, LMR models. Our rigorous experiments confirm the need for developing specific models for Arabic LMR in the disaster domain. Moreover, experiments show the promising domain and geographical generalizability of IDRISI-RA under zero-shot learning.

Exploring Large Language Models for Classical Philology
Frederick Riemenschneider and Anette Frank   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Recent advances in NLP have led to the creation of powerful language models for many languages including Ancient Greek and Latin. While prior work on Classical languages unanimously uses BERT, in this work we create four language models for Ancient Greek that vary along two dimensions to study their versatility for tasks of interest for Classical languages: we explore (i) encoder-only and encoder-decoder architectures using RoBERTa and T5 as strong model types, and create for each of them (ii) a monolingual Ancient Greek and a multilingual instance that includes Latin and English. We evaluate all models on morphological and syntactic tasks, including lemmatization, which demonstrates the added value of T5's decoding abilities. We further define two probing tasks to investigate the knowledge acquired by models pre-trained on Classical texts. Our experiments provide the first benchmarking analysis of existing models of Ancient Greek. Results show that our models provide significant improvements over the SoTA. The systematic analysis of model types can inform future research in designing language models for Classical languages, including the development of novel generative tasks. We make all our models available as community resources, along with a large curated pre-training corpus for Ancient Greek, to support the creation of a larger, comparable model zoo for Classical Philology.

The KITMUS Test: Evaluating Knowledge Integration from Multiple Sources
Akshatha Arodi, Martin Pömsl, Kaheer Suleman, Adam Trischler, Alexandra Olteanu and Jackie Chi Kit Cheung   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Many state-of-the-art natural language understanding (NLU) models are based on pretrained neural language models. These models often make inferences using information from multiple sources. An important class of such inferences are those that require both background knowledge, presumably contained in a model's pretrained parameters, and instance-specific information that is supplied at inference time. However, the integration and reasoning abilities of NLU models in the presence of multiple knowledge sources have been largely understudied. In this work, we propose a test suite of coreference resolution subtasks that require reasoning over multiple facts. These subtasks differ in terms of which knowledge sources contain the relevant facts. We also introduce subtasks where knowledge is present only at inference time using fictional knowledge. We evaluate state-of-the-art coreference resolution models on our dataset. Our results indicate that several models struggle to reason on-the-fly over knowledge observed both at pretrain time and at inference time. However, with task-specific training, a subset of models demonstrates the ability to integrate certain knowledge types from multiple sources. Still, even the best performing models seem to have difficulties with reliably integrating knowledge presented only at inference time.

Revisiting non-English Text Simplification: A Unified Multilingual Benchmark
Michael Joseph Ryan, Tarek Naous and Wei Xu   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Recent advancements in high-quality, large-scale English resources have pushed the frontier of English Automatic Text Simplification (ATS) research. However, less work has been done on multilingual text simplification due to the lack of a diverse evaluation benchmark that covers complex-simple sentence pairs in many languages. This paper introduces the MultiSim benchmark, a collection of 27 resources in 12 distinct languages containing over 1.7 million complex-simple sentence pairs. This benchmark will encourage research in developing more effective multilingual text simplification models and evaluation metrics. Our experiments using MultiSim with pre-trained multilingual language models reveal exciting performance improvements from multilingual training in non-English settings. We observe strong performance from Russian in zero-shot cross-lingual transfer to low-resource languages. We further show that few-shot prompting with BLOOM-176b achieves comparable quality to reference simplifications outperforming fine-tuned models in most languages. We validate these findings through human evaluation.

A Needle in a Haystack: An Analysis of High-Agreement Workers on MTurk for Summarization
Lining Zhang, Simon Mille, Yufang Hou, Daniel Deutsch, Elizabeth Clark, Yixin Liu, Saad Mahamood, Sebastian Gehrmann, Miruna Adriana Clinciu, Khyathi Raghavi Chandu and João Sedoc   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
To prevent the costly and inefficient use of resources on low-quality annotations, we want a method for creating a pool of dependable annotators who can effectively complete difficult tasks, such as evaluating automatic summarization. Thus, we investigate the recruitment of high-quality Amazon Mechanical Turk workers via a two-step pipeline. We show that we can successfully filter out subpar workers before they carry out the evaluations and obtain high-agreement annotations with similar constraints on resources. Although our workers demonstrate a strong consensus among themselves and CloudResearch workers, their alignment with expert judgments on a subset of the data is not as expected and needs further training in correctness. This paper still serves as a best practice for the recruitment of qualified annotators in other challenging annotation tasks.

DEplain: A German Parallel Corpus with Intralingual Translations into Plain Language for Sentence and Document Simplification
Regina Stodden, Omar Momen and Laura Kallmeyer   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Text simplification is an intralingual translation task in which documents, or sentences of a complex source text are simplified for a target audience. The success of automatic text simplification systems is highly dependent on the quality of parallel data used for training and evaluation. To advance sentence simplification and document simplification in German, this paper presents DEplain, a new dataset of parallel, professionally written and manually aligned simplifications in plain German "plain DE" or German: "Einfache Sprache". DEplain consists of a news-domain (approx. 500 document pairs, approx. 13k sentence pairs) and a web-domain corpus (approx. 150 aligned documents, approx. 2k aligned sentence pairs). In addition, we are building a web harvester and experimenting with automatic alignment methods to facilitate the integration of non-aligned and to be-published parallel documents. Using this approach, we are dynamically increasing the web-domain corpus, so it is currently extended to approx. 750 document pairs and approx. 3.5k aligned sentence pairs. We show that using DEplain to train a transformer-based seq2seq text simplification model can achieve promising results. We make available the corpus, the

adapted alignment methods for German, the web harvester and the trained models here: https://github.com/rstodden/DEPlain.

Naamapadam: A Large-Scale Named Entity Annotated Data for Indic Languages

Arnav Anil Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy and Anoop Kunchukuttan
14:00-15:30 (Frontenac Ballroom and Queen's Quay)
We present, *Naamapadam*, the largest publicly available Named Entity Recognition (NER) dataset for the 11 major Indian languages from two language families. The dataset contains more than 400k sentences annotated with a total of at least 100k entities from three standard entity categories (Person, Location, and, Organization) for 9 out of the 11 languages. The training dataset has been automatically created from the Samanantar parallel corpus by projecting automatically tagged entities from an English sentence to the corresponding Indian language translation. We also create manually annotated testsets for 9 languages. We demonstrate the utility of the obtained dataset on the Naamapadam-test dataset. We also release *IndicNER*, a multilingual IndicBERT model fine-tuned on Naamapadam training set. IndicNER achieves an F1 score of more than 80 for 7 out of 9 test languages. The dataset and models are available under open-source licences at https://ai4bharat.iitm.ac.in/naamapadam.

MDACE: MIMIC Documents Annotated with Code Evidence

*Hua Jafari, Rana Jafari, April D Russell, Russell Klopfer, Edmond Lu, Benjamin R Striner and Matthew R. Gormley*14:00-15:30 (Frontenac Ballroom and Queen's Quay)
We introduce a dataset for evidence/rationale extraction on an extreme multi-label classification task over long medical documents. One such task is Computer-Assisted Coding (CAC) which has improved significantly in recent years, thanks to advances in machine learning technologies. Yet simply predicting a set of final codes for a patient encounter is insufficient as CAC systems are required to provide supporting textual evidence to justify the billing codes. A model able to produce accurate and reliable supporting evidence for each code would be a tremendous benefit. However, a human annotated code evidence corpus is extremely difficult to create because it requires specialized knowledge. In this paper, we introduce MDACE, the first publicly available code evidence dataset, which is built on a subset of the MIMIC-III clinical records. The dataset – annotated by professional medical coders – consists of 302 Inpatient charts with 3,934 evidence spans and 52 Profee charts with 5,563 evidence spans. We implemented several evidence extraction methods based on the EffectiveCAN model (Liu et al., 2021) to establish baseline performance on this dataset. MDACE can be used to evaluate code evidence extraction methods for CAC systems, as well as the accuracy and interpretability of deep learning models for multi-label classification. We believe that the release of MDACE will greatly improve the understanding and application of deep learning technologies for medical coding and document classification.

Zero-shot Faithful Factual Error Correction

Kung-Hsiang Huang, Hou Pong Chan and Heng Ji   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Faithfully correcting factual errors is critical for maintaining the integrity of textual knowledge bases and preventing hallucinations in sequence-to-sequence models. Drawing on humans' ability to identify and correct factual errors, we present a zero-shot framework that formulates questions about input claims, looks for correct answers in the given evidence, and assesses the faithfulness of each correction based on its consistency with the evidence. Our zero-shot framework outperforms fully-supervised approaches, as demonstrated by experiments on the FEVER and SciFact datasets, where our outputs are shown to be more faithful. More importantly, the decomposability nature of our framework inherently provides interpretability. Additionally, to reveal the most suitable metrics for evaluating factual error corrections, we analyze the correlation between commonly used metrics with human judgments in terms of three different dimensions regarding intelligibility and faithfulness.

Improving Automatic Quotation Attribution in Literary Novels

Krishnapriya Vishnubhotla, Frank Rudzicz, Graeme Hirst and Adam Hammond   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Current models for quotation attribution in literary novels assume varying levels of available information in their training and test data, which poses a challenge for in-the-wild inference. Here, we approach quotation attribution as a set of four interconnected sub-tasks: character identification, coreference resolution, quotation identification, and speaker attribution. We benchmark state-of-the-art models on each of these sub-tasks independently, using a large dataset of annotated coreferences and quotations in literary novels (the Project Dialogism Novel Corpus). We also train and evaluate models for the speaker attribution task in particular, showing that a simple sequential prediction model achieves accuracy scores on par with state-of-the-art models.

Improving Domain Generalization for Prompt-Aware Essay Scoring via Disentangled Representation Learning

Zhiwei Jiang, Tianyi Gao, Yafeng Yin, Meng Liu, Hua Yu, Zifeng Cheng and Qing Gu   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Automated Essay Scoring (AES) aims to score essays written in response to specific prompts. Many AES models have been proposed, but most of them are either prompt-specific or prompt-adaptive and cannot generalize well on "unseen" prompts. This work focuses on improving the generalization ability of AES models from the perspective of domain generalization, where the data of target prompts cannot be accessed during training. Specifically, we propose a prompt-aware neural AES model to extract comprehensive representation for essay scoring, including both prompt-invariant and prompt-specific features. To improve the generalization of representation, we further propose a novel disentangled representation learning framework. In this framework, a contrastive norm-angular alignment strategy and a counterfactual self-training strategy are designed to disentangle the prompt-invariant information and prompt-specific information in representation. Extensive experimental results on datasets of both ASAP and TOEFL11 demonstrate the effectiveness of our method under the domain generalization setting.

LeXFiles and LegalLAMA: Facilitating English Multinational Legal Language Model Development

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz and Anders Søgaard 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
In this work, we conduct a detailed analysis on the performance of legal-oriented pre-trained language models (PLMs). We examine the interplay between their original objective, acquired knowledge, and legal language understanding capacities which we define as the upstream, probing, and downstream performance, respectively. We consider not only the models' size but also the pre-training corpora used as important dimensions in our study. To this end, we release a multinational English legal corpus (LeXFiles) and a legal knowledge probing benchmark (LegalLAMA) to facilitate training and detailed analysis of legal-oriented PLMs. We release two new legal PLMs trained on LeXFiles and evaluate them alongside others on LegalLAMA and LexGLUE. We find that probing performance strongly correlates with upstream performance in related legal topics. On the other hand, downstream performance is mainly driven by the model's size and prior legal knowledge which can be estimated by upstream and probing performance. Based on these findings, we can conclude that both dimensions are important for those seeking the development of domain-specific PLMs.

Interpretable Math Word Problem Solution Generation via Step-by-step Planning

mengxue zhang, Zichao Wang, Zhichao Yang, weiqi feng and Andrew Lan   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Solutions to math word problems (MWPs) with step-by-step explanations are valuable, especially in education, to help students better comprehend problem-solving strategies. Most existing approaches only focus on obtaining the final correct answer. A few recent approaches leverage intermediate solution steps to improve final answer correctness but often cannot generate coherent steps with a clear solution strategy. Contrary to existing work, we focus on improving the correctness and coherence of the intermediate solutions steps. We propose a step-by-step

planning approach for intermediate solution generation, which strategically plans the generation of the next solution step based on the MWP and the previous solution steps. Our approach first plans the next step by predicting the necessary math operation needed to proceed, given history steps, then generates the next step, token-by-token, by prompting a language model with the predicted math operation. Experiments on the GSM8K dataset demonstrate that our approach improves the accuracy and interpretability of the solution on both automatic metrics and human evaluation.

Vision Meets Definitions: Unsupervised Visual Word Sense Disambiguation Incorporating Gloss Information
Sunjae Kwon, Rishabh Garodia, Minhwa Lee, Zhichao Yang and hong yu   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Visual Word Sense Disambiguation (VWSD) is a task to find the image that most accurately depicts the correct sense of the target word for the given context. Previously, image-text matching models often suffered from recognizing polysemous words. This paper introduces an unsupervised VWSD approach that uses gloss information of an external lexical knowledge-base, especially the sense definitions. Specifically, we suggest employing Bayesian inference to incorporate the sense definitions when sense information of the answer is not provided. In addition, to ameliorate the out-of-dictionary (OOD) issue, we propose a context-aware definition generation with GPT-3. Experimental results show that the VWSD performance significantly increased with our Bayesian inference-based approach. In addition, our context-aware definition generation achieved prominent performance improvement in OOD examples exhibiting better performance than the existing definition generation method.

Logic-driven Indirect Supervision: An Application to Crisis Counseling
Mattia Medina Grespan, Meghan Broadbent, Xinyao Zhang, Katherine E Axford, Brent Kious, Zac Imel and Vivek Srikumar   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Ensuring the effectiveness of text-based crisis counseling requires observing ongoing conversations and providing feedback, both labor-intensive tasks. Automatic analysis of conversations—at the full chat and utterance levels—may help support counselors and provide better care. While some session-level training data (e.g., rating of patient risk) is often available from counselors, labeling utterances requires expensive post hoc annotation. But the latter can not only provide insights about conversation dynamics, but can also serve to support quality assurance efforts for counselors. In this paper, we examine if inexpensive—and potentially noisy—session-level annotation can help improve label utterances. To this end, we propose a logic-based indirect supervision approach that exploits declaratively stated structural dependencies between both levels of annotation to improve utterance modeling. We show that adding these rules gives an improvement of 3.5% f-score over a strong multi-task baseline for utterance-level predictions. We demonstrate via ablation studies how indirect supervision via logic rules also improves the consistency and robustness of the system.

BIC: Twitter Bot Detection with Text-Graph Interaction and Semantic Consistency
*Zhenyu Lei, Herun Wan, Wenqian Zhang, Shangbin Feng, Zilong Chen, Jundong Li, Qinghua Zheng and Minnan Luo*14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Twitter bots are automatic programs operated by malicious actors to manipulate public opinion and spread misinformation. Research efforts have been made to automatically identify bots based on texts and networks on social media. Existing methods only leverage texts or networks alone, and while few works explored the shallow combination of the two modalities, we hypothesize that the interaction and information exchange between texts and graphs could be crucial for holistically evaluating bot activities on social media. In addition, according to a recent survey (Cresci, 2020), Twitter bots are constantly evolving while advanced bots steal genuine users' tweets and dilute their malicious content to evade detection. This results in greater inconsistency across the timeline of novel Twitter bots, which warrants more attention. In light of these challenges, we propose BIC, a Twitter Bot detection framework with text-graph Interaction and semantic Consistency. Specifically, in addition to separately modeling the two modalities on social media, BIC employs a text-graph interaction module to enable information exchange across modalities in the learning process. In addition, given the stealing behavior of novel Twitter bots, BIC proposes to model semantic consistency in tweets based on attention weights while using it to augment the decision process. Extensive experiments demonstrate that BIC consistently outperforms state-of-the-art baselines on two widely adopted datasets. Further analyses reveal that text-graph interactions and modeling semantic consistency are essential improvements and help combat bot evolution.

PVGRU: Generating Diverse and Relevant Dialogue Responses via Pseudo-Variational Mechanism
Yongkang Liu, Shi Feng, Daling Wang, Yifei Zhang and Hinrich Schütze   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
We investigate response generation for multi-turn dialogue in generative chatbots. Existing generative models based on RNNs (Recurrent Neural Networks) usually employ the last hidden state to summarize the history, which makes models unable to capture the subtle variability observed in different dialogues and cannot distinguish the differences between dialogues that are similar in composition. In this paper, we propose Pseudo-Variational Gated Recurrent Unit (PVGRU). The key novelty of PVGRU is a recurrent summarizing variable that aggregates the accumulated distribution variations of subsequences. We train PVGRU without relying on posterior knowledge, thus avoiding the training-inference inconsistency problem. PVGRU can perceive subtle semantic variability through summarizing variables that are optimized by two objectives we employ for training: distribution consistency and reconstruction. In addition, we build a Pseudo-Variational Hierarchical Dialogue (PVHD) model based on PVGRU. Experimental results demonstrate that PVGRU can broadly improve the diversity and relevance of responses on two benchmark datasets.

Don't Forget Your ABC's: Evaluating the State-of-the-Art in Chat-Oriented Dialogue Systems
Sarah E. Finch, James D. Finch and Jinho D. Choi   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Despite tremendous advancements in dialogue systems, stable evaluation still requires human judgments producing notoriously high-variance metrics due to their inherent subjectivity. Moreover, methods and labels in dialogue evaluation are not fully standardized, especially for open-domain chats, with a lack of work to compare and assess the validity of those approaches. The use of inconsistent evaluation can misinform the performance of a dialogue system, which becomes a major hurdle to enhance it. Thus, a dimensional evaluation of chat-oriented open-domain dialogue systems that reliably measures several aspects of dialogue capabilities is desired. This paper presents a novel human evaluation method to estimate the rates of many{pasted macro 'LN'} dialogue system behaviors. Our method is used to evaluate four state-of-the-art open-domain dialogue systems and compared with existing approaches. The analysis demonstrates that our behavior method is more suitable than alternative Likert-style or comparative approaches for dimensional evaluation of these systems.

Listener Model for the PhotoBook Referential Game with CLIPScores as Implicit Reference Chain
Shih-Lun Wu, Yi-Hui Chou and Liangze Li   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
PhotoBook is a collaborative dialogue game where two players receive private, partially-overlapping sets of images and resolve which images they have in common. It presents machines with a great challenge to learn how people build common ground around multimodal context to communicate effectively. Methods developed in the literature, however, cannot be deployed to real gameplay since they only tackle some subtasks of the game, and they require additional reference chains inputs, whose extraction process is imperfect. Therefore, we propose a reference chain-free listener model that directly addresses the game's predictive task, i.e., deciding whether an image is shared with partner. Our DeBERTa-based listener model reads the full dialogue, and utilizes CLIPScore features to assess utterance-image relevance. We achieve >77% accuracy on unseen sets of images/game themes, outperforming baseline by >17 points.

Task-Aware Specialization for Efficient and Robust Dense Retrieval for Open-Domain Question Answering
Hao Cheng, Hao Fang, Xiaodong Liu and Jianfeng Gao 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Given its effectiveness on knowledge-intensive natural language processing tasks, dense retrieval models have become increasingly popular. Specifically, the de-facto architecture for open-domain question answering uses two isomorphic encoders that are initialized from the same pretrained model but separately parameterized for questions and passages. This biencoder architecture is parameter-inefficient in that there is no parameter sharing between encoders. Further, recent studies show that such dense retrievers underperform BM25 in various settings. We thus propose a new architecture, Task-Aware Specialization for dEnse Retrieval (TASER), which enables parameter sharing by interleaving shared and specialized blocks in a single encoder. Our experiments on five question answering datasets show that TASER can achieve superior accuracy, surpassing BM25, while using about 60% of the parameters as bi-encoder dense retrievers. In out-of-domain evaluations, TASER is also empirically more robust than bi-encoder dense retrievers. Our code is available at https://github.com/microsoft/taser.

MultiTool-CoT: GPT-3 Can Use Multiple External Tools with Chain of Thought Prompting
Tatsuro Inaba, Hirokazu Kiyomaru, Fei Cheng and Sadao Kurohashi 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Large language models (LLMs) have achieved impressive performance on various reasoning tasks. To further improve the performance, we propose MultiTool-CoT, a novel framework that leverages chain-of-thought (CoT) prompting to incorporate multiple external tools, such as a calculator and a knowledge retriever, during the reasoning process. We apply MultiTool-CoT to the Task 2 dataset of NumGLUE, which requires both numerical reasoning and domain-specific knowledge. The experiments show that our method significantly outperforms strong baselines and achieves state-of-the-art performance.

Learning to Simulate Natural Language Feedback for Interactive Semantic Parsing
Hao Yan, Saurabh Srivastava, Yintao Tai, Sida I. Wang, Wen-tau Yih and Ziyu Yao 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Interactive semantic parsing based on natural language (NL) feedback, where users provide feedback to correct the parser mistakes, has emerged as a more practical scenario than the traditional one-shot semantic parsing. However, prior work has heavily relied on human-annotated feedback data to train the interactive semantic parser, which is prohibitively expensive and not scalable. In this work, we propose a new task of simulating NL feedback for interactive semantic parsing. We accompany the task with a novel feedback evaluator. The evaluator is specifically designed to assess the quality of the simulated feedback, based on which we decide the best feedback simulator from our proposed variants. On a text-to-SQL dataset, we show that our feedback simulator can generate high-quality NL feedback to boost the error correction ability of a specific parser. In low-data settings, our feedback simulator can help achieve comparable error correction performance as trained using the costly, full set of human annotations.

BUCA: A Binary Classification Approach to Unsupervised Commonsense Question Answering
Jie He, Simon Chi Lok U, Victor Gutierrez-Basulto and Jeff Z. Pan 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Unsupervised commonsense reasoning (UCR) is becoming increasingly popular as the construction of commonsense reasoning datasets is expensive, and they are inevitably limited in their scope. A popular approach to UCR is to fine-tune language models with external knowledge (e.g., knowledge graphs), but this usually requires a large number of training examples. In this paper, we propose to transform the downstream multiple choice question answering task into a simpler binary classification task by ranking all candidate answers according to their reasonableness. To this end, for training the model, we convert the knowledge graph triples into reasonable and unreasonable texts. Extensive experimental results show the effectiveness of our approach on various multiple choice question answering benchmarks. Furthermore, compared with existing UCR approaches using KGs, ours is less data hungry.

Answering Ambiguous Questions via Iterative Prompting
Weiwei Sun, Hengyi Cai, Hongshen Chen, Pengjie Ren, Zhumin CHEN, Maarten de Rijke and Zhaochun Ren 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
In open-domain question answering, due to the ambiguity of questions, multiple plausible answers may exist. To provide feasible answers to an ambiguous question, one approach is to directly predict all valid answers, but this can struggle with balancing relevance and diversity. An alternative is to gather candidate answers and aggregate them, but this method can be computationally costly and may neglect dependencies among answers. In this paper, we present AmbigPrompt to address the imperfections of existing approaches to answering ambiguous questions. Specifically, we integrate an answering model with a prompting model in an iterative manner. The prompting model adaptively tracks the reading process and progressively triggers the answering model to compose distinct and relevant answers. Additionally, we develop a task-specific post-pretraining approach for both the answering model and the prompting model, which greatly improves the performance of our framework. Empirical studies on two commonly-used open benchmarks show that AmbigPrompt achieves state-of-the-art or competitive results while using less memory and having a lower inference latency than competing approaches. Additionally, AmbigPrompt also performs well in low-resource settings.

DisentQA: Disentangling Parametric and Contextual Knowledge with Counterfactual Question Answering
Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor and Omri Abend 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Question answering models commonly have access to two sources of "knowledge" during inference time: (1) parametric knowledge - the factual knowledge encoded in the model weights, and (2) contextual knowledge - external knowledge (e.g., a Wikipedia passage) given to the model to generate a grounded answer. Having these two sources of knowledge entangled together is a core issue for generative QA models as it is unclear whether the answer stems from the given non-parametric knowledge or not. This unclarity has implications on issues of trust, interpretability and factuality. In this work, we propose a new paradigm in which QA models are trained to disentangle the two sources of knowledge. Using counterfactual data augmentation, we introduce a model that predicts two answers for a given question: one based on given contextual knowledge and one based on parametric knowledge. Our experiments on the Natural Questions dataset show that this approach improves the performance of QA models by making them more robust to knowledge conflicts between the two knowledge sources, while generating useful disentangled answers.

Using contradictions improves question answering systems
Etienne Fortier-Dubois and Domenic Anthony Rosati 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
This work examines the use of contradiction in natural language inference (NLI) for question answering (QA). Typically, NLI systems help answer questions by determining if a potential answer is entailed (supported) by some background context. But is it useful to also determine if an answer contradicts the context? We test this in two settings, multiple choice and extractive QA, and find that systems that incorporate contradiction can slightly do better than entailment-only systems on certain datasets. However, the best performances come from using contradiction, entailment, and QA model confidence scores together. This has implications for the deployment of QA systems in domains such as medicine and science where safety is an issue.

Elaboration-Generating Commonsense Question Answering at Scale
Wenya Wang, Vivek Srikumar, Hannaneh Hajishirzi and Noah A. Smith 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
In question answering requiring common sense, language models (e.g., GPT-3) have been used to generate text expressing background

knowledge that helps improve performance. Yet the cost of working with such models is very high; in this work, we finetune smaller language models to generate useful intermediate context, referred to here as elaborations. Our framework alternates between updating two language models—an elaboration generator and an answer predictor—allowing each to influence the other. Using less than 0.5% of the parameters of GPT-3, our model outperforms alternatives with similar sizes and closes the gap with GPT-3 on four commonsense question answering benchmarks. Human evaluations show that the quality of the generated elaborations is high.

MatCha: Enhancing Visual Language Pretraining with Math Reasoning and Chart Derendering
Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier and Julian Martin Eisenschlos　　14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Visual language data such as plots, charts, and infographics are ubiquitous in the human world. However, state-of-the-art vision-language models do not perform well on these data. We propose MatCha (Math reasoning and Chart derendering pretraining) to enhance visual language models' capabilities in jointly modeling charts/plots and language data. Specifically, we propose several pretraining tasks that cover plot deconstruction and numerical reasoning which are the key capabilities in visual language modeling. We perform the MatCha pretraining starting from Pix2Struct, a recently proposed image-to-text visual language model. On standard benchmarks such as PlotQA and ChartQA, the MatCha model outperforms state-of-the-art methods by as much as nearly 20%. We also examine how well MatCha pretraining transfers to domains such as screenshots, textbook diagrams, and document figures and observe overall improvement, verifying the usefulness of MatCha pretraining on broader visual language tasks.

Modeling Appropriate Language in Argumentation
Timon Ziegenbein, Shahbaz Syed, Felix Lange, Martin Potthast and Henning Wachsmuth　　14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Online discussion moderators must make ad-hoc decisions about whether the contributions of discussion participants are appropriate or should be removed to maintain civility. Existing research on offensive language and the resulting tools cover only one aspect among many involved in such decisions. The question of what is considered appropriate in a controversial discussion has not yet been systematically addressed. In this paper, we operationalize appropriate language in argumentation for the first time. In particular, we model appropriateness through the absence of flaws, grounded in research on argument quality assessment, especially in aspects from rhetoric. From these, we derive a new taxonomy of 14 dimensions that determine inappropriate language in online discussions. Building on three argument quality corpora, we then create a corpus of 2191 arguments annotated for the 14 dimensions. Empirical analyses support that the taxonomy covers the concept of appropriateness comprehensively, showing several plausible correlations with argument quality dimensions. Moreover, results of baseline approaches to assessing appropriateness suggest that all dimensions can be modeled computationally on the corpus.

Measuring the Effect of Influential Messages on Varying Personas
Chenkai Sun, Jinning Li, Hou Pong Chan, ChengXiang Zhai and Heng Ji　　14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Predicting how a user responds to news events enables important applications such as allowing intelligent agents or content producers to estimate the effect on different communities and revise unreleased messages to prevent unexpected bad outcomes such as social conflict and moral injury. We present a new task, Response Forecasting on Personas for News Media, to estimate the response a persona (characterizing an individual or a group) might have upon seeing a news message. Compared to the previous efforts which only predict generic comments to news, the proposed task not only introduces personalization in the modeling but also predicts the sentiment polarity and intensity of each response. This enables more accurate and comprehensive inference on the mental state of the persona. Meanwhile, the generated sentiment dimensions make the evaluation and application more reliable. We create the first benchmark dataset, which consists of 13,357 responses to 3,847 news headlines from Twitter. We further evaluate the SOTA neural language models with our dataset. The empirical results suggest that the included persona attributes are helpful for the performance of all response dimensions. Our analysis shows that the best-performing models are capable of predicting responses that are consistent with the personas, and as a byproduct, the task formulation also enables many interesting applications in the analysis of social network groups and their opinions, such as the discovery of extreme opinion groups.

A Weakly Supervised Classifier and Dataset of White Supremacist Language
Michael Miller Yoder, Ahmad Diab, David West Brown and Kathleen M Carley　　14:00-15:30 (Frontenac Ballroom and Queen's Quay)
We present a dataset and classifier for detecting the language of white supremacist extremism, a growing issue in online hate speech. Our weakly supervised classifier is trained on large datasets of text from explicitly white supremacist domains paired with neutral and anti-racist data from similar domains. We demonstrate that this approach improves generalization performance to new domains. Incorporating anti-racist texts as counterexamples to white supremacist language mitigates bias.

Multilingual Event Extraction from Historical Newspaper Adverts
Nadav Borenstein, Natália da Silva Perez and Isabelle Augenstein　　14:00-15:30 (Frontenac Ballroom and Queen's Quay)
NLP methods can aid historians in analyzing textual materials in greater volumes than manually feasible. Developing such methods poses substantial challenges though. First, acquiring large, annotated historical datasets is difficult, as only domain experts can reliably label them. Second, most available off-the-shelf NLP models are trained on modern language texts, rendering them significantly less effective when applied to historical corpora. This is particularly problematic for less well studied tasks, and for languages other than English. This paper addresses these challenges while focusing on the under-explored task of event extraction from a novel domain of historical texts. We introduce a new multilingual dataset in English, French, and Dutch composed of newspaper ads from the early modern colonial period reporting on enslaved people who liberated themselves from enslavement. We find that: 1) even with scarce annotated data, it is possible to achieve surprisingly good results by formulating the problem as an extractive QA task and leveraging existing datasets and models for modern languages; and 2) cross-lingual low-resource learning for historical languages is highly challenging, and machine translation of the historical datasets to the considered target languages is, in practice, often the best-performing solution.

What is the Real Intention behind this Question? Dataset Collection and Intention Classification
Maryam Sadat Mirzaei, Kourosh Meshgi and Satoshi Sekine　　14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Asking and answering questions are inseparable parts of human social life. The primary purposes of asking questions are to gain knowledge or request help which has been the subject of question-answering studies. However, questions can also reflect negative intentions and include implicit offenses, such as highlighting one's lack of knowledge or bolstering an alleged superior knowledge, which can lead to conflict in conversations; yet has been scarcely researched. This paper is the first study to introduce a dataset (Question Intention Dataset) that includes questions with positive/neutral and negative intentions and the underlying intention categories within each group. We further conduct a meta-analysis to highlight tacit and apparent intents. We also propose a classification method using Transformers augmented by TF-IDF-based features and report the results of several models for classifying the main intention categories. We aim to highlight the importance of taking intentions into account, especially implicit and negative ones, to gain insight into conflict-evoking questions and better understand human-human communication on the web for NLP applications.

CoID Fusion: Collaborative Descent for Distributed Multitask Finetuning
Shachar Don-Yehiya, Elad Venezian, Colin Raffel, Noam Slonim and Leshem Choshen　　14:00-15:30 (Frontenac Ballroom and Queen's Quay)

Pretraining has been shown to scale well with compute, data size and data diversity. Multitask learning trains on a mixture of supervised datasets and produces improved performance compared to self-supervised pretraining. Until now, massively multitask learning required simultaneous access to all datasets in the mixture and heavy compute resources that are only available to well-resourced teams.

In this paper, we propose ColD Fusion, a method that provides the benefits of multitask learning but leverages distributed computation and requires limited communication and no sharing of data. Consequentially, ColD Fusion can create a synergistic loop, where finetuned models can be recycled to continually improve the pretrained model they are based on. We show that ColD Fusion yields comparable benefits to multitask training by producing a model that (a) attains strong performance on all of the datasets it was multitask trained on and (b) is a better starting point for finetuning on unseen datasets. We find ColD Fusion outperforms RoBERTa and even previous multitask models. Specifically, when training and testing on 35 diverse datasets, ColD Fusion-based model outperforms RoBERTa by 2.19 points on average without any changes to the architecture.

CELDA: Leveraging Black-box Language Model as Enhanced Classifier without Labels
Hyunsoo Cho, Youna Kim and Sang-goo Lee   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Utilizing language models (LMs) without internal access is becoming an attractive paradigm in the field of NLP as many cutting-edge LMs are released through APIs and boast a massive scale. The de-facto method in this type of black-box scenario is known as prompting, which has shown progressive performance enhancements in situations where data labels are scarce or unavailable. Despite their efficacy, they still fall short in comparison to fully supervised counterparts and are generally brittle to slight modifications. In this paper, we propose Clustering-enhanced Linear Discriminative Analysis (CELDA), a novel approach that improves the text classification accuracy with a very weak-supervision signal (i.e., name of the labels). Our framework draws a precise decision boundary without accessing weights or gradients of the LM model or data labels. The core ideas of CELDA are twofold: (1) extracting a refined pseudo-labeled dataset from an unlabeled dataset, and (2) training a lightweight and robust model on the top of LM, which learns an accurate decision boundary from an extracted noisy dataset. Throughout in-depth investigations on various datasets, we demonstrated that CELDA reaches new state-of-the-art in weakly-supervised text classification and narrows the gap with a fully-supervised model. Additionally, our proposed methodology can be applied universally to any LM and has the potential to scale to larger models, making it a more viable option for utilizing large LMs.

Tailoring Instructions to Student's Learning Levels Boosts Knowledge Distillation
Yuxin Ren, Zihan Zhong, Xingjian Shi, Yi Zhu, Chun Yuan and Mu Li   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
It has been commonly observed that a teacher model with superior performance does not necessarily result in a stronger student, highlighting a discrepancy between current teacher training practices and effective knowledge transfer. In order to enhance the guidance of the teacher training process, we introduce the concept of distillation influence to determine the impact of distillation from each training sample on the student's generalization ability. In this paper, we propose Learning Good Teacher Matters (LGTM), an efficient training technique for incorporating distillation influence into the teacher's learning process. By prioritizing samples that are likely to enhance the student's generalization ability, our LGTM outperforms 10 common knowledge distillation baselines on 6 text classification tasks in the GLUE benchmark.

Backpack Language Models
John Hewitt, John Thickstun, Christopher D. Manning and Percy Liang   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
We present Backpacks: a new neural architecture that marries strong modeling performance with an interface for interpretability and control. Backpacks learn multiple non-contextual sense vectors for each word in a vocabulary, and represent a word in a sequence as a context-dependent, non-negative linear combination of sense vectors in this sequence. We find that, after training, sense vectors specialize, each encoding a different aspect of a word. We can interpret a sense vector by inspecting its (non-contextual, linear) projection onto the output space, and intervene on these interpretable hooks to change the model's behavior in predictable ways. We train a 170M-parameter Backpack language model on OpenWebText, matching the loss of a GPT-2 small (124Mparameter) Transformer. On lexical similarity evaluations, we find that Backpack sense vectors outperform even a 6B-parameter Transformer LM's word embeddings. Finally, we present simple algorithms that intervene on sense vectors to perform controllable text generation and debiasing. For example, we can edit the sense vocabulary to tend more towards a topic, or localize a source of gender bias to a sense vector and globally suppress that sense.

Targeted Data Generation: Finding and Fixing Model Weaknesses
Zexue He, Marco Tulio Ribeiro and Fereshte Khani   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Even when aggregate accuracy is high, state-of-the-art NLP models often fail systematically on specific subgroups of data, resulting in unfair outcomes and eroding user trust. Additional data collection may not help in addressing these weaknesses, as such challenging subgroups may be unknown to users, and underrepresented in the existing and new data. We propose Targeted Data Generation (TDG), a framework that automatically identifies challenging subgroups, and generates new data for those subgroups using large language models (LLMs) with a human in the loop. TDG estimates the expected benefit and potential harm of data augmentation for each subgroup, and selects the ones most likely to improve within-group performance without hurting overall performance. In our experiments, TDG significantly improves the accuracy on challenging subgroups for state-of-the-art sentiment analysis and natural language inference models, while also improving overall test accuracy.

Verify-and-Edit: A Knowledge-Enhanced Chain-of-Thought Framework
Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin and Lidong Bing   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
As large language models (LLMs) have become the norm in NLP, demonstrating good performance in generation and reasoning tasks, one of its most fatal disadvantages is the lack of factual correctness. Generating unfactual texts not only leads to lower performances but also degrades the trust and validity of their applications. Chain-of-Thought (CoT) prompting improves trust and model performance on complex reasoning tasks by generating interpretable reasoning chains, but still suffers from factuality concerns in knowledge-intensive tasks. In this paper, we propose the Verify-and-Edit framework for CoT prompting, which seeks to increase prediction factuality by post-editing reasoning chains according to external knowledge. Building on top of GPT-3, our framework lead to accuracy improvements in multiple open-domain question-answering tasks.

Latent Positional Information is in the Self-Attention Variance of Transformer Language Models Without Positional Embeddings
Ta-Chung Chi, Ting-Han Fan, Li-Wei Chen, alexander rudnicky and Peter J Ramadge 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
The use of positional embeddings in transformer language models is widely accepted. However, recent research has called into question the necessity of such embeddings. We further extend this inquiry by demonstrating that a randomly initialized and frozen transformer language model, devoid of positional embeddings, inherently encodes strong positional information through the shrinkage of self-attention variance. To quantify this variance, we derive the underlying distribution of each step within a transformer layer. Through empirical validation using a fully pretrained model, we show that the variance shrinkage effect still persists after extensive gradient updates. Our findings serve to justify the decision to discard positional embeddings and thus facilitate more efficient pretraining of transformer language models.

Few-shot In-context Learning on Knowledge Base Question Answering
Tianle Li, Xueguang Ma, Alex Zhuang, Yu Gu, Yu Su and Wenhu Chen   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Question answering over knowledge bases is considered a difficult problem due to the challenge of generalizing to a wide variety of possible

natural language questions. Additionally, the heterogeneity of knowledge base schema items between different knowledge bases often necessitates specialized training for different knowledge base question-answering (KBQA) datasets. To handle questions over diverse KBQA datasets with a unified training-free framework, we propose KB-BINDER, which for the first time enables few-shot in-context learning over KBQA tasks. Firstly, KB-BINDER leverages large language models like Codex to generate logical forms as the draft for a specific question by imitating a few demonstrations. Secondly, KB-BINDER grounds on the knowledge base to bind the generated draft to an executable one with BM25 score matching. The experimental results on four public heterogeneous KBQA datasets show that KB-BINDER can achieve a strong performance with only a few in-context demonstrations. Especially on GraphQA and 3-hop MetaQA, KB-BINDER can even outperform the state-of-the-art trained models. On GrailQA and WebQSP, our model is also on par with other fully-trained models. We believe KB-BINDER can serve as an important baseline for future research. We plan to release all the code and data. Our code is available at https://github.com/ltl3A87/KB-BINDER.

MixCE: Training Autoregressive Language Models by Mixing Forward and Reverse Cross-Entropies
Shiyue Zhang, Shijie Wu, Ozan Irsoy, Steven Lu, Mohit Bansal, Mark Dredze and David Rosenberg   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Autoregressive language models are trained by minimizing the cross-entropy of the model distribution Q relative to the data distribution P – that is, minimizing the forward cross-entropy, which is equivalent to maximum likelihood estimation (MLE). We have observed that models trained in this way may "over-generalize", in the sense that they produce non-human-like text. Moreover, we believe that reverse cross-entropy, i.e., the cross-entropy of P relative to Q, is a better reflection of how a human would evaluate text generated by a model. Hence, we propose learning with MixCE, an objective that mixes the forward and reverse cross-entropies. We evaluate models trained with this objective on synthetic data settings (where P is known) and real data, and show that the resulting models yield better generated text without complex decoding strategies.

Do Models Really Learn to Follow Instructions? An Empirical Study of Instruction Tuning
Po-Nien Kung and Nanyun Peng   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Recent works on instruction tuning (IT) have achieved great performance with zero-shot generalizability to unseen tasks. With additional context (e.g., task definition, examples) provided to models for fine-tuning, they achieved much higher performance than untuned models. Despite impressive performance gains, what models learn from IT remains understudied. In this work, we analyze how models utilize instructions during IT by comparing model training with altered vs. original instructions. Specifically, we create simplified task definitions by removing all semantic components and only leaving the output space information, and delusive examples that contain incorrect input-output mapping. Our experiments show that models trained on simplified task definition or delusive examples can achieve comparable performance to the ones trained on the original instructions and examples. Furthermore, we introduce a random baseline to perform zeroshot classification tasks, and find it achieves similar performance (42.6% exact-match) as IT does (43% exact-match) in low resource setting, while both methods outperform naive T5 significantly (30% per exact-match). Our analysis provides evidence that the impressive performance gain of current IT models can come from picking up superficial patterns, such as learning the output format and guessing. Our study highlights the urgent need for more reliable IT methods and evaluation.

Say What You Mean! Large Language Models Speak Too Positively about Negative Commonsense Knowledge
Jiangjie Chen, Wei Shi, Ziquan Fu, Sijie Cheng, Lei Li and Yanghua Xiao   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Large language models (LLMs) have been widely studied for their ability to store and utilize positive knowledge. Positive knowledge, such as "lions don't live in the ocean", is also ubiquitous in the world but rarely mentioned explicitly in text. What do LLMs know about negative knowledge? This work examines the ability of LLMs on negative commonsense knowledge. We design a constrained keywords-to-sentence generation task (CG) and a Boolean question answering task (QA) to probe LLMs. Our experiments reveal that LLMs frequently fail to generate valid sentences grounded in negative commonsense knowledge, yet they can correctly answer polar yes-or-no questions. We term this phenomenon the belief conflict of LLMs. Our further analysis shows that statistical shortcuts and negation reporting bias from language modeling pre-training cause this conflict.

Mitigating Label Biases for In-context Learning
Yu Fei, Yifan Hou, Zeming Chen and Antoine Bosselut   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Various design settings for in-context learning (ICL), such as the choice and order of the in-context examples, can bias the model's predictions. While many studies discuss these design choices, there have been few systematic investigations into categorizing them and mitigating their impact. In this work, we define a typology for three types of label biases in ICL for text classification: vanilla-label bias, context-label bias, and domain-label bias (which we conceptualize and detect for the first time).

Our analysis demonstrates that prior label bias calibration methods fall short of addressing all three types of biases. Specifically, domain-label bias restricts LLMs to random-level performance on many tasks regardless of the choice of in-context examples. To mitigate the effect of these biases, we propose a simple bias calibration method that estimates a language model's label bias using random in-domain words from the task corpus. After controlling for this estimated bias when making predictions, our novel domain-context calibration significantly improves the ICL performance of GPT-J and GPT-3 on a wide range of tasks. The gain is substantial on tasks with large domain-label bias (up to 37% in Macro-F1). Furthermore, our results generalize to models with different scales, pretraining methods, and manually-designed task instructions, showing the prevalence of label biases in ICL.

Open-Domain Hierarchical Event Schema Induction by Incremental Prompting and Verification
Sha Li, Ruining Zhao, Manling Li, Heng Ji, Chris Callison-Burch and Jiawei Han   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Event schemas are a form of world knowledge about the typical progression of events. Recent methods for event schema induction use information extraction systems to construct a large number of event graph instances from documents, and then learn to generalize the schema from such instances. In contrast, we propose to treat event schemas as a form of commonsense knowledge that can be derived from large language models (LLMs). This new paradigm greatly simplifies the schema induction process and allows us to handle both hierarchical relations and temporal relations between events in a straightforward way. Since event schemas have complex graph structures, we design an incremental prompting and verification method IncPrompt to break down the construction of a complex event graph into three stages: event skeleton construction, event expansion, and event-event relation verification. Compared to directly using LLMs to generate a linearized graph, Inc-Schema can generate large and complex schemas with 7.2% F1 improvement in temporal relations and 31.0% F1 improvement in hierarchical relations. In addition, compared to the previous state-of-the-art closed-domain schema induction model, human assessors were able to cover 10% more events when translating the schemas into coherent stories and rated our schemas 1.3 points higher (on a 5-point scale) in terms of readability.

RL4F: Generating Natural Language Feedback with Reinforcement Learning for Repairing Model Outputs
Afra Feyza Akyurek, Ekin Akyurek, Ashwin Kalyan, Peter Clark, Derry Tanti Wijaya and Niket Tandon   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Despite their unprecedented success, even the largest language models make mistakes. Similar to how humans learn and improve using feedback, previous work proposed providing language models with natural language feedback to guide them in repairing their outputs. Be-

cause human-generated critiques are expensive to obtain, researchers have devised learned critique generators in lieu of human critics while assuming one can train downstream models to utilize generated feedback. However, this approach does not apply to black-box or limited access models such as ChatGPT, as they cannot be fine-tuned. Moreover, in the era of large general-purpose language agents, fine-tuning is neither computationally nor spatially efficient as it results in multiple copies of the network. In this work, we introduce RL4F (Reinforcement Learning for Feedback), a multi-agent collaborative framework where the critique generator is trained to maximize end-task performance of GPT-3, a fixed model more than 200 times its size. RL4F produces critiques that help GPT-3 revise its outputs. We study three datasets for action planning, summarization and alphabetization and show relative improvements up to 10% in multiple text similarity metrics over other learned, retrieval-augmented or prompting-based critique generators.

Training Trajectories of Language Models Across Scales
Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer and Veselin Stoyanov
14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Scaling up language models has led to unprecedented performance gains, but little is understood about how the training dynamics change as models get larger. How do language models of different sizes learn during pre-training? Why do larger language models demonstrate more desirable behaviors? In this paper, we analyze the intermediate training checkpoints of differently sized OPT models (Zhang et al., 2022)—from 125M to 175B parameters—on next-token prediction, sequence-level generation and downstream tasks. We find that 1) at a given perplexity and independent of model sizes, a similar subset of training tokens see the most significant reduction in loss, with the rest stagnating or showing double-descent behavior (Nakkiran et al., 2020); 2) early in training, all models learn to reduce the perplexity of grammatical sequences that contain hallucinations, with small models halting at this suboptimal distribution and larger ones eventually learning to assign these sequences lower probabilities; and 3) perplexity is a strong predictor of in-context learning performance on 74 multiple-choice tasks from BIG-Bench, and this holds independent of the model size. Together, these results show that perplexity is more predictive of model behaviors than model size or training computation.

NarrowBERT: Accelerating Masked Language Model Pretraining and Inference
Haoxin Li, Phillip Keung, Daniel Cheng, Jungo Kasai and Noah A. Smith 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Large-scale language model pretraining is a very successful form of self-supervised learning in natural language processing, but it is increasingly expensive to perform as the models and pretraining corpora have become larger over time. We propose NarrowBERT, a modified transformer encoder that increases the throughput for masked language model pretraining by more than 2x. NarrowBERT sparsifies the transformer model such that the self-attention queries and feedforward layers only operate on the masked tokens of each sentence during pretraining, rather than all of the tokens as with the usual transformer encoder. We also show that NarrowBERT increases the throughput at inference time by as much as 3.5x with minimal (or no) performance degradation on sentence encoding tasks like MNLI. Finally, we examine the performance of NarrowBERT on the IMDB and Amazon reviews classification and CoNLL NER tasks and show that it is also comparable to standard BERT performance.

CATS: A Pragmatic Chinese Answer-to-Sequence Dataset with Large Scale and High Quality
Liang Li, Ruiying Geng, Chengyang Fang, Bing Li, Can Ma, Rongyu Cao, Binhua Li, Fei Huang and Yongbin Li 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
There are three problems existing in the popular data-to-text datasets. First, the large-scale datasets either contain noise or lack real application scenarios. Second, the datasets close to real applications are relatively small in size. Last, current datasets bias in the English language while leaving other languages underexplored. To alleviate these limitations, in this paper, we present CATS, a pragmatic Chinese answer-to-sequence dataset with large scale and high quality. The dataset aims to generate textual descriptions for the answer in the practical TableQA system. Further, to bridge the structural gap between the input SQL and table and establish better semantic alignments, we propose a Unified Graph Transformation approach to establish a joint encoding space for the two hybrid knowledge resources and convert this task to a graph-to-text problem. The experiment results demonstrate the effectiveness of our proposed method. Further analysis on CATS attests to both the high quality and challenges of the dataset

WeCheck: Strong Factual Consistency Checker via Weakly Supervised Learning
Wenhao Wu, Wei Li, Xinyan Xiao, Jiachen Liu, Sujian Li and Yajuan Lyu 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
A crucial issue of current text generation models is that they often uncontrollably generate text that is factually inconsistent with inputs. Due to lack of annotated data, existing factual consistency metrics usually train evaluation models on synthetic texts or directly transfer from other related tasks, such as question answering (QA) and natural language inference (NLI). Bias in synthetic text or upstream tasks makes them perform poorly on text actually generated by language models, especially for general evaluation for various tasks. To alleviate this problem, we propose a weakly supervised framework named **WeCheck** that is directly trained on actual generated samples from language models with weakly annotated labels. WeCheck first utilizes a generative model to infer the factual labels of generated samples by aggregating weak labels from multiple resources. Next, we train a simple noise-aware classification model as the target metric using the inferred weakly supervised information. Comprehensive experiments on various tasks demonstrate the strong performance of WeCheck, achieving an average absolute improvement of 3.3% on the TRUE benchmark over 11B state-of-the-art methods using only 435M parameters. Furthermore, it is up to 30 times faster than previous evaluation methods, greatly improving the accuracy and efficiency of factual consistency evaluation.

TwistList: Resources and Baselines for Tongue Twister Generation
Tyler Loakman, Chen Tang and Chenghua Lin 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Previous work in phonetically-grounded language generation has mainly focused on domains such as lyrics and poetry. In this paper, we present work on the generation of tongue twisters - a form of language that is required to be phonetically conditioned to maximise sound overlap, whilst maintaining semantic consistency with an input topic, and still being grammatically correct. We present TwistList, a large annotated dataset of tongue twisters, consisting of 2.1K+ human-authored examples. We additionally present several benchmark systems (referred to as TwisterMisters) for the proposed task of tongue twister generation, including models that both do and do not require training on in-domain data. We present the results of automatic and human evaluation to demonstrate the performance of existing mainstream pre-trained models in this task with limited (or no) task specific training and data, and no explicit phonetic knowledge. We find that the task of tongue twister generation is challenging for models under these conditions, yet some models are still capable of generating acceptable examples of this language type.

Annotating and Detecting Fine-grained Factual Errors for Dialogue Summarization
Rongxin Zhu, Jianzhong Qi and Jey Han Lau 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
A series of datasets and models have been proposed for summaries generated for well-formatted documents such as news articles. Dialogue summaries, however, have been under explored. In this paper, we present the first dataset with fine-grained factual error annotations named DIASUMFACT. We define fine-grained factual error detection as a sentence-level multi-label classification problem, and we evaluate two state-of-the-art (SOTA) models on our dataset. Both models yield sub-optimal results, with a macro-averaged F1 score of around 0.25 over 6 error classes. We further propose an unsupervised model ENDERANKER via candidate ranking using pretrained encoder-decoder models. Our model performs on par with the SOTA models while requiring fewer resources. These observations confirm the challenges in detecting

factual errors from dialogue summaries, which call for further studies, for which our dataset and results offer a solid foundation.

Multi-Document Summarization with Centroid-Based Pretraining
Ratish Surendran Pudupully, Parag Jain, Nancy Chen and Mark Steedman   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
In Multi-Document Summarization (MDS), the input can be modeled as a set of documents, and the output is its summary. In this paper, we focus on pretraining objectives for MDS. Specifically, we introduce a novel pretraining objective, which involves selecting the ROUGE-based centroid of each document cluster as a proxy for its summary. Our objective thus does not require human written summaries and can be utilized for pretraining on a dataset consisting solely of document sets. Through zero-shot, few-shot, and fully supervised experiments on multiple MDS datasets, we show that our model *Centrum* is better or comparable to a state-of-the-art model. We make the pretrained and fine-tuned models freely available to the research communityhttps://github.com/ratishsp/centrum.

Abstractive Summarizers are Excellent Extractive Summarizers
Daniel Varab and Yumo Xu   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Extractive and abstractive summarization designs have historically been fragmented, limiting the benefits that often arise from compatible model architectures. In this paper, we explore the potential synergies of modeling extractive summarization with an abstractive summarization system and propose three novel inference algorithms using the sequence-to-sequence architecture. We evaluate them on the CNN & Dailymail dataset and show that recent advancements in abstractive system designs enable abstractive systems to not only compete, but even surpass the performance of extractive systems with custom architectures. To our surprise, abstractive systems achieve this without being exposed to extractive oracle summaries and, therefore, for the first time allow a single model to produce both abstractive and extractive summaries. This evidence questions our fundamental understanding of extractive system design, and the necessity for extractive labels while pathing the way for promising research directions in hybrid models.

On Improving Summarization Factual Consistency from Natural Language Feedback
Yixin Liu, Budhaditya Deb, Milagro Teruel, Aaron L Halfaker, Dragomir Radev and Ahmed Hassan Awadallah   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Despite the recent progress in language generation models, their outputs may not always meet user expectations. In this work, we study whether informational feedback in natural language can be leveraged to improve generation quality and user preference alignment. To this end, we consider factual consistency in summarization, the quality that the summary should only contain information supported by the input documents, as the user-expected preference. We collect a high-quality dataset, DeFacto, containing human demonstrations and informational natural language feedback consisting of corrective instructions, edited summaries, and explanations with respect to the factual consistency of the summary. Using our dataset, we study three natural language generation tasks: (1) editing a summary by following the human feedback, (2) generating human feedback for editing the original summary, and (3) revising the initial summary to correct factual errors by generating both the human feedback and edited summary. We show that DeFacto can provide factually consistent human-edited summaries and further insights into summarization factual consistency thanks to its informational natural language feedback. We further demonstrate that fine-tuned language models can leverage our dataset to improve the summary factual consistency, while large language models lack the zero-shot learning ability in our proposed tasks that require controllable text generation.

Towards Understanding Omission in Dialogue Summarization
*Yicheng Zou, Kaitao Song, Xu Tan, Zhongkai Fu, Qi Zhang, Dongsheng Li and Tao Gui*14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Dialogue summarization aims to condense the lengthy dialogue into a concise summary, and has recently achieved significant progress. However, the result of existing methods is still far from satisfactory. Previous works indicated that omission is a major factor in affecting the quality of summarization, but few of them have further explored the omission problem, such as how omission affects summarization results and how to detect omission, which is critical for reducing omission and improving summarization quality. Moreover, analyzing and detecting omission relies on summarization datasets with omission labels (i.e., which dialogue utterances are omitted in the summarization), which are not available in the current literature. In this paper, we propose the OLDS dataset, which provides high-quality omission labels for dialogue summarization. By analyzing this dataset, we find that a large improvement in summarization quality can be achieved by providing ground-truth omission labels for the summarization model to recover omission information, which demonstrates the importance of omission detection for omission mitigation in dialogue summarization. Therefore, we formulate an omission detection task and demonstrate our proposed dataset can support the training and evaluation of this task well. We also call for research action on omission detection based on our proposed datasets. Our dataset and codes are publicly available.

Accelerating Transformer Inference for Translation via Parallel Decoding
Andrea Santilli, Silvio Severino, Emilian Postolache, Valentino Maiorca, Michele Mancusi, Riccardo Marin and Emanuele Rodola   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Autoregressive decoding limits the efficiency of transformers for Machine Translation (MT). The community proposed specific network architectures and learning-based methods to solve this issue, which are expensive and require changes to the MT model, trading inference speed at the cost of the translation quality. In this paper, we propose to address the problem from the point of view of decoding algorithms, as a less explored but rather compelling direction. We propose to reframe the standard greedy autoregressive decoding of MT with a parallel formulation leveraging Jacobi and Gauss-Seidel fixed-point iteration methods for fast inference. This formulation allows to speed up existing models without training or modifications while retaining translation quality. We present three parallel decoding algorithms and test them on different languages and models showing how the parallelization introduces a speedup up to 38% w.r.t. the standard autoregressive decoding and nearly 2x when scaling the method on parallel resources. Finally, we introduce a decoding dependency graph visualizer (DDGviz) that let us see how the model has learned the conditional dependence between tokens and inspect the decoding procedure.

The Inside Story: Towards Better Understanding of Machine Translation Neural Evaluation Metrics
Ricardo Rei, Nuno M. Guerreiro, Marcos Treviso, Luisa Coheur, Alon Lavie and André Martins   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Neural metrics for machine translation evaluation, such as COMET, exhibit significant improvements in their correlation with human judgments, as compared to traditional metrics based on lexical overlap, such as BLEU. Yet, neural metrics are, to a great extent, "black boxes" returning a single sentence-level score without transparency about the decision-making process. In this work, we develop and compare several neural explainability methods and demonstrate their effectiveness for interpreting state-of-the-art fine-tuned neural metrics. Our study reveals that these metrics leverage token-level information that can be directly attributed to translation errors, as assessed through comparison of token-level neural saliency maps with Multidimensional Quality Metrics (MQM) annotations and with synthetically-generated critical translation errors. To ease future research, we release our code at: https://github.com/Unbabel/COMET/tree/explainable-metrics

A Simple Concatenation can Effectively Improve Speech Translation
Linlin Zhang, Kai Fan, Boxing Chen and Luo Si   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
A triple speech translation data comprises speech, transcription, and translation. In the end-to-end paradigm, text machine translation (MT) usually plays the role of a teacher model for the speech translation (ST) via knowledge distillation. Parameter sharing with the teacher is

often adopted to construct the ST model architecture, however, the two modalities are independently fed and trained via different losses. This situation does not match ST's properties across two modalities and also limits the upper bound of the performance. Inspired by the works of video Transformer, we propose a simple unified cross-modal ST method, which concatenates speech and text as the input, and builds a teacher that can utilize both cross-modal information simultaneously. Experimental results show that in our unified ST framework, models can effectively utilize the auxiliary information from speech and text, and achieve compelling results on MuST-C datasets.

INK: Injecting kNN Knowledge in Nearest Neighbor Machine Translation
Wenhao Zhu, Jingjing Xu, Shujian Huang, Lingpeng Kong and Jiajun CHEN   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Neural machine translation has achieved promising results on many translation tasks. However, previous studies have shown that neural models induce a non-smooth representation space, which harms its generalization results. Recently, kNN-MT has provided an effective paradigm to smooth the prediction based on neighbor representations during inference. Despite promising results, kNN-MT usually requires large inference overhead. We propose an effective training framework INK to directly smooth the representation space via adjusting representations of kNN neighbors with a small number of new parameters. The new parameters are then used to refresh the whole representation datastore to get new kNN knowledge asynchronously. This loop keeps running until convergence. Experiments on four benchmark datasets show that INK achieves average gains of 1.99 COMET and 1.0 BLEU, outperforming the state-of-the-art kNN-MT system with 0.02x memory space and 1.9x inference speedup.

Local Byte Fusion for Neural Machine Translation
Makesh Narsimhan Sreedhar, Xiangpeng Wan, Yu Cheng and Junjie Hu   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Subword tokenization schemes are the dominant technique used in current NLP models. However, such schemes can be rigid and tokenizers built on one corpus may not adapt well to other parallel corpora. It has also been observed that in multilingual corpora, subword tokenization schemes oversegment low-resource languages, leading to a drop in translation performance. An alternative to subword tokenizers is byte-based tokenization, i.e., tokenization into byte sequences using the UTF-8 encoding scheme. Byte tokens often represent inputs at a sub-character granularity, i.e., one character can be represented by a span of byte tokens. This results in much longer byte sequences that are hard to interpret without aggregating local information from multiple byte tokens.

In this paper, we propose a Local Byte Fusion (LOBEF) method for byte-based machine translation—utilizing byte n-gram and word boundaries—to aggregate local semantic information. Extensive experiments on multilingual translation, zero-shot cross-lingual transfer, and domain adaptation reveal a consistent improvement over vanilla byte-based models. Further analysis also indicates that our byte-based models are parameter-efficient and perform competitive to subword models.

More than Classification: A Unified Framework for Event Temporal Relation Extraction
Quzhe Huang, Yutong Hu, Shengqi Zhu, Yansong Feng, Chang Liu and Dongyan Zhao 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Event temporal relation extraction (ETRE) is usually formulated as a multi-label classification task, where each type of relation is simply treated as a one-hot label. This formulation ignores the meaning of relations and wipes out their intrinsic dependency. After examining the relation definitions in various ETRE tasks, we observe that all relations can be interpreted using the start and end time points of events. For example, relation *Includes* could be interpreted as event 1 starting no later than event 2 and ending no earlier than event 2. In this paper, we propose a unified event temporal relation extraction framework, which transforms temporal relations into logical expressions of time points and completes the ETRE by predicting the relations between certain time point pairs. Experiments on TB-Dense and MATRES show significant improvements over a strong baseline and outperform the state-of-the-art model by 0.3% on both datasets. By representing all relations in a unified framework, we can leverage the relations with sufficient data to assist the learning of other relations, thus achieving stable improvement in low-data scenarios. When the relation definitions are changed, our method can quickly adapt to the new ones by simply modifying the logic expressions that map time points to new event relations. The code is released at https://github.com/AndrewZhe/A-Unified-Framework-for-ETRE

Discriminative Reasoning with Sparse Event Representation for Document-level Event-Event Relation Extraction
Changsen Yuan, Heyan Huang, Yixin Cao and Yonggang Wen   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Document-level Event Causality Identification (DECI) aims to extract causal relations between events in a document. It challenges conventional sentence-level task (SECI) with difficult long-text understanding. In this paper, we propose a novel DECI model (SENDIR) for better document-level reasoning. Different from existing works that build an event graph via linguistic tools, SENDIR does not require any prior knowledge. The basic idea is to discriminate event pairs in the same sentence or span multiple sentences by assuming their different information density: 1) low density in the document suggests sparse attention to skip irrelevant information. Our module 1 designs various types of attention for event representation learning to capture long-distance dependence. 2) High density in a sentence makes SECI relatively easy. Module 2 uses different weights to highlight the roles and contributions of intra- and inter-sentential reasoning, which introduces supportive event pairs for joint modeling. Extensive experiments demonstrate great improvements in SENDIR and the effectiveness of various sparse attention for document-level representations. Codes will be released later.

Rethinking Multimodal Entity and Relation Extraction from a Translation Point of View
Changmeng Zheng, Junhao Feng, Yi Cai, Xiaoyong Wei and Qing Li   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
We revisit the multimodal entity and relation extraction from a translation point of view. Special attention is paid on the misalignment issue in text-image datasets which may mislead the learning. We are motivated by the fact that the cross-modal misalignment is a similar problem of cross-lingual divergence issue in machine translation. The problem can then be transformed and existing solutions can be borrowed by treating a text and its paired image as the translation to each other. We implement a multimodal back-translation using diffusion-based generative models for pseudo-paralleled pairs and a divergence estimator by constructing a high-resource corpora as a bridge for low-resource learners. Fine-grained confidence scores are generated to indicate both types and degrees of alignments with which better representations are obtained. The method has been validated in the experiments by outperforming 14 state-of-the-art methods in both entity and relation extraction tasks. The source code is available at https://github.com/thecharm/TMR.

When to Use What: An In-Depth Comparative Empirical Analysis of OpenIE Systems for Downstream Applications
Kevin Song Pei, Ishan Jindal, Kevin Chen-Chuan Chang, ChengXiang Zhai and Yunyao Li   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Open Information Extraction (OpenIE) has been used in the pipelines of various NLP tasks. Unfortunately, there is no clear consensus on which models to use in which tasks. Muddying things further is the lack of comparisons that take differing training sets into account. In this paper, we present an application-focused empirical survey of neural OpenIE models, training sets, and benchmarks in an effort to help users choose the most suitable OpenIE systems for their applications. We find that the different assumptions made by different models and datasets have a statistically significant effect on performance, making it important to choose the most appropriate model for one's applications. We demonstrate the applicability of our recommendations on a downstream Complex QA application.

Do Androids Laugh at Electric Sheep? Humor "Understanding" Benchmarks from The New Yorker Caption Contest
Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff and Yejin Choi   14:00-15:30 (Frontenac

Ballroom and Queen's Quay)
Large neural networks can now generate jokes, but do they really "understand" humor? We challenge AI models with three tasks derived from the New Yorker Cartoon Caption Contest: matching a joke to a cartoon, identifying a winning caption, and explaining why a winning caption is funny. These tasks encapsulate progressively more sophisticated aspects of "understanding" a cartoon; key elements are the complex, often surprising relationships between images and captions and the frequent inclusion of indirect and playful allusions to human experience and culture. We investigate both multimodal and language-only models: the former are challenged with the cartoon images directly, while the latter are given multifaceted descriptions of the visual scene to simulate human-level visual understanding. We find that both types of models struggle at all three tasks. For example, our best multimodal models fall 30 accuracy points behind human performance on the matching task, and, even when provided ground-truth visual scene descriptors, human-authored explanations are preferred head-to-head over the best machine-authored ones (few-shot GPT-4) in more than 2/3 of cases. We release models, code, leaderboard, and corpus, which includes newly-gathered annotations describing the image's locations/entities, what's unusual in the scene, and an explanation of the joke.

MetaVL: Transferring In-Context Learning Ability From Language Models to Vision-Language Models
Masoud Monajatipoor, Liunian Harold Li, Mozhdeh Rouhsedaghat, Lin Yang and Kai-Wei Chang   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Large-scale language models have shown the ability to adapt to a new task via conditioning on a few demonstrations (i.e., in-context learning). However, in the vision-language domain, most large-scale pre-trained vision-language (VL) models do not possess the ability to conduct in-context learning. How can we enable in-context learning for VL models? In this paper, we study an interesting hypothesis: can we transfer the in-context learning ability from the language domain to the VL domain? Specifically, we first meta-trains a language model to perform in-context learning on NLP tasks (as in MetaICL); then we transfer this model to perform VL tasks by attaching a visual encoder. Our experiments suggest that indeed in-context learning ability can be transferred cross modalities: our model considerably improves the in-context learning capability on VL tasks and can even compensate for the size of the model significantly. On VQA, OK-VQA, and GQA, our method could outperform the baseline model while having 20 times fewer parameters.

Cross-modal Attention Congruence Regularization for Vision-Language Relation Alignment
Rohan Pandey, Rulin Shao, Paul Pu Liang, Ruslan Salakhutdinov and Louis-Philippe Morency   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Despite recent progress towards scaling up multimodal vision-language models, these models are still known to struggle on compositional generalization benchmarks such as Winoground. We find that a critical component lacking from current vision-language models is relation-level alignment: the ability to match directional semantic relations in text (e.g., 'mug in grass') with spatial relationships in the image (e.g., the position of the mug relative to the grass). To tackle this problem, we show that relation alignment can be enforced by encouraging the language attention from 'mug' to 'grass' (capturing the semantic relation 'in') to match the visual attention from the mug to the grass (capturing the corresponding physical relation). Tokens and their corresponding objects are softly identified using a weighted mean of cross-modal attention. We prove that this notion of soft cross-modal equivalence is equivalent to enforcing congruence between vision and language attention matrices under a 'change of basis' provided by the cross-modal attention matrix. Intuitively, our approach projects visual attention into the language attention space to calculate its divergence from the actual language attention, and vice versa. We apply our Cross-modal Attention Congruence Regularization (CACR) loss to fine-tune UNITER and improve its Winoground Group score by 5.75 points.

CLAPSpeech: Learning Prosody from Text Context with Contrastive Language-Audio Pre-Training
Zhenhui Ye, Rongjie Huang, Yi Ren, Ziyue Jiang, Jinglin Liu, Jinzheng He, Xiang Yin and Zhou Zhao 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Improving text representation has attracted much attention to achieve expressive text-to-speech (TTS). However, existing works only implicitly learn the prosody with masked token reconstruction tasks, which leads to low training efficiency and difficulty in prosody modeling. We propose CLAPSpeech, a cross-modal contrastive pre-training framework that learns from the prosody variance of the same text token under different contexts. Specifically, 1) with the design of a text encoder and a prosody encoder, we encourage the model to connect the text context with its corresponding prosody pattern in the joint multi-modal space; 2) we introduce a multi-scale pre-training pipeline to capture prosody patterns in multiple levels. 3) we show how to incorporate CLAPSpeech into existing TTS models for better prosody. Experiments on three datasets not only show that CLAPSpeech could improve the prosody prediction for existing TTS methods, but also demonstrate its generalization ability to adapt to multiple languages and multi-speaker text-to-speech. We also deeply analyze the principle behind the performance of CLAPSpeech. Ablation studies demonstrate the necessity of each component in CLAPSpeech. Source code and audio samples are available at https://clapspeech.github.io.

Deep Active Learning for Morphophonological Processing
Seyed Morteza Mirbostani, Yasaman Boreshban, Salam Khalifa, SeyedAbolghasem Mirroshandel and Owen Rambow   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Building a system for morphological processing is a challenging task in morphologically complex languages like Arabic. Although there are some deep learning based models that achieve successful results, these models rely on a large amount of annotated data. Building such datasets, specially for some of the lower-resource Arabic dialects, is very difficult, time-consuming, and expensive. In addition, some parts of the annotated data do not contain useful information for training machine learning models. Active learning strategies allow the learner algorithm to select the most informative samples for annotation. There has been little research that focuses on applying active learning for morphological inflection and morphophonological processing. In this paper, we have proposed a deep active learning method for this task. Our experiments on Egyptian Arabic show that with only about 30% of annotated data, we achieve the same results as does the state-of-the-art model on the whole dataset.

Morphological Inflection with Phonological Features
David Guriel, Omer Goldman and Reut Tsarfaty   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Recent years have brought great advances into solving morphological tasks, mostly due to powerful neural models applied to various tasks as (re)inflection and analysis. Yet, such morphological tasks cannot be considered solved, especially when little training data is available or when generalizing to previously unseen lemmas. This work explores effects on performance obtained through various ways in which morphological models get access to sub-character phonological features that are often the targets of morphological processes. We design two methods to achieve this goal: one that leaves models as is but manipulates the data to include features instead of characters, and another that manipulates models to take phonological features into account when building representations for phonemes. We elicit phonemic data from standard graphemic data using language-specific grammars for languages with shallow grapheme-to-phoneme mapping, and we experiment with two reinflection models over eight languages. Our results show that our methods yield comparable results to the grapheme-based baseline overall, with minor improvements in some of the languages. All in all, we conclude that patterns in character distributions are likely to allow models to infer the underlying phonological characteristics, even when phonemes are not explicitly represented.

Improving Generalization in Language Model-based Text-to-SQL Semantic Parsing: Two Simple Semantic Boundary-based Techniques

Daking Rai, Bailin Wang, Yilun Zhou and Ziyu Yao   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Compositional and domain generalization present significant challenges in semantic parsing, even for state-of-the-art semantic parsers based on pre-trained language models (LMs). In this study, we empirically investigate improving an LM's generalization in semantic parsing with two simple techniques: at the token level, we introduce a token preprocessing method to preserve the semantic boundaries of tokens produced by LM tokenizers; at the sequence level, we propose to use special tokens to mark the boundaries of components aligned between input and output. Our experimental results on two text-to-SQL semantic parsing datasets show that our token preprocessing, although simple, can substantially improve the LM performance on both types of generalization, and our component boundary marking method is particularly helpful for compositional generalization.

Metaphor Detection via Explicit Basic Meanings Modelling
Yucheng Li, Shun Wang, Chenghua Lin and Frank Guerin   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
One noticeable trend in metaphor detection is the embrace of linguistic theories such as the metaphor identification procedure (MIP) for model architecture design. While MIP clearly defines that the metaphoricity of a lexical unit is determined based on the contrast between its contextual meaning and its basic meaning, existing work does not strictly follow this principle, typically using the aggregated meaning to approximate the basic meaning of target words. In this paper, we propose a novel metaphor detection method, which models the basic meaning of the word based on literal annotation from the training set, and then compares this with the contextual meaning in a target sentence to identify metaphors. Empirical results show that our method outperforms the state-of-the-art method significantly by 1.0% in F1 score. Moreover, our performance even reaches the theoretical upper bound on the VUA18 benchmark for targets with basic annotations, which demonstrates the importance of modelling basic meanings for metaphor detection.

HyPe: Better Pre-trained Language Model Fine-tuning with Hidden Representation Perturbation
Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang and Songfang Huang   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Language models with the Transformers structure have shown great performance in natural language processing. However, there still poses problems when fine-tuning pre-trained language models on downstream tasks, such as over-fitting or representation collapse. In this work, we propose HyPe, a simple yet effective fine-tuning technique to alleviate such problems by perturbing hidden representations of Transformers layers. Unlike previous works that only add noise to inputs or parameters, we argue that the hidden representations of Transformers layers convey more diverse and meaningful language information. Therefore, making the Transformers layers more robust to hidden representation perturbations can further benefit the fine-tuning of PLMs en bloc. We conduct extensive experiments and analyses on GLUE and other natural language inference datasets. Results demonstrate that HyPe outperforms vanilla fine-tuning and enhances generalization of hidden representations from different layers. In addition, HyPe acquires negligible computational overheads, and is better than and compatible with previous state-of-the-art fine-tuning techniques.

DISCO: Distilling Counterfactuals with Large Language Models
Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal and Kyle Richardson   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Models trained with counterfactually augmented data learn representations of the causal structure of tasks, enabling robust generalization. However, high-quality counterfactual data is scarce for most tasks and not easily generated at scale. When crowdsourced, such data is typically limited in scale and diversity; when generated using supervised methods, it is computationally expensive to extend to new counterfactual dimensions. In this work, we introduce DISCO (DIStilled COunterfactual Data), a new method for automatically generating high-quality counterfactual data at scale. DISCO engineers prompts to generate phrasal perturbations with a large general language model. Then, a task-specific teacher model filters these generations to distill high-quality counterfactual data. While task-agnostic, we apply our pipeline to the task of natural language inference (NLI) and find that on challenging evaluations such as the NLI stress test, comparatively smaller student models trained with DISCO generated counterfactuals are more robust (6% absolute) and generalize better across distributions (2%) compared to models trained without data augmentation. Furthermore, DISCO augmented models are 10% more consistent between counterfactual pairs on three evaluation sets, demonstrating that DISCO augmentation enables models to more reliably learn causal representations. Our repository are available at: https://github.com/eric11eca/disco

Bring More Attention to Syntactic Symmetry for Automatic Postediting of High-Quality Machine Translations
Baikjin Jung, Myungji Lee, Jong-Hyeok Lee and Yunsu Kim   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Automatic postediting (APE) is an automated process to refine a given machine translation (MT). Recent findings present that existing APE systems are not good at handling high-quality MTs even for a language pair with abundant data resources, English–German: the better the given MT is, the harder it is to decide what parts to edit and how to fix these errors. One possible solution to this problem is to instill deeper knowledge about the target language into the model. Thus, we propose a linguistically motivated method of regularization that is expected to enhance APE models' understanding of the target language: a loss function that encourages symmetric self-attention on the given MT. Our analysis of experimental results demonstrates that the proposed method helps improving the state-of-the-art architecture's APE quality for high-quality MTs.

Neural Unsupervised Reconstruction of Protolanguage Word Forms
Andre W He, Nicholas Tomlin and Dan Klein   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
We present a state-of-the-art neural approach to the unsupervised reconstruction of ancient word forms. Previous work in this domain used expectation-maximization to predict simple phonological changes between ancient word forms and their cognates in modern languages. We extend this work with neural models that can capture more complicated phonological and morphological changes. At the same time, we preserve the inductive biases from classical methods by building monotonic alignment constraints into the model and deliberately underfitting during the maximization step. We evaluate our performance on the task of reconstructing Latin from a dataset of cognates across five Romance languages, achieving a notable reduction in edit distance from the target word forms compared to previous methods.

Resolving Indirect Referring Expressions for Entity Selection
Mohammad Javad Hosseini, Filip Radlinski, Silvia Pareti and Annie Louis   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Recent advances in language modeling have enabled new conversational systems. In particular, it is often desirable for people to make choices among specified options when using such systems. We address the problem of reference resolution, when people use natural expressions to choose between real world entities. For example, given the choice 'Should we make a Simnel cake or a Pandan cake¿ a natural response from a non-expert may be indirect: 'let's make the green one'. Reference resolution has been little studied with natural expressions, thus robustly understanding such language has large potential for improving naturalness in dialog, recommendation, and search systems. We create AltEntities (Alternative Entities), a new public dataset of entity pairs and utterances, and develop models for the disambiguation problem. Consisting of 42K indirect referring expressions across three domains, it enables for the first time the study of how large language models can be adapted to this task. We find they achieve 82%-87% accuracy in realistic settings, which while reasonable also invites further advances.

(QA)^2: Question Answering with Questionable Assumptions
Najoung Kim, Phu Mon Htut, Samuel R. Bowman and Jackson Petty   14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Naturally occurring information-seeking questions often contain questionable assumptions—assumptions that are false or unverifiable. Ques-

tions containing questionable assumptions are challenging because they require a distinct answer strategy that deviates from typical answers for information-seeking questions. For instance, the question 'When did Marie Curie discover Uranium?' cannot be answered as a typical 'when' question without addressing the false assumption 'Marie Curie discovered Uranium'. In this work, we propose $(QA)^2$ (Question Answering with Questionable Assumptions), an open-domain evaluation dataset consisting of naturally occurring search engine queries that may or may not contain questionable assumptions. To be successful on $(QA)^2$, systems must be able to detect questionable assumptions and also be able to produce adequate responses for both typical information-seeking questions and ones with questionable assumptions. Through human rater acceptability on end-to-end QA with $(QA)^2$, we find that current models do struggle with handling questionable assumptions, leaving substantial headroom for progress.

Improving Self-training for Cross-lingual Named Entity Recognition with Contrastive and Prototype Learning
Ran Zhou, Xin Li, Lidong Bing, Erik Cambria and Chunyan Miao 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
In cross-lingual named entity recognition (NER), self-training is commonly used to bridge the linguistic gap by training on pseudo-labeled target-language data. However, due to sub-optimal performance on target languages, the pseudo labels are often noisy and limit the overall performance. In this work, we aim to improve self-training for cross-lingual NER by combining representation learning and pseudo label refinement in one coherent framework. Our proposed method, namely ContProto mainly comprises two components: (1) contrastive self-training and (2) prototype-based pseudo-labeling. Our contrastive self-training facilitates span classification by separating clusters of different classes, and enhances cross-lingual transferability by producing closely-aligned representations between the source and target language. Meanwhile, prototype-based pseudo-labeling effectively improves the accuracy of pseudo labels during training. We evaluate ContProto on multiple transfer pairs, and experimental results show our method brings substantial improvements over current state-of-the-art methods.

DAMP: Doubly Aligned Multilingual Parser for Task-Oriented Dialogue
William Held, Christopher Hidey, Fei Liu, Eric Y Zhu, Rahul Goel, Diyi Yang and Rushin Shah 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Modern virtual assistants use internal semantic parsing engines to convert user utterances to actionable commands. However, prior work has demonstrated multilingual models are less robust for semantic parsing compared to other tasks. In global markets such as India and Latin America, robust multilingual semantic parsing is critical as codeswitching between languages is prevalent for bilingual users. In this work we dramatically improve the zero-shot performance of a multilingual and codeswitched semantic parsing system using two stages of multilingual alignment. First, we show that contrastive alignment pretraining improves *both* English performance and transfer efficiency. We then introduce a constrained optimization approach for hyperparameter-free adversarial alignment during finetuning. Our Doubly Aligned Multilingual Parser (DAMP) improves mBERT transfer performance by 3x, 6x, and 81x on the Spanglish, Hinglish and Multilingual Task Oriented Parsing benchmarks respectively and outperforms XLM-R and mT5-Large using 3.2x fewer parameters.

Analyzing Text Representations by Measuring Task Alignment
Cesar Gonzalez-Gutierrez, Audi Primadhanty, Francesco Cazzaro and Ariadna Julieta Quattoni 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Textual representations based on pre-trained language models are key, especially in few-shot learning scenarios. What makes a representation good for text classification? Is it due to the geometric properties of the space or because it is well aligned with the task? We hypothesize the second claim. To test it, we develop a task alignment score based on hierarchical clustering that measures alignment at different levels of granularity. Our experiments on text classification validate our hypothesis by showing that task alignment can explain the classification performance of a given representation.

REV: Information-Theoretic Evaluation of Free-Text Rationales
Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi and Swabha Swayamdipta 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Generating free-text rationales is a promising step towards explainable NLP, yet evaluating such rationales remains a challenge. Existing metrics have mostly focused on measuring the association between the rationale and a given label. We argue that an ideal metric should focus on the new information uniquely provided in the rationale that is otherwise not provided in the input or the label. We investigate this research problem from an information-theoretic perspective using conditional V-information (Hewitt et al., 2021). More concretely, we propose a metric called REV (Rationale Evaluation with conditional V-information), to quantify the amount of new, label-relevant information in a rationale beyond the information already available in the input or the label. Experiments across four benchmarks with reasoning tasks, including chain-of-thought, demonstrate the effectiveness of REV in evaluating rationale-label pairs, compared to existing metrics. We further demonstrate REV is consistent with human judgments on rationale evaluations and provides more sensitive measurements of new information in free-text rationales. When used alongside traditional performance metrics, REV provides deeper insights into models' reasoning and prediction processes.

Improving Syntactic Probing Correctness and Robustness with Control Tasks
Weicheng Ma, Brian C Wang, Hefan Zhang, Lili Wang, Rolando Coto-Solano, Saeed Hassanpour and Soroush Vosoughi 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Syntactic probing methods have been used to examine whether and how pre-trained language models (PLMs) encode syntactic features. However, the probing methods are usually biased by the PLMs' memorization of common word co-occurrences, even if they do not form syntactic relations. This paper presents a random-word-substitution and random-label-matching control task to reduce these biases and improve the robustness of syntactic probing methods. Our control tasks are also shown to notably improve the consistency of probing results between different probing methods and make the methods more robust with respect to the text attributes of the probing instances. Our control tasks make syntactic probing methods better at reconstructing syntactic features and more generalizable to unseen text domains. Our experiments show that our proposed control tasks are effective on different PLMs, probing methods, and syntactic features.

SeeGULL: A Stereotype Benchmark with Broad Geo-Cultural Coverage Leveraging Generative Models
Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran and Sunipa Dev 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Stereotype benchmark datasets are crucial to detect and mitigate social stereotypes about groups of people in NLP models. However, existing datasets are limited in size and coverage, and are largely restricted to stereotypes prevalent in the Western society. This is especially problematic as language technologies gain hold across the globe. To address this gap, we present SeeGULL, a broad-coverage stereotype dataset, built by utilizing generative capabilities of large language models such as PaLM, and GPT-3, and leveraging a globally diverse rater pool to validate the prevalence of those stereotypes in society. SeeGULL is in English, and contains stereotypes about identity groups spanning 178 countries across 8 different geo-political regions across 6 continents, as well as state-level identities within the US and India. We also include fine-grained offensiveness scores for different stereotypes and demonstrate their global disparities. Furthermore, we include comparative annotations about the same groups by annotators living in the region vs. those that are based in North America, and demonstrate that within-region stereotypes about groups differ from those prevalent in North America.

The Tail Wagging the Dog: Dataset Construction Biases of Social Bias Benchmarks

Nikil Roashan Selvam, Sunipa Dev, Daniel Khashabi, Tushar Khot and Kai-Wei Chang 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
How reliably can we trust the scores obtained from social bias benchmarks as faithful indicators of problematic social biases in a given model? In this work, we study this question by contrasting social biases with non-social biases that stem from choices made during dataset construction (which might not even be discernible to the human eye). To do so, we empirically simulate various alternative constructions for a given benchmark based on seemingly innocuous modifications (such as paraphrasing or random-sampling) that maintain the essence of their social bias. On two well-known social bias benchmarks (Winogender and BiasNLI), we observe that these shallow modifications have a surprising effect on the resulting degree of bias across various models and consequently the relative ordering of these models when ranked by measured bias.
We hope these troubling observations motivate more robust measures of social biases.

Are Sample-Efficient NLP Models More Robust?

Nelson F. Liu, Ananya Kumar, Percy Liang and Robin Jia 14:00-15:30 (Frontenac Ballroom and Queen's Quay)
Recent results in image classification and extractive question answering have observed that pre-trained models trained on less in-distribution data have better out-ofdistribution performance. However, it is unclear how broadly these trends hold. We conduct a large empirical study across three tasks, three broadly-applicable modeling interventions (increasing model size, using a different adaptation method, and pre-training on more data), and 14 diverse datasets to investigate the relationship between sample efficiency (amount of data needed to reach a given ID accuracy) and robustness (how models fare on OOD evaluation). We find that higher sample efficiency is only correlated with better average OOD robustness on some modeling interventions and tasks, but not others. On individual datasets, models with lower sample efficiency can even be more robust. These results suggest that general-purpose methods for improving sample efficiency are unlikely to yield universal OOD robustness improvements, since such improvements are highly dataset- and task-dependent. Even in an era of large, multi-purpose pre-trained models, task-specific decisions may often be necessary for OOD generalization.

Sentiment Analysis, Stylistic Analysis, and Argument Mining

14:00-15:30 (Pier 2&3)

Guiding Computational Stance Detection with Expanded Stance Triangle Framework

Zhengyuan Liu, Yong Keong Yap, Hai Leong Chieu and Nancy Chen 14:00-14:15 (Pier 2&3)
Stance detection determines whether the author of a piece of text is in favor of, against, or neutral towards a specified target, and can be used to gain valuable insights into social media. The ubiquitous indirect referral of targets makes this task challenging, as it requires computational solutions to model semantic features and infer the corresponding implications from a literal statement. Moreover, the limited amount of available training data leads to subpar performance in out-of-domain and cross-target scenarios, as data-driven approaches are prone to rely on superficial and domain-specific features. In this work, we decompose the stance detection task from a linguistic perspective, and investigate key components and inference paths in this task. The stance triangle is a generic linguistic framework previously proposed to describe the fundamental ways people express their stance. We further expand it by characterizing the relationship between explicit and implicit objects. We then use the framework to extend one single training corpus with additional annotation. Experimental results show that strategically-enriched data can significantly improve the performance on out-of-domain and cross-target evaluation.

A New Direction in Stance Detection: Target-Stance Extraction in the Wild

Yingjie Li, Krishna K Garg and Cornelia Caragea 14:15-14:30 (Pier 2&3)
Stance detection aims to detect the stance toward a corresponding target. Existing works use the assumption that the target is known in advance, which is often not the case in the wild. Given a text from social media platforms, the target information is often unknown due to implicit mentions in the source text and it is infeasible to have manual target annotations at a large scale. Therefore, in this paper, we propose a new task Target-Stance Extraction (TSE) that aims to extract the (target, stance) pair from the text. We benchmark the task by proposing a two-stage framework that first identifies the relevant target in the text and then detects the stance given the predicted target and text. Specifically, we first propose two different settings: Target Classification and Target Generation, to identify the potential target from a given text. Then we propose a multi-task approach that takes target prediction as the auxiliary task to detect the stance toward the predicted target. We evaluate the proposed framework on both in-target stance detection in which the test target is always seen in the training stage and zero-shot stance detection that needs to detect the stance for the targets that are unseen during the training phase. The new TSE task can facilitate future research in the field of stance detection.

Node Placement in Argument Maps: Modeling Unidirectional Relations in High & Low-Resource Scenarios

Iman Jundi, Neele Falk, Eva Maria Vecchi and Gabriella Lapesa 14:30-14:45 (Pier 2&3)
Argument maps structure discourse into nodes in a tree with each node being an argument that supports or opposes its parent argument. This format is more comprehensible and less redundant compared to an unstructured one. Exploring those maps and maintaining their structure by placing new arguments under suitable parents is more challenging for users with huge maps that are typical in online discussions.
To support those users, we introduce the task of node placement: suggesting candidate nodes as parents for a new contribution. We establish an upper-bound of human performance, and conduct experiments with models of various sizes and training strategies. We experiment with a selection of maps from Kialo, drawn from a heterogeneous set of domains.
Based on an annotation study, we highlight the ambiguity of the task that makes it challenging for both humans and models. We examine the unidirectional relation between tree nodes and show that encoding a node into different embeddings for each of the parent and child cases improves performance. We further show the few-shot effectiveness of our approach.

Topic-Guided Sampling For Data-Efficient Multi-Domain Stance Detection

Erik Arakelyan, Arnav Arora and Isabelle Augenstein 14:45-15:00 (Pier 2&3)
The task of Stance Detection is concerned with identifying the attitudes expressed by an author towards a target of interest. This task spans a variety of domains ranging from social media opinion identification to detecting the stance towards a legal claim. However, the framing of the task varies within these domains in terms of the data collection protocol, the label dictionary and the number of available annotations. Furthermore, these stance annotations are significantly imbalanced on a per-topic and inter-topic basis. These make multi-domain stance detection challenging, requiring standardization and domain adaptation. To overcome this challenge, we propose Topic Efficient StancE Detection (TESTED), consisting of a topic-guided diversity sampling technique used for creating a multi-domain data efficient training set and a contrastive objective that is used for fine-tuning a stance classifier using the produced set. We evaluate the method on an existing benchmark of 16 datasets with in-domain, i.e. all topics seen and out-of-domain, i.e. unseen topics, experiments. The results show that the method outperforms the state-of-the-art with an average of 3.5 F1 points increase in-domain and is more generalizable with an averaged 10.2 F1 on out-of-domain

evaluation while using <10% of the training data. We show that our sampling technique mitigates both inter- and per-topic class imbalances. Finally, our analysis demonstrates that the contrastive learning objective allows the model for a more pronounced segmentation of samples with varying labels.

C-STANCE: A Large Dataset for Chinese Zero-Shot Stance Detection

Chenye Zhao, Yingjie Li and Cornelia Caragea 15:00-15:15 (Pier 2&3)

Zero-shot stance detection (ZSSD) aims to determine whether the author of a text is in favor of, against, or neutral toward a target that is unseen during training. Despite the growing attention on ZSSD, most recent advances in this task are limited to English and do not pay much attention to other languages such as Chinese. To support ZSSD research, in this paper, we present C-STANCE that, to our knowledge, is the first Chinese dataset for zero-shot stance detection. We introduce two challenging subtasks for ZSSD: target-based ZSSD and domain-based ZSSD. Our dataset includes both noun-phrase targets and claim targets, covering a wide range of domains. We provide a detailed description and analysis of our dataset. To establish results on C-STANCE, we report performance scores using state-of-the-art deep learning models. We publicly release our dataset and code to facilitate future research.

[CL] Comparing Selective Masking Methods for Depression Detection in Social Media

Chanapa Pananookooln, Jakrapop Akaranee and Chaklam Silpasuwanchai 15:15-15:30 (Pier 2&3)

Identifying those at risk for depression is a crucial issue and social media provides an excellent platform for examining the linguistic patterns of depressed individuals. A significant challenge in depression classification problem is ensuring that prediction models are not overly dependent on topic keywords (i.e., depression keywords) such that it fails to predict when such keywords are unavailable. One promising approach is masking—that is, by selectively masking various words and asking the model to predict the masked words, the model is forced to learn the inherent language patterns of depression. This study evaluates seven masking techniques. Moreover, predicting the masked words during pre-training or fine-tuning phase was also examined. Last, six class imbalance ratios were compared to determine the robustness of masked word selection methods. Key findings demonstrate that selective masking outperforms random masking in terms of F1-score. The most accurate and robust models are identified. Our research also indicates that reconstructing the masked words during the pre-training phase is more advantageous than during the fine-tuning phase. Further discussion and implications are discussed. This is the first study to comprehensively compare masked word selection methods, which has broad implications for the field of depression classification and general NLP. Our code can be found at: https://github.com/chanapapan/Depression-Detection.

Language Grounding to Vision, Robotics, and Beyond

14:00-15:30 (Pier 4&5)

Cross2StrA: Unpaired Cross-lingual Image Captioning with Cross-lingual Cross-modal Structure-pivoted Alignment

Shengqiong Wu, Hao Fei, Wei Ji and Tat-Seng Chua 14:00-14:15 (Pier 4&5)

Unpaired cross-lingual image captioning has long suffered from irrelevancy and disfluency issues, due to the inconsistencies of the semantic scene and syntax attributes during transfer. In this work, we propose to address the above problems by incorporating the scene graph (SG) structures and the syntactic constituency (SC) trees. Our captioner contains the semantic structure-guided image-to-pivot captioning and the syntactic structure-guided pivot-to-target translation, two of which are joined via pivot language. We then take the SG and SC structures as pivoting, performing cross-modal semantic structure alignment and cross-lingual syntactic structure alignment learning. We further introduce cross-lingual&cross-modal back-translation training to fully align the captioning and translation stages. Experiments on English-Chinese transfers show that our model shows great superiority in improving captioning relevancy and fluency.

Why Did the Chicken Cross the Road? Rephrasing and Analyzing Ambiguous Questions in VQA

Elias Stengel-Eskin, Jimena Guallar-Blasco, Yi Zhou and Benjamin Van Durme 14:15-14:30 (Pier 4&5)

Natural language is ambiguous. Resolving ambiguous questions is key to successfully answering them. Focusing on questions about images, we create a dataset of ambiguous examples. We annotate these, grouping answers by the underlying question they address and rephrasing the question for each group to reduce ambiguity. Our analysis reveals a linguistically-aligned ontology of reasons for ambiguity in visual questions. We then develop an English question-generation model which we demonstrate via automatic and human evaluation produces less ambiguous questions. We further show that the question generation objective we use allows the model to integrate answer group information without any direct supervision.

VLN-Trans: Translator for the Vision and Language Navigation Agent

Yue Zhang and Parisa Kordjamshidi 14:30-14:45 (Pier 4&5)

Language understanding is essential for the navigation agent to follow instructions. We observe two kinds of issues in the instructions that can make the navigation task challenging: 1. The mentioned landmarks are not recognizable by the navigation agent due to the different vision abilities of the instructor and the modeled agent. 2. The mentioned landmarks are applicable to multiple targets, thus not distinctive for selecting the target among the candidate viewpoints. To deal with these issues, we design a translator module for the navigation agent to convert the original instructions into easy-to-follow sub-instruction representations at each step. The translator needs to focus on the recognizable and distinctive landmarks based on the agent's visual abilities and the observed visual environment. To achieve this goal, we create a new synthetic sub-instruction dataset and design specific tasks to train the translator and the navigation agent. We evaluate our approach on Room2Room (R2R), Room4room (R4R), and Room2Room Last (R2R-Last) datasets and achieve state-of-the-art results on multiple benchmarks.

Visually-augmented pretrained language models for NLP tasks without images

Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Qinyu Zhang and Ji-Rong Wen 14:45-15:00 (Pier 4&5)

Although pre-trained language models (PLMs) have shown impressive performance by text-only self-supervised training, they are found lack of visual semantics or commonsense. Existing solutions often rely on explicit images for visual knowledge augmentation (requiring time-consuming retrieval or generation), and they also conduct the augmentation for the whole input text, without considering whether it is actually needed in specific inputs or tasks. To address these issues, we propose a novel **V**isually-**A**ugmented fine-tuning approach that can be generally applied to various PLMs or NLP tasks, **W**ithout using any retrieved or generated **I**mages, namely **VAWI**. Experimental results show that our approach can consistently improve the performance of BERT, RoBERTa, BART, and T5 at different scales, and outperform several competitive baselines on ten tasks. Our codes and data are publicly available at https://github.com/RUCAIBox/VAWI.

Gloss-Free End-to-End Sign Language Translation

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, bang zhang and Yi Yang 15:00-15:15 (Pier 4&5)

In this paper, we tackle the problem of sign language translation (SLT) without gloss annotations. Although intermediate representation like

gloss has been proven effective, gloss annotations are hard to acquire, especially in large quantities. This limits the domain coverage of translation datasets, thus handicapping real-world applications. To mitigate this problem, we design the Gloss-Free End-to-end sign language translation framework (GloFE). Our method improves the performance of SLT in the gloss-free setting by exploiting the shared underlying semantics of signs and the corresponding spoken translation. Common concepts are extracted from the text and used as a weak form of intermediate representation. The global embedding of these concepts is used as a query for cross-attention to find the corresponding information within the learned visual features. In a contrastive manner, we encourage the similarity of query results between samples containing such concepts and decrease those that do not. We obtained state-of-the-art results on large-scale datasets, including OpenASL and How2Sign.

VisText: A Benchmark for Semantically Rich Chart Captioning
Benny J. Tang, Angie Boggust and Arvind Satyanarayan — 15:15-15:30 (Pier 4&5)
Captions that describe or explain charts help improve recall and comprehension of the depicted data and provide a more accessible medium for people with visual disabilities. However, current approaches for automatically generating such captions struggle to articulate the perceptual or cognitive features that are the hallmark of charts (e.g., complex trends and patterns). In response, we introduce VisText: a dataset of 12,441 pairs of charts and captions that describe the charts' construction, report key statistics, and identify perceptual and cognitive phenomena. In VisText, a chart is available as three representations: a rasterized image, a backing data table, and a *scene graph* — a hierarchical representation of a chart's visual elements akin to a web page's Document Object Model (DOM). To evaluate the impact of VisText, we fine-tune state-of-the-art language models on our chart captioning task and apply prefix-tuning to produce captions that vary the semantic content they convey. Our models generate coherent, semantically rich captions and perform on par with state-of-the-art chart captioning models across machine translation and text generation metrics. Through qualitative analysis, we identify six broad categories of errors that our models make that can inform future work.

Syntax: Tagging, Chunking, and Parsing
14:00-15:30 (Pier 7&8)

Holographic CCG Parsing
Ryosuke Yamaki, Tadahiro Taniguchi and Daichi Mochihashi — 14:00-14:15 (Pier 7&8)
We propose a method for formulating CCG as a recursive composition in a continuous vector space. Recent CCG supertagging and parsing models generally demonstrate high performance, yet rely on black-box neural architectures to implicitly model phrase structure dependencies. Instead, we leverage the method of holographic embeddings as a compositional operator to explicitly model the dependencies between words and phrase structures in the embedding space. Experimental results revealed that holographic composition effectively improves the supertagging accuracy to achieve state-of-the-art parsing performance when using a C&C parser. The proposed span-based parsing algorithm using holographic composition achieves performance comparable to state-of-the-art neural parsing with Transformers. Furthermore, our model can semantically and syntactically infill text at the phrase level due to the decomposability of holographic composition.

Don't Parse, Choose Spans! Continuous and Discontinuous Constituency Parsing via Autoregressive Span Selection
Songlin Yang and Kewei Tu — 14:15-14:30 (Pier 7&8)
We present a simple and unified approach for both continuous and discontinuous constituency parsing via autoregressive span selection. Constituency parsing aims to produce a set of non-crossing spans so that they can form a constituency parse tree. We sort gold spans using a predefined order and leverage a pointer network to autoregressively select spans by that order. To deal with discontinuous spans, we consecutively select their subspans from left to right, label all but last subspans with special discontinuous labels and the last subspan as the whole discontinuous spans' labels. We use simple heuristic to output valid trees so that our approach is able to predict all possible continuous and discontinuous constituency trees without sacrificing data coverage and without the need to use expensive chart-based parsing algorithms. Experiments on multiple continuous and discontinuous benchmarks show that our model achieves state-of-the-art or competitive performance.

Contextual Distortion Reveals Constituency: Masked Language Models are Implicit Parsers
Jiaxi Li and Wei Lu — 14:30-14:45 (Pier 7&8)
Recent advancements in pre-trained language models (PLMs) have demonstrated that these models possess some degree of syntactic awareness. To leverage this knowledge, we propose a novel chart-based method for extracting parse trees from masked language models (LMs) without the need to train separate parsers. Our method computes a score for each span based on the distortion of contextual representations resulting from linguistic perturbations. We design a set of perturbations motivated by the linguistic concept of constituency tests, and use these to score each span by aggregating the distortion scores. To produce a parse tree, we use chart parsing to find the tree with the minimum score. Our method consistently outperforms previous state-of-the-art methods on English with masked LMs, and also demonstrates superior performance in a multilingual setting, outperforming the state-of-the-art in 6 out of 8 languages. Notably, although our method does not involve parameter updates or extensive hyperparameter search, its performance can even surpass some unsupervised parsing methods that require fine-tuning. Our analysis highlights that the distortion of contextual representation resulting from syntactic perturbation can serve as an effective indicator of constituency across languages.

Unsupervised Discontinuous Constituency Parsing with Mildly Context-Sensitive Grammars
Songlin Yang, Roger Levy and Yoon Kim — 14:45-15:00 (Pier 7&8)
We study grammar induction with mildly context-sensitive grammars for unsupervised discontinuous parsing. Using the probabilistic linear context-free rewriting system (LCFRS) formalism, our approach fixes the rule structure in advance and focuses on parameter learning with maximum likelihood. To reduce the computational complexity of both parsing and parameter estimation, we restrict the grammar formalism to LCFRS-2 (i.e., binary LCFRS with fan-out two) and further discard rules that require $O(l^6)$ time to parse, reducing inference to $O(l^5)$. We find that using a large number of nonterminals is beneficial and thus make use of tensor decomposition-based rank-space dynamic programming with an embedding-based parameterization of rule probabilities to scale up the number of nonterminals. Experiments on German and Dutch show that our approach is able to induce linguistically meaningful trees with continuous and discontinuous structures.

[CL] Cross-Lingual Transfer with Language-Specific Subnetworks for Low-Resource Dependency Parsing
Rochelle Choenni, Dan Garrette and Ekaterina Shutova — 15:00-15:15 (Pier 7&8)
Large multilingual language models typically share their parameters across all languages, which enables cross-lingual task transfer, but learning can also be hindered when training updates from different languages are in conflict. In this article, we propose novel methods for using language-specific subnetworks, which control cross-lingual parameter sharing, to reduce conflicts and increase positive transfer during fine-tuning. We introduce dynamic subnetworks, which are jointly updated with the model, and we combine our methods with meta-learning, an established, but complementary, technique for improving cross-lingual transfer. Finally, we provide extensive analyses of how each of our methods affects the models.

Hexatagging: Projective Dependency Parsing as Tagging
Afra Amini, Tianyu Liu and Ryan Cotterell 15:15-15:30 (Pier 7&8)
We introduce a novel dependency parser, the hexatagger, that constructs dependency trees by tagging the words in a sentence with elements from a finite set of possible tags. In contrast to many approaches to dependency parsing, our approach is fully parallelizable at training time, i.e., the structure-building actions needed to build a dependency parse can be predicted in parallel to each other. Additionally, exact decoding is linear in time and space complexity. Furthermore, we derive a probabilistic dependency parser that predicts hexatags using no more than a linear model with features from a pretrained language model, i.e., we forsake a bespoke architecture explicitly designed for the task. Despite the generality and simplicity of our approach, we achieve state-of-the-art performance of 96.4 LAS and 97.4 UAS on the Penn Treebank test set. Additionally, our parser's linear time complexity and parallelism significantly improve computational efficiency, with a roughly 10-times speed-up over previous state-of-the-art models during decoding.

Spotlight Session - 19:00-21:00

Findings Spotlights I

19:00-21:00 (Metropolitan East)

OpenPI-C: A Better Benchmark and Stronger Baseline for Open-Vocabulary State Tracking
Xueqing Wu, Sha Li and Heng Ji 19:00-21:00 (Metropolitan East)
Open-vocabulary state tracking is a more practical version of state tracking that aims to track state changes of a process without restricting the state space and entity space. OpenPI (Tandon et al., 2020) is to date the only dataset annotated for open-vocabulary state tracking. However, we identify issues with the dataset quality and evaluation metric. For the dataset, we categorize 3 types of problems on the procedure level, step level and state change level respectively, and build a clean dataset OpenPI-C using multiple rounds of human judgment. For the evaluation metric, we propose a cluster-based metric to fix the original metric's preference for repetition.
Model-wise, we enhance the seq2seq generation baseline by reinstating two key properties for state tracking: temporal dependency and entity awareness. The state of the world after an action is inherently dependent on the previous state. We model this dependency through a dynamic memory bank and allow the model to attend to the memory slots during decoding. On the other hand, the state of the world is naturally a union of the states of involved entities. Since the entities are unknown in the open-vocabulary setting, we propose a two-stage model that refines the state change prediction conditioned on entities predicted from the first stage. Empirical results show the effectiveness of our proposed model, especially on the cleaned dataset and the cluster-based metric. The code and data are released at https://github.com/shirley-wu/openpi-c

Exploiting Hierarchically Structured Categories in Fine-grained Chinese Named Entity Recognition
Jiuding Yang, Jinwen Luo, Weidong Guo, Di Niu and Yu Xu 19:00-21:00 (Metropolitan East)
Chinese Named Entity Recognition (CNER) is a widely used technology in various applications. While recent studies have focused on utilizing additional information of the Chinese language and characters to enhance CNER performance, this paper focuses on a specific aspect of CNER known as fine-grained CNER (FG-CNER). FG-CNER involves the use of hierarchical, fine-grained categories (e.g. Person-MovieStar) to label named entities. To promote research in this area, we introduce the FiNE dataset, a dataset for FG-CNER consisting of 30,000 sentences from various domains and containing 67,651 entities in 54 fine-grained flattened hierarchical categories. Additionally, we propose SoftFiNE, a novel approach for FG-CNER that utilizes a custom-designed relevance scoring function based on label structures to learn the potential relevance between different flattened hierarchical labels. Our experimental results demonstrate that the proposed SoftFiNE method outperforms the state-of-the-art baselines on the FiNE dataset. Furthermore, we conduct extensive experiments on three other datasets, including OntoNotes 4.0, Weibo, and Resume, where SoftFiNE achieved state-of-the-art performance on all three datasets.

Correction of Errors in Preference Ratings from Automated Metrics for Text Generation
Jan Deriu, Pius von Däniken, Don Tuggener and Mark Cieliebak 19:00-21:00 (Metropolitan East)
A major challenge in the field of Text Generation is evaluation: Human evaluations are cost-intensive, and automated metrics often display considerable disagreements with human judgments. In this paper, we propose to apply automated metrics for Text Generation in a preference-based evaluation protocol. The protocol features a statistical model that incorporates various levels of uncertainty to account for the error-proneness of the metrics. We show that existing metrics are generally over-confident in assigning significant differences between systems. As a remedy, the model allows to combine human ratings with automated ratings. We show that it can reduce the required amounts of human ratings to arrive at robust and statistically significant results by more than 50%, while yielding the same evaluation outcome as the pure human evaluation in 95% of cases. We showcase the benefits of the evaluation protocol for three text generation tasks: dialogue systems, machine translation, and text summarization.

HeGeL: A Novel Dataset for Geo-Location from Hebrew Text
Tzuf Paz-Argaman, Tal Bauman, Itai Mondshine, Itzhak Omer, Sagi Dalyot and Reut Tsarfaty 19:00-21:00 (Metropolitan East)
The task of textual geolocation — retrieving the coordinates of a place based on a free-form language description — calls for not only grounding but also natural language understanding and geospatial reasoning. Even though there are quite a few datasets in English used for geolocation, they are currently based on open-source data (Wikipedia and Twitter), where the location of the described place is mostly implicit, such that the location retrieval resolution is limited. Furthermore, there are no datasets available for addressing the problem of textual geolocation in morphologically rich and resource-poor languages, such as Hebrew. In this paper, we present the Hebrew Geo-Location (HeGeL) corpus, designed to collect literal place descriptions and analyze lingual geospatial reasoning. We crowdsourced 5,649 literal Hebrew place descriptions of various place types in three cities in Israel. Qualitative and empirical analysis show that the data exhibits abundant use of geospatial reasoning and requires a novel environmental representation.

An Annotated Dataset for Explainable Interpersonal Risk Factors of Mental Disturbance in Social Media Posts
Muskan Garg, Amirmohammad Shahbandegan, Amrit Chadha and Vijay Mago 19:00-21:00 (Metropolitan East)
With a surge in identifying suicidal risk and its severity in social media posts, we argue that a more consequential and explainable research is required for optimal impact on clinical psychology practice and personalized mental healthcare. The success of computational intelligence techniques for inferring mental illness from social media resources, points to natural language processing as a lens for determining Interpersonal Risk Factors (IRF) in human writings. Motivated with limited availability of datasets for social NLP research community, we construct and release a new annotated dataset with human-labelled explanations and classification of IRF affecting mental disturbance on social media: (i) Thwarted Belongingness (TBe), and (ii) Perceived Burdensomeness (PBu). We establish baseline models on our dataset facilitating future research directions to develop real-time personalized AI models by detecting patterns of TBe and PBu in emotional spectrum of user's histor-

ical social media profile.

Echoes from Alexandria: A Large Resource for Multilingual Book Summarization

Alessandro Scirè, Simone Conia, Simone Ciciliano and Roberto Navigli 19:00-21:00 (Metropolitan East)

In recent years, research in text summarization has mainly focused on the news domain, where texts are typically short and have strong layout features. The task of full-book summarization presents additional challenges which are hard to tackle with current resources, due to their limited size and availability in English only. To overcome these limitations, we present "Echoes from Alexandria", or in shortened form, "Echoes", a large resource for multilingual book summarization. Echoes features three novel datasets: i) Echo-Wiki, for multilingual book summarization, ii) Echo-XSum, for extremely-compressive multilingual book summarization, and iii) Echo-FairySum, for extractive book summarization. To the best of our knowledge, Echoes – with its thousands of books and summaries – is the largest resource, and the first to be multilingual, featuring 5 languages and 25 language pairs. In addition to Echoes, we also introduce a new extractive-then-abstractive baseline, and, supported by our experimental results and manual analysis of the summaries generated, we argue that this baseline is more suitable for book summarization than purely-abstractive approaches. We release our resource and software at https://github.com/Babelscape/echoes-from-alexandria in the hope of fostering innovative research in multilingual book summarization.

A Unified Evaluation Framework for Novelty Detection and Accommodation in NLP with an Instantiation in Authorship Attribution

Neeraj Varshney, Himanshu Gupta, Eric Robertson, Bing Liu and Chitta Baral 19:00-21:00 (Metropolitan East)

State-of-the-art natural language processing models have been shown to achieve remarkable performance in 'closed-world' settings where all the labels in the evaluation set are known at training time. However, in real-world settings, 'novel' instances that do not belong to any known class are often observed. This renders the ability to deal with novelties crucial. To initiate a systematic research in this important area of 'dealing with novelties', we introduce NoveltyTask, a multi-stage task to evaluate a system's performance on pipelined novelty 'detection' and 'accommodation' tasks. We provide mathematical formulation of NoveltyTask and instantiate it with the authorship attribution task that pertains to identifying the correct author of a given text. We use amazon reviews corpus and compile a large dataset (consisting of 250k instances across 200 authors/labels) for NoveltyTask. We conduct comprehensive experiments and explore several baseline methods for the task. Our results show that the methods achieve considerably low performance making the task challenging and leaving sufficient room for improvement. Finally, we believe our work will encourage research in this underexplored area of dealing with novelties, an important step en route to developing robust systems.

RQUGE: Reference-Free Metric for Evaluating Question Generation by Answering the Question

Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan, James Henderson and Marzieh Saeidi 19:00-21:00 (Metropolitan East)

Existing metrics for evaluating the quality of automatically generated questions such as BLEU, ROUGE, BERTScore, and BLEURT compare the reference and predicted questions, providing a high score when there is a considerable lexical overlap or semantic similarity between the candidate and the reference questions. This approach has two major shortcomings. First, we need expensive human-provided reference questions. Second, it penalises valid questions that may not have high lexical or semantic similarity to the reference questions. In this paper, we propose a new metric, RQUGE, based on the answerability of the candidate question given the context. The metric consists of a question-answering and a span scorer modules, using pre-trained models from existing literature, thus it can be used without any further training. We demonstrate that RQUGE has a higher correlation with human judgment without relying on the reference question. Additionally, RQUGE is shown to be more robust to several adversarial corruptions. Furthermore, we illustrate that we can significantly improve the performance of QA models on out-of-domain datasets by fine-tuning on synthetic data generated by a question generation model and reranked by RQUGE.

C-XNLI: Croatian Extension of XNLI Dataset

Leo Obadić, Andrej Jertec, Marko Rajnović and Branimir Dropuljić 19:00-21:00 (Metropolitan East)

Comprehensive multilingual evaluations have been encouraged by emerging cross-lingual benchmarks and constrained by existing parallel datasets. To partially mitigate this limitation, we extended the Cross-lingual Natural Language Inference (XNLI) corpus with Croatian. The development and test sets were translated by a professional translator, and we show that Croatian is consistent with other XNLI dubs. The train set is translated using Facebook's 1.2B parameter m2m_100 model. We thoroughly analyze the Croatian train set and compare its quality with the existing machine-translated German set. The comparison is based on 2000 manually scored sentences per language using a variant of the Direct Assessment (DA) score commonly used at the Conference on Machine Translation (WMT). Our findings reveal that a less-resourced language like Croatian is still lacking in translation quality of longer sentences compared to German. However, both sets have a substantial amount of poor quality translations, which should be considered in translation-based training or evaluation setups.

ANALOGICAL - A Novel Benchmark for Long Text Analogy Evaluation in Large Language Models

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal Gajera, Shreeyash Mukul Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth and Amitava Das 19:00-21:00 (Metropolitan East)

Over the past decade, analogies, in the form of word-level analogies, have played a significant role as an intrinsic measure of evaluating the quality of word embedding methods such as word2vec. Modern large language models (LLMs), however, are primarily evaluated on extrinsic measures based on benchmarks such as GLUE and SuperGLUE, and there are only a few investigations on whether LLMs can draw analogies between long texts. In this paper, we present ANALOGICAL, a new benchmark to intrinsically evaluate LLMs across a taxonomy of analogies of long text with six levels of complexity – (i) word, (ii) word vs. sentence, (iii) syntactic, (iv) negation, (v) entailment, and (vi) metaphor. Using thirteen datasets and three different distance measures, we evaluate the abilities of eight LLMs in identifying analogical pairs in the semantic vector space. Our evaluation finds that it is increasingly challenging for LLMs to identify analogies when going up the analogy taxonomy.

LEDA: a Large-Organization Email-Based Decision-Dialogue-Act Analysis Dataset

Mladen Karan, Prashant Khare, Ravi Shekhar, Stephen McQuistin, Ignacio Castro, Gareth Tyson, Colin Perkins, Patrick G.T. Healey and Matthew Purver 19:00-21:00 (Metropolitan East)

Collaboration increasingly happens online. This is especially true for large groups working on global tasks, with collaborators all around the globe. The size and distributed nature of such groups makes decision-making challenging. This paper proposes a set of dialog acts for the study of decision-making mechanisms in such groups, and provides a new annotated dataset based on real-world data from the public mail-archives of one such organisation – the Internet Engineering Task Force (IETF). We provide an initial data analysis showing that this dataset can be used to better understand decision-making in such organisations. Finally, we experiment with a preliminary transformer-based dialog act tagging model.

Varta: A Large-Scale Headline-Generation Dataset for Indic Languages

Rahul Aralikatte, Ziling Cheng, Sumanth Doddapaneni and Jackie Chi Kit Cheung 19:00-21:00 (Metropolitan East)

We present Varta, a large-scale multilingual dataset for headline generation in Indic languages. This dataset includes more than 41 million pairs of headlines and articles in 14 different Indic languages (and English), which come from a variety of high-quality news sources. To the best of our knowledge, this is the largest collection of curated news articles for Indic languages currently available. We use the collected

data in a series of experiments to answer important questions related to Indic NLP and multilinguality research in general. We show that the dataset is challenging even for state-of-the-art abstractive models and that they perform only slightly better than extractive baselines. Owing to its size, we also show that the dataset can be used to pre-train strong language models that outperform competitive baselines in both NLU and NLG benchmarks.

NusaCrowd: Open Source Initiative for Indonesian NLP Resources
Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Indra Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Halim Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Akbar Septiandri, James Jaya, Kaus-tubh Dhole, Arie Suryani, Rifki Afina Putri, Dan Su, Keith David Stevens, Made Nindyatama Nityasya, Muhammad Farid Adilazuarda, Ryan Ignatius Hadiwijaya, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Inastra Damapuspita, Haryo Akbarianto Wibowo, Cuk Tho, Ichwanul Muslim Karo Karo, Tirana Noor Fatyanosa, Ziwei Ji, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Pascale Fung, Herry Sujaini, Sakriani Sakti and Ayu Purwarianti 19:00-21:00 (Metropolitan East)
We present NusaCrowd, a collaborative initiative to collect and unify existing resources for Indonesian languages, including opening access to previously non-public resources. Through this initiative, we have brought together 137 datasets and 118 standardized data loaders. The quality of the datasets has been assessed manually and automatically, and their value is demonstrated through multiple experiments. NusaCrowd's data collection enables the creation of the first zero-shot benchmarks for natural language understanding and generation in Indonesian and the local languages of Indonesia. Furthermore, NusaCrowd brings the creation of the first multilingual automatic speech recognition benchmark in Indonesian and the local languages of Indonesia. Our work strives to advance natural language processing (NLP) research for languages that are under-represented despite being widely spoken.

ORCA: A Challenging Benchmark for Arabic Language Understanding
AbdelRahim Elmadany, ElMoatez Billah Nagoudi and Muhammad Abdul-Mageed 19:00-21:00 (Metropolitan East)
Due to the crucial role pretrained language models play in modern NLP, several benchmarks have been proposed to evaluate their perfor-mance. In spite of these efforts, no public benchmark of diverse nature currently exists for evaluating Arabic NLU. This makes it challenging to measure progress for both Arabic and multilingual language models. This challenge is compounded by the fact that any benchmark target-ing Arabic needs to take into account the fact that Arabic is not a single language but rather a collection of languages and language varieties. In this work, we introduce a publicly available benchmark for Arabic language understanding evaluation dubbed ORCA. It is carefully con-structed to cover diverse Arabic varieties and a wide range of challenging Arabic understanding tasks exploiting 60 different datasets (across seven NLU task clusters). To measure current progress in Arabic NLU, we use ORCA to offer a comprehensive comparison between 18 mul-tilingual and Arabic language models. We also provide a public leaderboard with a unified single-number evaluation metric (ORCA score) to facilitate future research.

InfoSync: Information Synchronization across Multilingual Semi-structured Tables
Siddharth Hemant Khincha, Chelsi Jain, Vivek Gupta, Tushar Kataria and Shuo Zhang 19:00-21:00 (Metropolitan East)
Information Synchronization of semi-structured data across languages is challenging. For example, Wikipedia tables in one language need to be synchronized with others. To address this problem, we introduce a new dataset InfoSync and a two-step method for tabular synchroniza-tion. InfoSync contains 100K entity-centric tables (Wikipedia Infoboxes) across 14 languages, of which a subset (3.5K pairs) are manually annotated. The proposed method includes 1) Information Alignment to map rows and 2) Information Update for updating missing/outdated information for aligned tables across multilingual tables. When evaluated on InfoSync, information alignment achieves an F1 score of 87.91 (en <-> non-en). To evaluate information updation, we perform human-assisted Wikipedia edits on Infoboxes for 532 table pairs. Our ap-proach obtains an acceptance rate of 77.28% on Wikipedia, showing the effectiveness of the proposed method.

Take a Break in the Middle: Investigating Subgoals towards Hierarchical Script Generation
XINZE LI, Yixin Cao, Muhao Chen and Aixin Sun 19:00-21:00 (Metropolitan East)
Goal-oriented Script Generation is a new task of generating a list of steps that can fulfill the given goal. In this paper, we propose to extend the task from the perspective of cognitive theory. Instead of a simple flat structure, the steps are typically organized hierarchically — Human often decompose a complex task into subgoals, where each subgoal can be further decomposed into steps. To establish the benchmark, we contribute a new dataset, propose several baseline methods, and set up evaluation metrics. Both automatic and human evaluation verify the high-quality of dataset, as well as the effectiveness of incorporating subgoals into hierarchical script generation. Furthermore, We also design and evaluate the model to discover subgoal, and find that it is a bit more difficult to decompose the goals than summarizing from segmented steps.

ISLTranslate: Dataset for Translating Indian Sign Language
Abhinav Joshi, Susmit Agrawal and Ashutosh Modi 19:00-21:00 (Metropolitan East)
Sign languages are the primary means of communication for many hard-of-hearing people worldwide. Recently, to bridge the communication gap between the hard-of-hearing community and the rest of the population, several sign language translation datasets have been proposed to enable the development of statistical sign language translation systems. However, there is a dearth of sign language resources for the Indian sign language. This resource paper introduces ISLTranslate, a translation dataset for continuous Indian Sign Language (ISL) consisting of 31k ISL-English sentence/phrase pairs. To the best of our knowledge, it is the largest translation dataset for continuous Indian Sign Language. We provide a detailed analysis of the dataset. To validate the performance of existing end-to-end Sign language to spoken language translation systems, we benchmark the created dataset with a transformer-based model for ISL translation.

PropSegmEnt: A Large-Scale Corpus for Proposition-Level Segmentation and Entailment Recognition
Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, Dan Roth and Tal Schuster 19:00-21:00 (Metropolitan East)
The widely studied task of Natural Language Inference (NLI) requires a system to recognize whether one piece of text is textually entailed by another, i.e. whether the entirety of its meaning can be inferred from the other. In current NLI datasets and models, textual entailment relations are typically defined on the sentence- or paragraph-level. However, even a simple sentence often contains multiple propositions, i.e. distinct units of meaning conveyed by the sentence. As these propositions can carry different truth values in the context of a given premise, we argue for the need to recognize the textual entailment relation of each proposition in a sentence individually. We propose PropSegmEnt, a corpus of over 45K propositions annotated by expert human raters. Our dataset structure resembles the tasks of (1) segmenting sentences within a document to the set of propositions, and (2) classifying the entailment relation of each proposition with respect to a different yet topically-aligned document, i.e. documents describing the same event or entity. We establish strong baselines for the segmentation and entailment tasks. Through case studies on summary hallucination detection and document-level NLI, we demonstrate that our conceptual framework is potentially useful for understanding and explaining the compositionality of NLI labels.

Revisiting Sample Size Determination in Natural Language Understanding
Ernie Chang, Muhammad Hassan Rashid, Pin-Jie Lin, Changsheng Zhao, Vera Demberg, Yangyang Shi and Vikas Chandra 19:00-21:00 (Metropolitan East)

Knowing exactly how many data points need to be labeled to achieve a certain model performance is a hugely beneficial step towards reducing the overall budgets for annotation. It pertains to both active learning and traditional data annotation, and is particularly beneficial for low resource scenarios. Nevertheless, it remains a largely under-explored area of research in NLP. We therefore explored various techniques for estimating the training sample size necessary to achieve a targeted performance value. We derived a simple yet effective approach to predict the maximum achievable model performance based on small amount of training samples – which serves as an early indicator during data annotation for data quality and sample size determination. We performed ablation studies on four language understanding tasks, and showed that the proposed approach allows us to forecast model performance within a small margin of mean absolute error (0.9%) with only 10% data.

Dynamic Structured Neural Topic Model with Self-Attention Mechanism
Nozomu Miyamoto, Masaru Isonuma, Sho Takase, Junichiro Mori and Ichiro Sakata   19:00-21:00 (Metropolitan East)
This study presents a dynamic structured neural topic model, which can handle the time-series development of topics while capturing their dependencies. Our model captures the topic branching and merging processes by modeling topic dependencies based on a self-attention mechanism. Additionally, we introduce citation regularization, which induces attention weights to represent citation relations by modeling text and citations jointly. Our model outperforms a prior dynamic embedded topic model regarding perplexity and coherence, while maintaining sufficient diversity across topics. Furthermore, we confirm that our model can potentially predict emerging topics from academic literature.

Discrete Prompt Optimization via Constrained Generation for Zero-shot Re-ranker
Sukmin Cho, Soyeong Jeong, Jeong yeon Seo and Jong Park   19:00-21:00 (Metropolitan East)
Re-rankers, which order retrieved documents with respect to the relevance score on the given query, have gained attention for the information retrieval (IR) task. Rather than fine-tuning the pre-trained language model (PLM), the large-scale language model (LLM) is utilized as a zero-shot re-ranker with excellent results. While LLM is highly dependent on the prompts, the impact and the optimization of the prompts for the zero-shot re-ranker are not explored yet. Along with highlighting the impact of optimization on the zero-shot re-ranker, we propose a novel discrete prompt optimization method, Constrained Prompt generation (Co-Prompt), with the metric estimating the optimum for re-ranking. Co-Prompt guides the generated texts from PLM toward optimal prompts based on the metric without parameter update. The experimental results demonstrate that Co-Prompt leads to outstanding re-ranking performance against the baselines. Also, Co-Prompt generates more interpretable prompts for humans against other prompt optimization methods.

DynaMiTE: Discovering Explosive Topic Evolutions with User Guidance
Nishant Balepur, Shivam Agarwal, Karthik Venkat Ramanan, Susik Yoon, Diyi Yang and Jiawei Han   19:00-21:00 (Metropolitan East)
Dynamic topic models (DTMs) analyze text streams to capture the evolution of topics. Despite their popularity, existing DTMs are either fully supervised, requiring expensive human annotations, or fully unsupervised, producing topic evolutions that often do not cater to a user's needs. Further, the topic evolutions produced by DTMs tend to contain generic terms that are not indicative of their designated time steps. To address these issues, we propose the task of discriminative dynamic topic discovery. This task aims to discover topic evolutions from temporal corpora that distinctly align with a set of user-provided category names and uniquely capture topics at each time step. We solve this task by developing DynaMiTE, a framework that ensembles semantic similarity, category indicative, and time indicative scores to produce informative topic evolutions. Through experiments on three diverse datasets, including the use of a newly-designed human evaluation experiment, we demonstrate that DynaMiTE is a practical and efficient framework for helping users discover high-quality topic evolutions suited to their interests.

Large Language Models are Built-in Autoregressive Search Engines
Noah Ziems, Wenhao Yu, Zhihan Zhang and Meng Jiang   19:00-21:00 (Metropolitan East)
Document retrieval is a key stage of standard Web search engines. Existing dual-encoder dense retrievers obtain representations for questions and documents independently, allowing for only shallow interactions between them. To overcome this limitation, recent autoregressive search engines replace the dual-encoder architecture by directly generating identifiers for relevant documents in the candidate pool. However, the training cost of such autoregressive search engines rises sharply as the number of candidate documents increases. In this paper, we find that large language models (LLMs) can follow human instructions to directly generate URLs for document retrieval.

Surprisingly, when providing a few Query-URL pairs as in-context demonstrations, LLMs can generate Web URLs where nearly 90% of the corresponding documents contain correct answers to open-domain questions. In this way, LLMs can be thought of as built-in search engines, since they have not been explicitly trained to map questions to document identifiers. Experiments demonstrate that our method can consistently achieve better retrieval performance than existing retrieval approaches by a significant margin on three open-domain question answering benchmarks, under both zero and few-shot settings. The code for this work can be found at https://github.com/Ziems/llm-url.

Nonparametric Decoding for Generative Retrieval
*Hyunji Lee, JaeYoung Kim, Hoyeon Chang, hanseok Oh, Sohee Yang, Vladimir Karpukhin, Yi Lu and Minjoon Seo*19:00-21:00 (Metropolitan East)
The generative retrieval model depends solely on the information encoded in its model parameters without external memory, its information capacity is limited and fixed. To overcome the limitation, we propose Nonparametric Decoding (Np Decoding) which can be applied to existing generative retrieval models. Np Decoding uses nonparametric contextualized vocab embeddings (external memory) rather than vanilla vocab embeddings as decoder vocab embeddings. By leveraging the contextualized vocab embeddings, the generative retrieval model is able to utilize both the parametric and nonparametric space. Evaluation over 9 datasets (8 single-hop and 1 multi-hop) in the document retrieval task shows that applying Np Decoding to generative retrieval models significantly improves the performance. We also show that Np Decoding is data- and parameter-efficient, and shows high performance in the zero-shot setting.

Recurrent Attention Networks for Long-text Modeling
Xianming Li, Zongxi Li, Xiaotian Luo, Haoran Xie, Xing Lee, Yingbin Zhao, Fu Lee Wang and Qing Li   19:00-21:00 (Metropolitan East)
Self-attention-based models have achieved remarkable progress in short-text mining. However, the quadratic computational complexities restrict their application in long text processing. Prior works have adopted the chunking strategy to divide long documents into chunks and stack a self-attention backbone with the recurrent structure to extract semantic representation. Such an approach disables parallelization of the attention mechanism, significantly increasing the training cost and raising hardware requirements. Revisiting the self-attention mechanism and the recurrent structure, this paper proposes a novel long-document encoding model, Recurrent Attention Network (RAN), to enable the recurrent operation of self-attention. Combining the advantages from both sides, the well-designed RAN is capable of extracting global semantics in both token-level and document-level representations, making it inherently compatible with both sequential and classification tasks, respectively. Furthermore, RAN is computationally scalable as it supports parallelization on long document processing. Extensive experiments demonstrate the long-text encoding ability of the proposed RAN model on both classification and sequential tasks, showing its potential for a wide range of applications.

SamToNe: Improving Contrastive Loss for Dual Encoder Retrieval Models with Same Tower Negatives
Fedor Moiseev, Gustavo Hernandez Abrego, Peter Dornbach, Imed Zitouni, Enrique Alfonseca and Zhe Dong   19:00-21:00 (Metropolitan East)

Dual encoders have been used for retrieval tasks and representation learning with good results. A standard way to train dual encoders is using a contrastive loss with in-batch negatives. In this work, we propose an improved contrastive learning objective by adding queries or documents from the same encoder towers to the negatives, for which we name it as "contrastive loss with SAMe TOwer NEgatives" (SamToNe). By evaluating on question answering retrieval benchmarks from MS MARCO and MultiReQA, and heterogenous zero-shot information retrieval benchmarks (BEIR), we demonstrate that SamToNe can effectively improve the retrieval quality for both symmetric and asymmetric dual encoders. By directly probing the embedding spaces of the two encoding towers via the t-SNE algorithm (van der Maaten and Hinton, 2008), we observe that SamToNe ensures the alignment between the embedding spaces from the two encoder towers. Based on the analysis of the embedding distance distributions of the top-1 retrieved results, we further explain the efficacy of the method from the perspective of regularisation.

Task-aware Retrieval with Instructions
Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi and Wen-tau Yih   19:00-21:00 (Metropolitan East)
We study the problem of retrieval with instructions, where users provide explicit descriptions of their intent along with their queries to guide a retrieval system. Our solution is a general-purpose task-aware retrieval system, trained using multi-task instruction tuning and can follow human-written instructions to find relevant documents to a given query. We introduce the first large-scale collection of 37 retrieval datasets with instructions, BERRI, and present TART, a single multi-task retrieval system trained on BERRI with instructions that can adapt to a new task without any parameter updates. TART advances the state of the art on two zero-shot retrieval benchmarks, BEIR and LOTTE, outperforming models up to three times larger. We further introduce a new evaluation setup, $X\hat{2}$-Retrieval, to better reflect real-world scenarios in which diverse domains and tasks are pooled. TART significantly outperforms competitive baselines in this setup, further highlighting the effectiveness of guiding retrieval with instructions.

Detecting Adversarial Samples through Sharpness of Loss Landscape
Rui Zheng, Shihan Dou, Yuhao Zhou, Qin Liu, Tao Gui, Qi Zhang, Zhongyu Wei, Xuanjing Huang and Menghan Zhang   19:00-21:00 (Metropolitan East)
Deep neural networks (DNNs) have been proven to be sensitive towards perturbations on input samples, and previous works highlight that adversarial samples are even more vulnerable than normal ones. In this work, this phenomenon is illustrated frWe first show that adversarial samples locate in steep and narrow local minima of the loss landscape (high sharpness) while normal samples, which differs distinctly from adversarial ones, reside in the loss surface that is more flatter (low sharpness).om the perspective of sharpness via visualizing the input loss landscape of models. Based on this, we propose a simple and effective sharpness-based detector to distinct adversarial samples by maximizing the loss increment within the region where the inference sample is located. Considering that the notion of sharpness of a loss landscape is relative, we further propose an adaptive optimization strategy in an attempt to fairly compare the relative sharpness among different samples. Experimental results show that our approach can outperform previous detection methods by large margins (average +6.6 F1 score) for four advanced attack strategies considered in this paper across three text classification tasks.

Score It All Together: A Multi-Task Learning Study on Automatic Scoring of Argumentative Essays
Yuning Ding, Marie Bexte and Andrea Horbach   19:00-21:00 (Metropolitan East)
When scoring argumentative essays in an educational context, not only the presence or absence of certain argumentative elements but also their quality is important. On the recently published student essay dataset PERSUADE, we first show that the automatic scoring of argument quality benefits from additional information about context, writing prompt and argument type. We then explore the different combinations of three tasks: automated span detection, type and quality prediction. Results show that a multi-task learning approach combining the three tasks outperforms sequential approaches that first learn to segment and then predict the quality/type of a segment.

An Exploration of Encoder-Decoder Approaches to Multi-Label Classification for Legal and Biomedical Text
Yova Kementchedjhieva and Ilias Chalkidis   19:00-21:00 (Metropolitan East)
Standard methods for multi-label text classification largely rely on encoder-only pre-trained language models, whereas encoder-decoder models have proven more effective in other classification tasks. In this study, we compare four methods for multi-label classification, two based on an encoder only, and two based on an encoder-decoder. We carry out experiments on four datasets—two in the legal domain and two in the biomedical domain, each with two levels of label granularity— and always depart from the same pre-trained model, T5. Our results show that encoder-decoder methods outperform encoder-only methods, with a growing advantage on more complex datasets and labeling schemes of finer granularity. Using encoder-decoder models in a non-autoregressive fashion, in particular, yields the best performance overall, so we further study this approach through ablations to better understand its strengths.

Prompt- and Trait Relation-aware Cross-prompt Essay Trait Scoring
Heejin Do, Yunsu Kim and Gary Geunbae Lee   19:00-21:00 (Metropolitan East)
Automated essay scoring (AES) aims to score essays written for a given prompt, which defines the writing topic. Most existing AES systems assume to grade essays of the same prompt as used in training and assign only a holistic score. However, such settings conflict with real-education situations; pre-graded essays for a particular prompt are lacking, and detailed trait scores of sub-rubrics are required. Thus, predicting various trait scores of unseen-prompt essays (called cross-prompt essay trait scoring) is a remaining challenge of AES. In this paper, we propose a robust model: prompt- and trait relation-aware cross-prompt essay trait scorer. We encode prompt-aware essay representation by essay-prompt attention and utilizing the topic-coherence feature extracted by the topic-modeling mechanism without access to labeled data; therefore, our model considers the prompt adherence of an essay, even in a cross-prompt setting. To facilitate multi-trait scoring, we design trait-similarity loss that encapsulates the correlations of traits. Experiments prove the efficacy of our model, showing state-of-the-art results for all prompts and traits. Significant improvements in low-resource-prompt and inferior traits further indicate our model's strength.

Towards Diverse and Effective Question-Answer Pair Generation from Children Storybooks
Sugyeong Eo, Hyeonseok Moon, Jinsung Kim, Yuna Hur, Jeongwook Kim, SongEun Lee, Changwoo Chun, SUNGSOO PARK and Heuiseok Lim   19:00-21:00 (Metropolitan East)
Recent advances in QA pair generation (QAG) have raised interest in applying this technique in the educational field. However, the diversity of QA types remains a challenge despite its contributions to comprehensive learning and assessment of children. In this paper, we propose a QAG framework that enhances QA type diversity by producing different interrogative sentences and implicit/explicit answers. Our framework comprises a QFS-based answer generator, an iterative QA generator, and a relevancy-aware ranker. The two generators aim to expand the number of candidates while covering various types. The ranker trained on the in-context negative samples clarifies the top-N outputs based on the ranking score. Extensive evaluations and detailed analyses demonstrate that our approach outperforms previous state-of-the-art results by significant margins, achieving improved diversity and quality. Our task-oriented processes are consistent with real-world demand, which highlights our system's high applicability.

Similarity-Based Content Scoring - A more Classroom-Suitable Alternative to Instance-Based Scoring?
Marie Bexte, Andrea Horbach and Torsten Zesch   19:00-21:00 (Metropolitan East)

Automatically scoring student answers is an important task that is usually solved using instance-based supervised learning. Recently, similarity-based scoring has been proposed as an alternative approach yielding similar perfor- mance. It has hypothetical advantages such as a lower need for annotated training data and better zero-shot performance, both of which are properties that would be highly beneficial when applying content scoring in a realistic classroom setting. In this paper we take a closer look at these alleged advantages by comparing different instance-based and similarity-based methods on multiple data sets in a number of learning curve experiments. We find that both the demand on data and cross-prompt performance is similar, thus not confirming the former two suggested advantages. The by default more straightforward possibility to give feedback based on a similarity-based approach may thus tip the scales in favor of it, although future work is needed to explore this advantage in practice.

Sequential Path Signature Networks for Personalised Longitudinal Language Modeling
Talia Tseriotou, Adam Tsakalidis, Peter Foster, Terence J Lyons and Maria Liakata   19:00-21:00 (Metropolitan East)
Longitudinal user modeling can provide a strong signal for various downstream tasks. Despite the rapid progress in representation learning, dynamic aspects of modelling individuals' language have only been sparsely addressed. We present a novel extension of neural sequential models using the notion of path signatures from rough path theory, which constitute graduated summaries of continuous paths and have the ability to capture non-linearities in trajectories. By combining path signatures of users' history with contextual neural representations and recursive neural networks we can produce compact time-sensitive user representations. Given the magnitude of mental health conditions with symptoms manifesting in language, we show the applicability of our approach on the task of identifying changes in individuals' mood by analysing their online textual content. By directly integrating signature transforms of users' history in the model architecture we jointly address the two most important aspects of the task, namely sequentiality and temporality. Our approach achieves state-of-the-art performance on macro-average F1 score on the two available datasets for the task, outperforming or performing on-par with state-of-the-art models utilising only historical posts and even outperforming prior models which also have access to future posts of users.

Contrastive Training Improves Zero-Shot Classification of Semi-structured Documents
Muhammad Khalifa, Yogarshi Vyas, Shuai Wang, Graham Horwood, Sunil Mallya and Miguel Ballesteros   19:00-21:00 (Metropolitan East)
We investigate semi-structured document classification in a zero-shot setting. Classification of semi-structured documents is more challenging than that of standard unstructured documents, as positional, layout, and style information play a vital role in interpreting such documents. The standard classification setting where categories are fixed during both training and testing falls short in dynamic environments where new classification categories could potentially emerge. We focus exclusively on the zero-shot learning setting where inference is done on new unseen classes. To address this task, we propose a matching-based approach that relies on a pairwise contrastive objective for both pretraining and fine-tuning. Our results show a significant boost in Macro F1 from the proposed pretraining step and comparable performance of the contrastive fine-tuning to a standard prediction objective in both supervised and unsupervised zero-shot settings.

GVdoc - Graph-based Visual DOcument Classification
Fnu Mohbat, Mohammed J Zaki, Catherine Finegan-Dollak and Ashish Verma   19:00-21:00 (Metropolitan East)
The robustness of a model for real-world deployment is decided by how well it performs on unseen data and distinguishes between in-domain and out-of-domain samples. Visual document classifiers have shown impressive performance on in-distribution test sets. However, they tend to have a hard time correctly classifying and differentiating out-of-distribution examples. Image-based classifiers lack the text component, whereas multi-modality transformer-based models face the token serialization problem in visual documents due to their diverse layouts. They also require a lot of computing power during inference, making them impractical for many real-world applications. We propose, GVdoc, a graph-based document classification model that addresses both of these challenges. Our approach generates a document graph based on its layout, and then trains a graph neural network to learn node and graph embeddings. Through experiments, we show that our model, even with fewer parameters, outperforms state-of-the-art models on out-of-distribution data while retaining comparable performance on the in-distribution test set.

Distractor Generation based on Text2Text Language Models with Pseudo Kullback-Leibler Divergence Regulation
Hui-Juan Wang, Kai-Yu Hsieh, Han-Cheng Yu, Jui-Ching Tsou, Yu An Shih, Chen-Hua Huang and Yao-Chung Fan   19:00-21:00
(Metropolitan East)
In this paper, we address the task of cloze-style multiple choice question (MCQs) distractor generation. Our study is featured by the following designs. First, we propose to formulate the cloze distractor generation as a Text2Text task. Second, we propose pseudo Kullback-Leibler Divergence for regulating the generation to consider the item discrimination index in education evaluation. Third, we explore the candidate augmentation strategy and multi-tasking training with cloze-related tasks to further boost the generation performance. Through experiments with benchmarking datasets, our best performing model advances the state-of-the-art result from 10.81 to 22.00 (p@1 score).

Financial Numeric Extreme Labelling: A dataset and benchmarking
Soumya Sharma, Subhendu Khatuya, Manjunath Hegde, Afreen Shaikh, Koustuv Dasgupta, Pawan Goyal and Niloy Ganguly   19:00-21:00
(Metropolitan East)
The U.S. Securities and Exchange Commission (SEC) mandates all public companies to file periodic financial statements that should contain numerals annotated with a particular label from a taxonomy. In this paper, we formulate the task of automating the assignment of a label to a particular numeral span in a sentence from an extremely large label set. Towards this task, we release a dataset, Financial Numeric Extreme Labelling (FNXL), annotated with 2,794 labels. We benchmark the performance of the FNXL dataset by formulating the task as (a) a sequence labelling problem and (b) a pipeline with span extraction followed by Extreme Classification. Although the two approaches perform comparably, the pipeline solution provides a slight edge for the least frequent labels.

Scientific Fact-Checking: A Survey of Resources and Approaches
Juraj Vladika and Florian Matthes   19:00-21:00 (Metropolitan East)
The task of fact-checking deals with assessing the veracity of factual claims based on credible evidence and background knowledge. In particular, scientific fact-checking is the variation of the task concerned with verifying claims rooted in scientific knowledge. This task has received significant attention due to the growing importance of scientific and health discussions on online platforms. Automated scientific fact-checking methods based on NLP can help combat the spread of misinformation, assist researchers in knowledge discovery, and help individuals understand new scientific breakthroughs. In this paper, we present a comprehensive survey of existing research in this emerging field and its related tasks. We provide a task description, discuss the construction process of existing datasets, and analyze proposed models and approaches. Based on our findings, we identify intriguing challenges and outline potential future directions to advance the field.

Understanding Programs by Exploiting (Fuzzing) Test Cases
Jianyu Zhao, Yuyang Rong, Yiwen Guo, Yifeng He and Hao Chen   19:00-21:00 (Metropolitan East)
Semantic understanding of programs has attracted great attention in the community. Inspired by recent successes of large language models (LLMs) in natural language understanding, tremendous progress has been made by treating programming language as another sort of natural language and training LLMs on corpora of program code. However, programs are essentially different from texts after all, in a sense that they are normally heavily structured and syntax-strict. In particular, programs and their basic units (i.e., functions and subroutines) are designed

to demonstrate a variety of behaviors and/or provide possible outputs, given different inputs. The relationship between inputs and possible outputs/behaviors represents the functions/subroutines and profiles the program as a whole. Hence, we propose to incorporate such a relationship into learning, for achieving a deeper semantic understanding of programs. To obtain inputs that are representative enough to trigger the execution of most part of the code, we resort to fuzz testing and propose fuzz tuning to boost the performance of program understanding and code representation learning, given a pre-trained LLM. The effectiveness of the proposed method is verified on two program understanding tasks including code clone detection and code classification, and it outperforms current state-of-the-arts by large margins. Code is available at https://github.com/rabbitjy/FuzzTuning.

Dialogue Planning via Brownian Bridge Stochastic Process for Goal-directed Proactive Dialogue
Jian Wang, Dongding Lin and Wenjie Li 19:00-21:00 (Metropolitan East)
Goal-directed dialogue systems aim to proactively reach a pre-determined target through multi-turn conversations. The key to achieving this task lies in planning dialogue paths that smoothly and coherently direct conversations towards the target. However, this is a challenging and under-explored task. In this work, we propose a coherent dialogue planning approach that uses a stochastic process to model the temporal dynamics of dialogue paths. We define a latent space that captures the coherence of goal-directed behavior using a Brownian bridge process, which allows us to incorporate user feedback flexibly in dialogue planning. Based on the derived latent trajectories, we generate dialogue paths explicitly using pre-trained language models. We finally employ these paths as natural language prompts to guide dialogue generation. Our experiments show that our approach generates more coherent utterances and achieves the goal with a higher success rate.

Injecting Comparison Skills in Task-Oriented Dialogue Systems for Database Search Results Disambiguation
Yongil Kim, Yerin Hwang, Joongbo Shin, Hyunkyung Bae and Kyomin Jung 19:00-21:00 (Metropolitan East)
In task-oriented dialogue (TOD) systems designed to aid users accomplish specific goals in one or more domains, the agent retrieves entities that satisfy user constraints from the database. However, when multiple database search results exist, an ambiguity occurs regarding which results to select and present to the user. Existing TOD systems handle this ambiguity by randomly selecting one or few results and presenting their names to the user. However, in a real scenario, users do not always accept a randomly recommended entity, and users should have access to more comprehensive information about the search results. To address this limitation, we propose a new task called Comparison-Based database search Ambiguity handling (CBA), which handles ambiguity in database search results by comparing the properties of multiple entities to enable users to choose according to their preferences. Accordingly, we introduce a new framework for automatically collecting high-quality dialogue data along with the Disambiguating Schema-guided Dialogue (DSD) dataset, an augmented version of the SGD dataset. Experimental studies on the DSD dataset demonstrate that training baseline models with the dataset effectively address the CBA task. Our dataset and code will be publicized.

CausalDialogue: Modeling Utterance-level Causality in Conversations
Yi-Lin Tuan, Alon Albalak, Wenda Xu, Michael S Saxon, Connor F Pryor, Lise Getoor and William Yang Wang 19:00-21:00 (Metropolitan East)
Despite their widespread adoption, neural conversation models have yet to exhibit natural chat capabilities with humans. In this research, we examine user utterances as causes and generated responses as effects, recognizing that changes in a cause should produce a different effect. To further explore this concept, we have compiled and expanded upon a new dataset called CausalDialogue through crowd-sourcing. This dataset includes multiple cause-effect pairs within a directed acyclic graph (DAG) structure. Our analysis reveals that traditional loss functions struggle to effectively incorporate the DAG structure, leading us to propose a causality-enhanced method called Exponential Maximum Average Treatment Effect (ExMATE) to enhance the impact of causality at the utterance level in training neural conversation models. To evaluate the needs of considering causality in dialogue generation, we built a comprehensive benchmark on CausalDialogue dataset using different models, inference, and training methods. Through experiments, we find that a causality-inspired loss like ExMATE can improve the diversity and agility of conventional loss function and there is still room for improvement to reach human-level quality on this new dataset.

Multi-Domain Dialogue State Tracking with Disentangled Domain-Slot Attention
Longfei Jin, Jiyi Li, Sheng Li and Takahiro Shinozaki 19:00-21:00 (Metropolitan East)
As the core of task-oriented dialogue systems, dialogue state tracking (DST) is designed to track the dialogue state through the conversation between users and systems. Multi-domain DST has been an important challenge in which the dialogue states across multiple domains need to consider. In recent mainstream approaches, each domain and slot are aggregated and regarded as a single query feeding into attention with the dialogue history to obtain domain-slot specific representations. In this work, we propose disentangled domain-slot attention for multi-domain dialogue state tracking. The proposed approach disentangles the domain-slot specific information extraction in a flexible and context-dependent manner by separating the query about domains and slots in the attention component. Through a series of experiments on MultiWOZ 2.0 and MultiWOZ 2.4 datasets, we demonstrate that our proposed approach outperforms the standard multi-head attention with aggregated domain-slot query.

Intent Discovery with Frame-guided Semantic Regularization and Augmentation
yajing sun, Rui Zhang, Jingyuan Yang and Wei Peng 19:00-21:00 (Metropolitan East)
Most existing intent discovery methods leverage representation learning and clustering to transfer the prior knowledge of known intents to unknown ones. The learned representations are limited to the syntactic forms of sentences, therefore, fall short of recognizing adequate variations under the same meaning of unknown intents. This paper proposes an approach utilizing frame knowledge as conceptual semantic guidance to bridge the gap between known intents representation learning and unknown intents clustering. Specifically, we employ semantic regularization to minimize the bidirectional KL divergence between model predictions for frame-based and sentence-based samples. Moreover, we construct a frame-guided data augmenter to capture intent-friendly semantic information and implement contrastive clustering learning for unsupervised sentence embedding. Extensive experiments on two benchmark datasets show that our method achieves substantial improvements in accuracy (5%+) compared to solid baselines.

DEnsity: Open-domain Dialogue Evaluation Metric using Density Estimation
ChaeHun Park, Seungil Chad Lee, Daniel Rim and Jaegul Choo 19:00-21:00 (Metropolitan East)
Despite the recent advances in open-domain dialogue systems, building a reliable evaluation metric is still a challenging problem. Recent studies proposed learnable metrics based on classification models trained to distinguish the correct response. However, neural classifiers are known to make overly confident predictions for examples from unseen distributions. We propose DENSITY, which evaluates a response by utilizing density estimation on the feature space derived from a neural classifier. Our metric measures how likely a response would appear in the distribution of human conversations. Moreover, to improve the performance of DENSITY, we utilize contrastive learning to further compress the feature space. Experiments on multiple response evaluation datasets show that DENSITY correlates better with human evaluations than the existing metrics.

Multimodal Recommendation Dialog with Subjective Preference: A New Challenge and Benchmark
Yuxing Long, Binyuan Hui, Caixia Yuan, Fei Huang, Yongbin Li and Xiaojie WANG 19:00-21:00 (Metropolitan East)
Existing multimodal task-oriented dialog data fails to demonstrate the diverse expressions of user subjective preferences and recommendation

acts in the real-life shopping scenario. This paper introduces a new dataset SURE (Multimodal Recommendation Dialog with Subjective Preference), which contains 12K shopping dialogs in complex store scenes. The data is built in two phases with human annotations to ensure quality and diversity. SURE is well-annotated with subjective preferences and recommendation acts proposed by sales experts. A comprehensive analysis is given to reveal the distinguishing features of SURE. Three benchmark tasks are then proposed on the data to evaluate the capability of multimodal recommendation agents. Basing on the SURE, we propose a baseline model, powered by a state-of-the-art multimodal model, for these tasks.

Multi3NLU++: A Multilingual, Multi-Intent, Multi-Domain Dataset for Natural Language Understanding in Task-Oriented Dialogue

Nikita Moghe, Evgeniia Razumovskaia, Liane K Guillou, Ivan Vulić, Anna Korhonen and Alexandra Birch   19:00-21:00 (Metropolitan East)
Task-oriented dialogue (ToD) systems have been widely deployed in many industries as they deliver more efficient customer support. These systems are typically constructed for a single domain or language and do not generalise well beyond this. To support work on Natural Language Understanding (NLU) in ToD across multiple languages and domains simultaneously, we constructed Multi3NLU++, a multilingual, multi-intent, multi-domain dataset. Multi3NLU++ extends the English-only NLU++ dataset to include manual translations into a range of high, medium, and low resource languages (Spanish, Marathi, Turkish and Amharic), in two domains (banking and hotels). Because of its multi-intent property, Multi3NLU++ represents complex and natural user goals, and therefore allows us to measure the realistic performance of ToD systems in a varied set of the world's languages. We use Multi3NLU++ to benchmark state-of-the-art multilingual models for the NLU tasks of intent detection and slot labeling for ToD systems in the multilingual setting. The results demonstrate the challenging nature of the dataset, particularly in the low-resource language setting, offering ample room for future experimentation in multi-domain multilingual ToD setups.

Prompted LLMs as Chatbot Modules for Long Open-domain Conversation

Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos and Kangwook Lee   19:00-21:00 (Metropolitan East)
In this paper, we propose MPC (Modular Prompted Chatbot), a new approach for creating high-quality conversational agents without the need for fine-tuning. Our method utilizes pre-trained large language models (LLMs) as individual modules for long-term consistency and flexibility, by using techniques such as few-shot prompting, chain-of-thought (CoT), and external memory. Our human evaluation results show that MPC is on par with fine-tuned chatbot models in open-domain conversations, making it an effective solution for creating consistent and engaging chatbots.

Zero-Shot Prompting for Implicit Intent Prediction and Recommendation with Commonsense Reasoning

Hui-Chi Kuo and Yun-Nung Chen   19:00-21:00 (Metropolitan East)
The current generation of intelligent assistants require explicit user requests to perform tasks or services, often leading to lengthy and complex conversations. In contrast, human assistants can infer multiple implicit intents from utterances via their commonsense knowledge, thereby simplifying interactions. To bridge this gap, this paper proposes a framework for multi-domain dialogue systems. This framework automatically infers implicit intents from user utterances, and prompts a large pre-trained language model to suggest suitable task-oriented bots. By leveraging commonsense knowledge, our framework recommends associated bots in a zero-shot manner, enhancing interaction efficiency and effectiveness. This approach substantially reduces interaction complexity, seamlessly integrates various domains and tasks, and represents a significant step towards creating more human-like intelligent assistants that can reason about implicit intents, offering a superior user experience.

Boosting Distress Support Dialogue Responses with Motivational Interviewing Strategy

Anuradha Welivita and Pearl Pu   19:00-21:00 (Metropolitan East)
AI-driven chatbots have become an emerging solution to address psychological distress. Due to the lack of psychotherapeutic data, researchers use dialogues scraped from online peer support forums to train them. But since the responses in such platforms are not given by professionals, they contain both conforming and non-conforming responses. In this work, we attempt to recognize these conforming and non-conforming response types present in online distress-support dialogues using labels adapted from a well-established behavioral coding scheme named Motivational Interviewing Treatment Integrity (MITI) code and show how some response types could be rephrased into a more MI adherent form that can, in turn, enable chatbot responses to be more compliant with the MI strategy. As a proof of concept, we build several rephrasers by fine-tuning Blender and GPT3 to rephrase MI non-adherent Advise without permission responses into Advise with permission. We show how this can be achieved with the construction of pseudo-parallel corpora avoiding costs for human labor. Through automatic and human evaluation we show that in the presence of less training data, techniques such as prompting and data augmentation can be used to produce substantially good rephrasings that reflect the intended style and preserve the content of the original text.

Diverse Retrieval-Augmented In-Context Learning for Dialogue State Tracking

Brendan King and Jeffrey Flanigan   19:00-21:00 (Metropolitan East)
There has been significant interest in zero and few-shot learning for dialogue state tracking (DST) due to the high cost of collecting and annotating task-oriented dialogues. Recent work has demonstrated that in-context learning requires very little data and zero parameter updates, and even outperforms trained methods in the few-shot setting. We propose RefPyDST, which advances the state of the art with three advancements to in-context learning for DST. First, we formulate DST as a Python programming task, explicitly modeling language coreference as variable reference in Python. Second, since in-context learning depends highly on the context examples, we propose a method to retrieve a diverse set of relevant examples to improve performance. Finally, we introduce a novel re-weighting method during decoding that takes into account probabilities of competing surface forms, and produces a more accurate dialogue state prediction. We evaluate our approach using MultiWOZ and achieve state-of-the-art multi-domain joint-goal accuracy in zero and few-shot settings.

The Whole Truth and Nothing But the Truth: Faithful and Controllable Dialogue Response Generation with Dataflow Transduction and Constrained Decoding

Hao Fang, Anusha Balakrishnan, Harsh Jhamtani, John Bufe, Jean Crawford, Jayant Krishnamurthy, Adam Pauls, Jason Eisner, Jacob Andreas and Dan Klein   19:00-21:00 (Metropolitan East)
In a real-world dialogue system, generated text must be truthful and informative while remaining fluent and adhering to a prescribed style. Satisfying these constraints simultaneously is difficult for the two predominant paradigms in language generation: neural language modeling and rule-based generation. We describe a hybrid architecture for dialogue response generation that combines the strengths of both paradigms. The first component of this architecture is a rule-based content selection model defined using a new formal framework called dataflow transduction, which uses declarative rules to transduce a dialogue agent's actions and their results (represented as dataflow graphs) into context-free grammars representing the space of contextually acceptable responses. The second component is a constrained decoding procedure that uses these grammars to constrain the output of a neural language model, which selects fluent utterances. Our experiments show that this system outperforms both rule-based and learned approaches in human evaluations of fluency, relevance, and truthfulness.

Imagination is All You Need! Curved Contrastive Learning for Abstract Sequence Modeling Utilized on Long Short-Term Dialogue Planning

Justus-Jonas Erker, Stefan Schaffer and Gerasimos Spanakis   19:00-21:00 (Metropolitan East)
Inspired by the curvature of space-time, we introduce Curved Contrastive Learning (CCL), a novel representation learning technique for learning the relative turn distance between utterance pairs in multi-turn dialogues. The resulting bi-encoder models can guide transformers as a response ranking model towards a goal in a zero-shot fashion by projecting the goal utterance and the corresponding reply candidates into a latent space. Here the cosine similarity indicates the distance/reachability of a candidate utterance toward the corresponding goal. Furthermore, we explore how these forward-neutral language representations can be utilized for assessing the likelihood of sequences by the entailment strength i.e. through the cosine similarity of its individual members (encoded separately) as an emergent property in the curved space. These non-local properties allow us to imagine the likelihood of future patterns in dialogues, specifically by ordering/identifying future goal utterances that are multiple turns away, given a dialogue context. As part of our analysis, we investigate characteristics that make conversations (un)plannable and find strong evidence of planning capability over multiple turns (in 61.56% over 3 turns) in conversations from the DailyDialog dataset. Finally, we show how we achieve higher efficiency in sequence modeling tasks compared to previous work thanks to our relativistic approach, where only the last utterance needs to be encoded and computed during inference.

Leveraging Explicit Procedural Instructions for Data-Efficient Action Prediction
Julia Isabel White, Arushi Raghuvanshi and Yada Pruksachatkun   19:00-21:00 (Metropolitan East)
Task-oriented dialogues often require agents to enact complex, multi-step procedures in order to meet user requests. While large language models have found success automating these dialogues in constrained environments, their widespread deployment is limited by the substantial quantities of task-specific data required for training. The following paper presents a data-efficient solution to constructing dialogue systems, leveraging explicit instructions derived from agent guidelines, such as company policies or customer service manuals. Our proposed Knowledge-Augmented Dialogue System (KADS) combines a large language model with a knowledge retrieval module that pulls documents outlining relevant procedures from a predefined set of policies, given a user-agent interaction. To train this system, we introduce a semi-supervised pre-training scheme that employs dialogue-document matching and action-oriented masked language modeling with partial parameter freezing. We evaluate the effectiveness of our approach on prominent task-oriented dialogue datasets, Action-Based Conversations Dataset and Schema-Guided Dialogue, for two dialogue tasks: action state tracking and workflow discovery. Our results demonstrate that procedural knowledge augmentation improves accuracy predicting in- and out-of-distribution actions while preserving high performance in settings with low or sparse data.

Improving Cross-task Generalization of Unified Table-to-text Models with Compositional Task Configurations
Jifan Chen, Yuhao Zhang, Lan Liu, Rui Dong, Xinchi Chen, Patrick Ng, William Yang Wang and zhiheng huang   19:00-21:00 (Metropolitan East)
There has been great progress in unifying various table-to-text tasks using a single encoder-decoder model trained via multi-task learning (Xie et al., 2022). However, existing methods typically encode task information with a simple dataset name as a prefix to the encoder. This not only limits the effectiveness of multi-task learning, but also hinders the model's ability to generalize to new domains or tasks that were not seen during training, which is crucial for real-world applications. In this paper, we propose compositional task configurations, a set of prompts prepended to the encoder to improve cross-task generalization of unified models. We design the task configurations to explicitly specify the task type, as well as its input and output types. We show that this not only allows the model to better learn shared knowledge across different tasks at training, but also allows us to control the model by composing new configurations that apply novel input-output combinations in a zero-shot manner. We demonstrate via experiments over ten table-to-text tasks that our method outperforms the UnifiedSKG baseline by noticeable margins in both in-domain and zero-shot settings, with average improvements of +0.5 and +12.6 from using a T5-large backbone, respectively.

Optimizing Test-Time Query Representations for Dense Retrieval
Mujeen Sung, Jungsoo Park, Jaewoo Kang, Danqi Chen and Jinhyuk Lee   19:00-21:00 (Metropolitan East)
Recent developments of dense retrieval rely on quality representations of queries and contexts from pre-trained query and context encoders. In this paper, we introduce TOUR (Test-Time Optimization of Query Representations), which further optimizes instance-level query representations guided by signals from test-time retrieval results. We leverage a cross-encoder re-ranker to provide fine-grained pseudo labels over retrieval results and iteratively optimize query representations with gradient descent. Our theoretical analysis reveals that TOUR can be viewed as a generalization of the classical Rocchio algorithm for pseudo relevance feedback, and we present two variants that leverage pseudo-labels as hard binary or soft continuous labels. We first apply TOUR on phrase retrieval with our proposed phrase re-ranker, and also evaluate its effectiveness on passage retrieval with an off-the-shelf re-ranker. TOUR greatly improves end-to-end open-domain question answering accuracy, as well as passage retrieval performance. TOUR also consistently improves direct re-ranking by up to 2.0% while running 1.3–2.4x faster with an efficient implementation.

TimelineQA: A Benchmark for Question Answering over Timelines
Wang-Chiew Tan, Jane Dwivedi-Yu, Yuliang Li, Lambert Mathias, Marzieh Saeidi, Jing Nathan Yan and Alon Y. Halevy   19:00-21:00 (Metropolitan East)
Lifelogs are descriptions of experiences that a person had during their life. Lifelogs are created by fusing data from the multitude of digital services, such as online photos, maps, shopping and content streaming services. Question answering over lifelogs can offer personal assistants a critical resource when they try to provide advice in context. However, obtaining answers to questions over lifelogs is beyond the current state of the art of question answering techniques for a variety of reasons, the most pronounced of which is that lifelogs combine free text with some degree of structure such as temporal and geographical information.

We create and publicly release TimelineQA, a benchmark for accelerating progress on querying lifelogs. TimelineQA generates lifelogs of imaginary people. The episodes in the lifelog range from major life episodes such as high school graduation to those that occur on a daily basis such as going for a run. We describe a set of experiments on TimelineQA with several state-of-the-art QA models. Our experiments reveal that for atomic queries, an extractive QA system significantly out-performs a state-of-the-art retrieval-augmented QA system. For multi-hop queries involving aggregates, we show that the best result is obtained with a state-of-the-art table QA technique, assuming the ground truth set of episodes for deriving the answer is available.

Phrase Retrieval for Open Domain Conversational Question Answering with Conversational Dependency Modeling via Contrastive Learning
Soyeong Jeong, Jinheon Baek, Sung Ju Hwang and Jong Park   19:00-21:00 (Metropolitan East)
Open-Domain Conversational Question Answering (ODConvQA) aims at answering questions through a multi-turn conversation based on a retriever-reader pipeline, which retrieves passages and then predicts answers with them. However, such a pipeline approach not only makes the reader vulnerable to the errors propagated from the retriever, but also demands additional effort to develop both the retriever and the reader, which further makes it slower since they are not runnable in parallel. In this work, we propose a method to directly predict answers with a phrase retrieval scheme for a sequence of words, reducing the conventional two distinct subtasks into a single one. Also, for the first time, we study its capability for ODConvQA tasks. However, simply adopting it is largely problematic, due to the dependencies between previous and current turns in a conversation. To address this problem, we further introduce a novel contrastive learning strategy, making sure to reflect previous turns when retrieving the phrase for the current context, by maximizing representational similarities of consecutive turns in a conver-

sation while minimizing irrelevant conversational contexts. We validate our model on two ODConvQA datasets, whose experimental results show that it substantially outperforms the relevant baselines with the retriever-reader. Code is available at: https://github.com/starsuzi/PRO-ConvQA.

DePlot: One-shot visual language reasoning by plot-to-table translation
Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier and Yasemin Altun   19:00-21:00 (Metropolitan East)
Visual language such as charts and plots is ubiquitous in the human world. Comprehending plots and charts requires strong reasoning skills. Prior state-of-the-art (SOTA) models require at least tens of thousands of training examples and their reasoning capabilities are still much limited, especially on complex human-written queries. This paper presents the first one-shot solution to visual language reasoning. We decompose the challenge of visual language reasoning into two steps: (1) plot-to-text translation, and (2) reasoning over the translated text. The key in this method is a modality conversion module, named as DePlot, which translates the image of a plot or chart to a linearized table. The output of DePlot can then be directly used to prompt a pretrained large language model (LLM), exploiting the few-shot reasoning capabilities of LLMs. To obtain DePlot, we standardize the plot-to-table task by establishing unified task formats and metrics, and train DePlot end-to-end on this task. DePlot can then be used off-the-shelf together with LLMs in a plug-and-play fashion. Compared with a SOTA model finetuned on more than thousands of data points, DePlot+LLM with just one-shot prompting achieves a 29.4% improvement over finetuned SOTA on human-written queries from the task of chart QA.

Coupling Large Language Models with Logic Programming for Robust and General Reasoning from Text
Zhun Yang, Adam Ishay and Joohyung Lee   19:00-21:00 (Metropolitan East)
While large language models (LLMs), such as GPT-3, appear to be robust and general, their reasoning ability is not at a level to compete with the best models trained for specific natural language reasoning problems. In this study, we observe that a large language model can serve as a highly effective few-shot semantic parser. It can convert natural language sentences into a logical form that serves as input for answer set programs, a logic-based declarative knowledge representation formalism. The combination results in a robust and general system that can handle multiple question-answering tasks without requiring retraining for each new task. It only needs a few examples to guide the LLM's adaptation to a specific task, along with reusable ASP knowledge modules that can be applied to multiple tasks. We demonstrate that this method achieves state-of-the-art performance on several NLP benchmarks, including bAbI, StepGame, CLUTRR, and gSCAN. Additionally, it successfully tackles robot planning tasks that an LLM alone fails to solve.

World Models for Math Story Problems
Andreas Opedal, Niklas Stoehr, Abulhair Saparov and Mrinmaya Sachan   19:00-21:00 (Metropolitan East)
Solving math story problems is a complex task for students and NLP models alike, requiring them to understand the world as described in the story and reason over it to compute an answer. Recent years have seen impressive performance on automatically solving these problems with large pre-trained language models and innovative techniques to prompt them. However, it remains unclear if these models possess accurate representations of mathematical concepts. This leads to lack of interpretability and trustworthiness which impedes their usefulness in various applications. In this paper, we consolidate previous work on categorizing and representing math story problems and develop MathWorld, which is a graph-based semantic formalism specific for the domain of math story problems. With MathWorld, we can assign world models to math story problems which represent the situations and actions introduced in the text and their mathematical relationships. We combine math story problems from several existing datasets and annotate a corpus of 1,019 problems and 3,204 logical forms with MathWorld. Using this data, we demonstrate the following use cases of MathWorld: (1) prompting language models with synthetically generated question-answer pairs to probe their reasoning and world modeling abilities, and (2) generating new problems by using the world models as a design space.

The Dangers of trusting Stochastic Parrots: Faithfulness and Trust in Open-domain Conversational Question Answering
Sabrina Chiesurin, Dimitris Dimakopoulos, Marco Antonio Sobrevilla Cabezudo, Arash Eshghi, Ioannis Papaioannou, Verena Rieser and Ioannis Konstas   19:00-21:00 (Metropolitan East)
Large language models are known to produce output which sounds fluent and convincing, but is also often wrong, e.g. "unfaithful" with respect to a rationale as retrieved from a knowledge base. In this paper, we show that task-based systems which exhibit certain advanced linguistic dialog behaviors, such as lexical alignment (repeating what the user said), are in fact preferred and trusted more, whereas other phenomena, such as pronouns and ellipsis are dis-preferred. We use open-domain question answering systems as our test-bed for task based dialog generation and compare several open- and closed-book models. Our results highlight the danger of systems that appear to be trustworthy by parroting user input while providing an unfaithful response.

Expand, Rerank, and Retrieve: Query Reranking for Open-Domain Question Answering
Yung-Sung Chuang, Wei Fang, Shang-Wen Li, Wen-tau Yih and James Glass   19:00-21:00 (Metropolitan East)
We propose EAR, a query Expansion And Reranking approach for improving passage retrieval, with the application to open-domain question answering. EAR first applies a query expansion model to generate a diverse set of queries, and then uses a query reranker to select the ones that could lead to better retrieval results. Motivated by the observation that the best query expansion often is not picked by greedy decoding, EAR trains its reranker to predict the rank orders of the gold passages when issuing the expanded queries to a given retriever. By connecting better the query expansion model and retriever, EAR significantly enhances a traditional sparse retrieval method, BM25. Empirically, EAR improves top-5/20 accuracy by 3-8 and 5-10 points in in-domain and out-of-domain settings, respectively, when compared to a vanilla query expansion model, GAR, and a dense retrieval model, DPR.

Hybrid Hierarchical Retrieval for Open-Domain Question Answering
Manoj Ghuhan Arivazhagan, Lan Liu, Peng Qi, Xinchi Chen, William Yang Wang and zhiheng huang   19:00-21:00 (Metropolitan East)
Retrieval accuracy is crucial to the performance of open-domain question answering (ODQA) systems. Recent work has demonstrated that dense hierarchical retrieval (DHR), which retrieves document candidates first and then relevant passages from the refined document set, can significantly outperform the single stage dense passage retriever (DPR). While effective, this approach requires document structure information to learn document representation and is hard to adopt to other domains without this information. Additionally, the dense retrievers tend to generalize poorly on out-of-domain data comparing with sparse retrievers such as BM25. In this paper, we propose Hybrid Hierarchical Retrieval (HHR) to address the existing limitations. Instead of relying solely on dense retrievers, we can apply sparse retriever, dense retriever, and a combination of them in both stages of document and passage retrieval. We perform extensive experiments on ODQA benchmarks and observe that our framework not only brings in-domain gains, but also generalizes better to zero-shot TriviaQA and Web Questions datasets with an average of 4.69% improvement on recall@100 over DHR. We also offer practical insights to trade off between retrieval accuracy, latency, and storage cost. The code is available on github.

RobustQA: Benchmarking the Robustness of Domain Adaptation for Open-Domain Question Answering
Rujun Han, Peng Qi, Yuhao Zhang, Lan Liu, Juliette Burger, William Yang Wang, zhiheng huang, Bing Xiang and Dan Roth   19:00-21:00 (Metropolitan East)
Open-domain question answering (ODQA) is a crucial task in natural language processing. A typical ODQA system relies on a retriever mod-

ule to select relevant contexts from a large corpus for a downstream reading comprehension model. Existing ODQA datasets consist mainly of Wikipedia corpus, and are insufficient to study models' generalizability across diverse domains as models are trained and evaluated on the same genre of data. We propose **RobustQA**, a novel benchmark consisting of datasets from 8 different domains, which facilitates the evaluation of ODQA's domain robustness. To build **RobustQA**, we annotate QA pairs in retrieval datasets with rigorous quality control. We further examine improving QA performances by incorporating unsupervised learning methods with target-domain corpus and adopting large generative language models. These methods can effectively improve model performances on **RobustQA**. However, experimental results demonstrate a significant gap from in-domain training, suggesting that **RobustQA** is a challenging benchmark to evaluate ODQA domain robustness.

KoRC: Knowledge Oriented Reading Comprehension Benchmark for Deep Text Understanding
Zijun Yao, Yantao Liu, Xin Lv, Shulin Cao, Jifan Yu, Juanzi Li and Lei Hou   19:00-21:00 (Metropolitan East)
Deep text understanding, which requires the connections between a given document and prior knowledge beyond its text, has been highlighted by many benchmarks in recent years. However, these benchmarks have encountered two major limitations. On the one hand, most of them require human annotation of knowledge, which leads to limited knowledge coverage. On the other hand, they usually use choices or spans in the texts as the answers, which results in narrow answer space. To overcome these limitations, we build a new challenging benchmark named KoRC in this paper. Compared with previous benchmarks, KoRC has two advantages, i.e., broad knowledge coverage and flexible answer format. Specifically, we utilize massive knowledge bases to guide annotators or large language models (LLMs) to construct knowledgable questions. Moreover, we use labels in knowledge bases rather than spans or choices as the final answers. We test state-of-the-art models on KoRC and the experimental results show that the strongest baseline only achieves 68.3% and 30.0% F1 measure in the IID and OOD test set, respectively. These results indicate that deep text understanding is still an unsolved challenge. We will release our dataset and baseline methods upon acceptance.

Distinguish Before Answer: Generating Contrastive Explanation as Knowledge for Commonsense Question Answering
Qianglong Chen, Guohai Xu, Ming Yan, Ji Zhang, Fei Huang, Luo Si and Yin Zhang   19:00-21:00 (Metropolitan East)
Existing knowledge-enhanced methods have achieved remarkable results in certain Q&A tasks via obtaining diverse knowledge from different knowledge bases. However, limited by the properties of retrieved knowledge, they still have trouble benefiting from both the knowledge relevance and distinguishment simultaneously. To address the challenge, we propose **CPACE**, a **C**oncept-centric **P**rompt-b**A**sed **C**ontrastive **E**xplanation Generation model, which aims to convert obtained symbolic knowledge into the contrastive explanation for better distinguishing the differences among given candidates. Firstly, following previous works, we retrieve different types of symbolic knowledge with a concept-centric knowledge extraction module. After that, we generate corresponding contrastive explanation using acquired symbolic knowledge and prompt as guidance for better modeling the knowledge distinguishment and interpretability. Finally, we regard the generated contrastive explanation as external knowledge for downstream task enhancement. We conduct a series of experiments on three widely-used question-answering datasets: CSQA, QASC, and OBQA. Experimental results demonstrate that with the help of generated contrastive explanation, our CPACE model achieves new SOTA on CSQA (89.8% on the testing set, 0.9% higher than human performance), and gains impressive improvement on QASC and OBQA (4.2% and 3.5%, respectively).

An Empirical Comparison of LM-based Question and Answer Generation Methods
Asahi Ushio, Fernando Alva-Manchego and Jose Camacho-Collados   19:00-21:00 (Metropolitan East)
Question and answer generation (QAG) consists of generating a set of question-answer pairs given a context (e.g. a paragraph). This task has a variety of applications, such as data augmentation for question answering (QA) models, information retrieval and education. In this paper, we establish baselines with three different QAG methodologies that leverage sequence-to-sequence language model (LM) fine-tuning. Experiments show that an end-to-end QAG model, which is computationally light at both training and inference times, is generally robust and outperforms other more convoluted approaches. However, there are differences depending on the underlying generative LM. Finally, our analysis shows that QA models fine-tuned solely on generated question-answer pairs can be competitive when compared to supervised QA models trained on human-labeled data.

Don't Lose Yourself! Empathetic Response Generation via Explicit Self-Other Awareness
Weixiang Zhao, Yanyan Zhao, Xin Lu and Bing Qin   19:00-21:00 (Metropolitan East)
As a critical step to achieve human-like chatbots, empathetic response generation has attained increasing interests. Previous attempts are incomplete and not sufficient enough to elicit empathy because they only stay on the initial stage of empathy to automatically sense and simulate the feelings and thoughts of others via other-awareness. However, they ignore to include self-awareness to consider the own views of the self in their responses, which is a crucial process to achieve the empathy. To this end, we propose to generate Empathetic response with explicit Self-Other Awareness (EmpSOA). Specifically, three stages, self-other differentiation, self-other modulation and self-other generation, are devised to clearly maintain, regulate and inject the self-other aware information into the process of empathetic response generation. Both automatic and human evaluations on the benchmark dataset demonstrate the superiority of EmpSOA to generate more empathetic responses. Our source code will be publicly available.

A Unified One-Step Solution for Aspect Sentiment Quad Prediction
Junxian Zhou, Haiqin Yang, Yuxuan He, Hao Mou and JunBo Yang   19:00-21:00 (Metropolitan East)
Aspect sentiment quad prediction (ASQP) is a challenging yet significant subtask in aspectbased sentiment analysis as it provides a complete aspect-level sentiment structure. However, existing ASQP datasets are usually small and low-density, hindering technical advancement. To expand the capacity, in this paper, we release two new datasets for ASQP, which contain the following characteristics: larger size, more words per sample, and higher density. With such datasets, we unveil the shortcomings of existing strong ASQP baselines and therefore propose a unified one-step solution for ASQP, namely One-ASQP, to detect the aspect categories and to identify the aspectopinion-sentiment (AOS) triplets simultaneously. Our One-ASQP holds several unique advantages: (1) by separating ASQP into two subtasks and solving them independently and simultaneously, we can avoid error propagation in pipeline-based methods and overcome slow training and inference in generation-based methods; (2) by introducing sentiment-specific horns tagging schema in a token-pair-based two-dimensional matrix, we can exploit deeper interactions between sentiment elements and efficiently decode the AOS triplets; (3) we design "[NULL]" token can help us effectively identify the implicit aspects or opinions. Experiments on two benchmark datasets and our released two datasets demonstrate the advantages of our One-ASQP. The two new datasets are publicly released at https://www.github.com/Datastory-CN/ASQP-Datasets.

Few-shot Joint Multimodal Aspect-Sentiment Analysis Based on Generative Multimodal Prompt
Xiaocui Yang, Shi Feng, Daling Wang, Qi Sun, Wenfang Wu, Yifei Zhang, Pengfei Hong and Soujanya Poria 19:00-21:00 (Metropolitan East)
We have witnessed the rapid proliferation of multimodal data on numerous social media platforms. Conventional studies typically require massive labeled data to train models for Multimodal Aspect-Based Sentiment Analysis (MABSA). However, collecting and annotating finegrained multimodal data for MABSA is tough. To alleviate the above issue, we perform three MABSA-related tasks with quite a small number of labeled multimodal samples. We first build diverse and comprehensive multimodal few-shot datasets according to the data distribution. To capture the specific prompt for each aspect term in a few-shot scenario, we propose a novel Generative Multimodal Prompt (GMP) model for MABSA, which includes the Multimodal Encoder module and the N-Stream Decoders module. We further introduce a subtask to predict

the number of aspect terms in each instance to construct the multimodal prompt. Extensive experiments on two datasets demonstrate that our approach outperforms strong baselines on two MABSA-related tasks in the few-shot setting.

A Simple Yet Strong Domain-Agnostic De-bias Method for Zero-Shot Sentiment Classification

Yang Zhao, Tetsuya Nasukawa, Masayasu Muraoka and Bishwaranjan Bhattacharjee   19:00-21:00 (Metropolitan East)

Zero-shot prompt-based learning has made much progress in sentiment analysis, and considerable effort has been dedicated to designing high-performing prompt templates. However, two problems exist; First, large language models are often biased to their pre-training data, leading to poor performance in prompt templates that models have rarely seen. Second, in order to adapt to different domains, re-designing prompt templates is usually required, which is time-consuming and inefficient. To remedy both shortcomings, we propose a simple yet strong data construction method to de-bias a given prompt template, yielding a large performance improvement in sentiment analysis tasks across different domains, pre-trained language models, and prompt templates. Also, we demonstrate the advantage of using domain-agnostic generic responses over the in-domain ground-truth data.

TransESC: Smoothing Emotional Support Conversation via Turn-Level State Transition

Weixiang Zhao, Yanyan Zhao, Shilong Wang and Bing Qin   19:00-21:00 (Metropolitan East)

Emotion Support Conversation (ESC) is an emerging and challenging task with the goal of reducing the emotional distress of people. Previous attempts fail to maintain smooth transitions between utterances in ESC because they ignoring to grasp the fine-grained transition information at each dialogue turn. To solve this problem, we propose to take into account turn-level state Transitions of ESC (TransESC) from three perspectives, including semantics transition, strategy transition and emotion transition, to drive the conversation in a smooth and natural way. Specifically, we construct the state transition graph with a two-step way, named transit-then-interact, to grasp such three types of turn-level transition information. Finally, they are injected into the transition aware decoder to generate more engaging responses. Both automatic and human evaluations on the benchmark dataset demonstrate the superiority of TransESC to generate more smooth and effective supportive responses. Our source code will be publicly available.

Multilingual Multi-Figurative Language Detection

Huiyuan Lai, Antonio Toral and Malvina Nissim   19:00-21:00 (Metropolitan East)

Figures of speech help people express abstract concepts and evoke stronger emotions than literal expressions, thereby making texts more creative and engaging. Due to its pervasive and fundamental character, figurative language understanding has been addressed in Natural Language Processing, but it's highly understudied in a multilingual setting and when considering more than one figure of speech at the same time. To bridge this gap, we introduce multilingual multi-figurative language modelling, and provide a benchmark for sentence-level figurative language detection, covering three common figures of speech and seven languages. Specifically, we develop a framework for figurative language detection based on template-based prompt learning. In so doing, we unify multiple detection tasks that are interrelated across multiple figures of speech and languages, without requiring task- or language-specific modules. Experimental results show that our framework outperforms several strong baselines and may serve as a blueprint for the joint modelling of other interrelated tasks.

Counterfactual Probing for the Influence of Affect and Specificity on Intergroup Bias

Venkata Subrahmanyan Govindarajan, David I Beaver, Kyle Mahowald and Junyi Jessy Li   19:00-21:00 (Metropolitan East)

While existing work on studying bias in NLP focues on negative or pejorative language use, Govindarajan et al. (2023) offer a revised framing of bias in terms of intergroup social context, and its effects on language behavior. In this paper, we investigate if two pragmatic features (specificity and affect) systematically vary in different intergroup contexts — thus connecting this new framing of bias to language output. Preliminary analysis finds modest correlations between specificity and affect of tweets with supervised intergroup relationship (IGR) labels. Counterfactual probing further reveals that while neural models finetuned for predicting IGR reliably use affect in classification, the model's usage of specificity is inconclusive.

Measuring Intersectional Biases in Historical Documents

Nadav Borenstein, Karolina Stanczak, Thea Rolskov, Natacha Klein Käfer, Natália da Silva Perez and Isabelle Augenstein   19:00-21:00 (Metropolitan East)

Data-driven analyses of biases in historical texts can help illuminate the origin and development of biases prevailing in modern society. However, digitised historical documents pose a challenge for NLP practitioners as these corpora suffer from errors introduced by optical character recognition (OCR) and are written in an archaic language. In this paper, we investigate the continuities and transformations of bias in historical newspapers published in the Caribbean during the colonial era (18th to 19th centuries). Our analyses are performed along the axes of gender, race, and their intersection. We examine these biases by conducting a temporal study in which we measure the development of lexical associations using distributional semantics models and word embeddings. Further, we evaluate the effectiveness of techniques designed to process OCR-generated data and assess their stability when trained on and applied to the noisy historical newspapers. We find that there is a trade-off between the stability of the word embeddings and their compatibility with the historical dataset. We provide evidence that gender and racial biases are interdependent, and their intersection triggers distinct effects. These findings align with the theory of intersectionality, which stresses that biases affecting people with multiple marginalised identities compound to more than the sum of their constituents.

It's not Sexually Suggestive; It's Educative | Separating Sex Education from Suggestive Content on TikTok videos

Enfa Rose George and Mihai Surdeanu   19:00-21:00 (Metropolitan East)

We introduce SexTok, a multi-modal dataset composed of TikTok videos labeled as sexually suggestive (from the annotator's point of view), sex-educational content, or neither. Such a dataset is necessary to address the challenge of distinguishing between sexually suggestive content and virtual sex education videos on TikTok. Children's exposure to sexually suggestive videos has been shown to have adversarial effects on their development (Collins et al. 2017). Meanwhile, virtual sex education, especially on subjects that are more relevant to the LGBTQIA+ community, is very valuable (Mitchell et al. 2014). The platform's current system removes/punishes some of both types of videos, even though they serve different purposes. Our dataset contains video URLs, and it is also audio transcribed. To validate its importance, we explore two transformer-based models for classifying the videos. Our preliminary results suggest that the task of distinguishing between these two types of videos is learnable but challenging. These experiments suggest that this dataset is meaningful and invites further study on the subject.

Contrastive Learning of Sociopragmatic Meaning in Social Media

Chiyu Zhang, Muhammad Abdul-Mageed and Ganesh Jawahar   19:00-21:00 (Metropolitan East)

Recent progress in representation and contrastive learning in NLP has not widely considered the class of sociopragmatic meaning (i.e., meaning in interaction within different language communities). To bridge this gap, we propose a novel framework for learning task-agnostic representations transferable to a wide range of sociopragmatic tasks (e.g., emotion, hate speech, humor, sarcasm). Our framework outperforms other contrastive learning frameworks for both in-domain and out-of-domain data, across both the general and few-shot settings. For example, compared to two popular pre-trained language models, our model obtains an improvement of 11.66 average F1 on 16 datasets when fine-tuned on only 20 training samples per dataset. We also show that our framework improves uniformity and preserves the semantic structure of representations. Our code is available at: https://github.com/UBC-NLP/infodcl

Words as Gatekeepers: Measuring Discipline-specific Terms and Meanings in Scholarly Publications
Li Lucy, Jesse Dodge, David Bamman and Katherine A Keith 19:00-21:00 (Metropolitan East)
Scholarly text is often laden with jargon, or specialized language that can facilitate efficient in-group communication within fields but hinder understanding for out-groups. In this work, we develop and validate an interpretable approach for measuring scholarly jargon from text. Expanding the scope of prior work which focuses on word types, we use word sense induction to also identify words that are widespread but overloaded with different meanings across fields. We then estimate the prevalence of these discipline-specific words and senses across hundreds of subfields, and show that word senses provide a complementary, yet unique view of jargon alongside word types. We demonstrate the utility of our metrics for science of science and computational sociolinguistics by highlighting two key social implications. First, though most fields reduce their use of jargon when writing for general-purpose venues, and some fields (e.g., biological sciences) do so less than others. Second, the direction of correlation between jargon and citation rates varies among fields, but jargon is nearly always negatively correlated with interdisciplinary impact. Broadly, our findings suggest that though multidisciplinary venues intend to cater to more general audiences, some fields' writing norms may act as barriers rather than bridges, and thus impede the dispersion of scholarly ideas.

Modeling Cross-Cultural Pragmatic Inference with Codenames Duet
Omar Shaikh, Caleb Ziems, William Held, Aryan J Pariani, Fred Morstatter and Diyi Yang 19:00-21:00 (Metropolitan East)
Pragmatic reference enables efficient interpersonal communication. Prior work uses simple reference games to test models of pragmatic reasoning, often with unidentified speakers and listeners. In practice, however, speakers' sociocultural background shapes their pragmatic assumptions. For example, readers of this paper assume NLP refers to Natural Language Processing, and not "Neuro-linguistic Programming." This work introduces the Cultural Codes dataset, which operationalizes sociocultural pragmatic inference in a simple word reference game. Cultural Codes is based on the multi-turn collaborative two-player game, Codenames Duet. Our dataset consists of 794 games with 7,703 turns, distributed across 153 unique players. Alongside gameplay, we collect information about players' personalities, values, and demographics. Utilizing theories of communication and pragmatics, we predict each player's actions via joint modeling of their sociocultural priors and the game context. Our experiments show that accounting for background characteristics significantly improves model performance for tasks related to both clue-giving and guessing, indicating that sociocultural priors play a vital role in gameplay decisions.

Causal Matching with Text Embeddings: A Case Study in Estimating the Causal Effects of Peer Review Policies
Raymond Zhang, Neha Nayak Kennard, Daniel S Smith, Daniel A McFarland, Andrew McCallum and Katherine A Keith 19:00-21:00 (Metropolitan East)
A promising approach to estimate the causal effects of peer review policies is to analyze data from publication venues that shift policies from single-blind to double-blind from one year to the next. However, in these settings the content of the manuscript is a confounding variable—each year has a different distribution of scientific content which may naturally affect the distribution of reviewer scores. To address this textual confounding, we extend variable ratio nearest neighbor matching to incorporate text embeddings. We compare this matching method to a widely-used causal method of stratified propensity score matching and a baseline of randomly selected matches. For our case study of the ICLR conference shifting from single- to double-blind review from 2017 to 2018, we find human judges prefer manuscript matches from our method in 70% of cases. While the unadjusted estimate of the average causal effect of reviewers' scores is -0.25, our method shifts the estimate to -0.17, a slightly smaller difference between the outcomes of single- and double-blind policies. We hope this case study enables exploration of additional text-based causal estimation methods and domains in the future.

Dramatic Conversation Disentanglement
Kent K. Chang, Danica Chen and David Bamman 19:00-21:00 (Metropolitan East)
We present a new dataset for studying conversation disentanglement in movies and TV series. While previous work has focused on conversation disentanglement in IRC chatroom dialogues, movies and TV shows provide a space for studying complex pragmatic patterns of floor and topic change in face-to-face multi-party interactions. In this work, we draw on theoretical research in sociolinguistics, sociology, and film studies to operationalize a conversational thread (including the notion of a floor change) in dramatic texts, and use that definition to annotate a dataset of 10,033 dialogue turns (comprising 2,209 threads) from 831 movies. We compare the performance of several disentanglement models on this dramatic dataset, and apply the best-performing model to disentangle 808 movies. We see that, contrary to expectation, average thread lengths do not decrease significantly over the past 40 years, and characters portrayed by actors who are women, while underrepresented, initiate more new conversational threads relative to their speaking time.

Responsibility Perspective Transfer for Italian Femicide News
Gosse Minnema, Huiyuan Lai, Benedetta Muscato and Malvina Nissim 19:00-21:00 (Metropolitan East)
Different ways of linguistically expressing the same real-world event can lead to different perceptions of what happened. Previous work has shown that different descriptions of gender-based violence (GBV) influence the reader's perception of who is to blame for the violence, possibly reinforcing stereotypes which see the victim as partly responsible, too. As a contribution to raise awareness on perspective-based writing, and to facilitate access to alternative perspectives, we introduce the novel task of automatically rewriting GBV descriptions as a means to alter the perceived level of blame on the perpetrator. We present a quasi-parallel dataset of sentences with low and high perceived responsibility levels for the perpetrator, and experiment with unsupervised (mBART-based), zero-shot and few-shot (GPT3-based) methods for rewriting sentences. We evaluate our models using a questionnaire study and a suite of automatic metrics.

On Text-based Personality Computing: Challenges and Future Directions
Qixiang Fang, Anastasia Giachanou, Ayoub Bagheri, Laura Boeschoten, Erik-Jan van Kesteren, Mahdi Shafiee Kamalabad and Daniel Oberski 19:00-21:00 (Metropolitan East)
Text-based personality computing (TPC) has gained many research interests in NLP. In this paper, we describe 15 challenges that we consider deserving the attention of the NLP research community. These challenges are organized by the following topics: personality taxonomies, measurement quality, datasets, performance evaluation, modelling choices, as well as ethics and fairness. When addressing each challenge, not only do we combine perspectives from both NLP and social sciences, but also offer concrete suggestions. We hope to inspire more valid and reliable TPC research.

Findings Spotlights II

19:00-21:00 (Metropolitan Centre)

On the Expressivity Role of LayerNorm in Transformers' Attention
Shaked Brody, Uri Alon and Eran Yahav 19:00-21:00 (Metropolitan Centre)
Layer Normalization (LayerNorm) is an inherent component in all Transformer-based models. In this paper, we show that LayerNorm is crucial to the expressivity of the multi-head attention layer that follows it. This is in contrast to the common belief that LayerNorm's only role

is to normalize the activations during the forward pass, and their gradients during the backward pass.
We consider a geometric interpretation of LayerNorm and show that it consists of two components: (a) projection of the input vectors to a d-1 space that is orthogonal to the [1,1,...,1] vector, and (b) scaling of all vectors to the same norm of root d. We show that each of these components is important for the attention layer that follows it in Transformers: (a) projection allows the attention mechanism to create an attention query that attends to all keys equally, offloading the need to learn this operation in the attention; and (b) scaling allows each key to potentially receive the highest attention, and prevents keys from being "un-select-able". We show empirically that Transformers do indeed benefit from these properties of LayeNorm in general language modeling and even in computing simple functions such as "majority".

EmbedTextNet: Dimension Reduction with Weighted Reconstruction and Correlation Losses for Efficient Text Embedding
Dae Yon Hwang, Bilal Taha and Yaroslav Nechaev   19:00-21:00 (Metropolitan Centre)
The size of embeddings generated by large language models can negatively affect system latency and model size in certain downstream practical applications (e.g. KNN search). In this work, we propose EmbedTextNet, a light add-on network that can be appended to an arbitrary language model to generate a compact embedding without requiring any changes in its architecture or training procedure. Specifically, we use a correlation penalty added to the weighted reconstruction loss that better captures the informative features in the text embeddings, which improves the efficiency of the language models. We evaluated EmbedTextNet on three different downstream tasks: text similarity, language modelling, and text retrieval. Empirical results on diverse benchmark datasets demonstrate the effectiveness and superiority of EmbedTextNet compared to state-of-art methodologies in recent works, especially in extremely low dimensional embedding sizes. The developed code for reproducibility is included in the supplementary material.

A Memory Model for Question Answering from Streaming Data Supported by Rehearsal and Anticipation of Coreference Information
Vladimir Araujo, Alvaro M Soto and Marie-Francine Moens   19:00-21:00 (Metropolitan Centre)
Existing question answering methods often assume that the input content (e.g., documents or videos) is always available to solve the task. Alternatively, memory networks were introduced to mimic the human process of incremental comprehension and compression of the information in a fixed-capacity memory. However, these models only learn how to maintain memory by backpropagating errors in the answers through the entire network. Instead, it has been suggested that humans have effective mechanisms to boost their memorization capacities, such as rehearsal and anticipation. Drawing inspiration from these, we propose a memory model that performs rehearsal and anticipation while processing inputs to memorize important information for solving question answering tasks from streaming data. The proposed mechanisms are applied self-supervised during training through masked modeling tasks focused on coreference information. We validate our model on a short-sequence (bAbI) dataset as well as large-sequence textual (NarrativeQA) and video (ActivityNet-QA) question answering datasets, where it achieves substantial improvements over previous memory network approaches. Furthermore, our ablation study confirms the proposed mechanisms' importance for memory models.

CFL: Causally Fair Language Models Through Token-level Attribute Controlled Generation
Rahul Madhavan, Rishabh Garg, Kahini Wadhawan and Sameep Mehta   19:00-21:00 (Metropolitan Centre)
We propose a method to control the attributes of Language Models (LMs) for the text generation task using Causal Average Treatment Effect (ATE) scores and counterfactual augmentation. We explore this method, in the context of LM detoxification, and propose the Causally Fair Language (CFL) architecture for detoxifying pre-trained LMs in a plug-and-play manner. Our architecture is based on a Structural Causal Model (SCM) that is mathematically transparent and computationally efficient as compared with many existing detoxification techniques. We also propose several new metrics that aim to better understand the behaviour of LMs in the context of toxic text generation. Further, we achieve state of the art performance for toxic degeneration, which are computed using Real Toxicity Prompts. Our experiments show that CFL achieves such a detoxification without much impact on the model perplexity. We also show that CFL mitigates the unintended bias problem through experiments on the BOLD dataset.

Text Augmentation Using Dataset Reconstruction for Low-Resource Classification
Adir Rahamim, Guy Uziel, Esther Goldbraich and Ateret Anaby Tavor   19:00-21:00 (Metropolitan Centre)
In the deployment of real-world text classification models, label scarcity is a common problem and as the number of classes increases, this problem becomes even more complex. An approach to addressing this problem is by applying text augmentation methods.
One of the more prominent methods involves using the text-generation capabilities of language models. In this paper, we propose Text AUgmentation by Dataset Reconstruction (TAU-DR), a novel method of data augmentation for text classification. We conduct experiments on several multi-class datasets, showing that our approach improves the current state-of-the-art techniques for data augmentation.

Low-Rank Updates of pre-trained Weights for Multi-Task Learning
Alexandre Daniel AUDIBERT, Massih R Amini, Konstantin Usevich and Marianne Clausel   19:00-21:00 (Metropolitan Centre)
Multi-Task Learning used with pre-trained models has been quite popular in the field of Natural Language Processing in recent years. This framework remains still challenging due to the specificity of the tasks and the challenges associated with fine-tuning large pre-trained models. In this paper, we propose a new approach for Multi-task learning which is based on stacking the weights of Neural Networks as a tensor. We show that low-rank updates in the canonical polyadic tensor decomposition of this tensor of weights lead to a simple, yet efficient algorithm, which without loss of performance allows to reduce considerably the model parameters. We investigate the interactions between tasks inside the model as well as the inclusion of sparsity to find the best tensor rank and to increase the compression rate. Our strategy is consistent with recent efforts that attempt to use constraints to fine-tune some model components. More precisely, we achieve equivalent performance as the state-of-the-art on the General Language Understanding Evaluation benchmark by training only 0.3 of the parameters per task while not modifying the baseline weights.

AutoMoE: Heterogeneous Mixture-of-Experts with Adaptive Computation for Efficient Neural Machine Translation
Ganesh Jawahar, Subhabrata Mukherjee, Xiaodong Liu, Young Jin Kim, Muhammad Abdul-Mageed, Laks Lakshmanan, V.S., Ahmed Hassan Awadallah, Sebastien Bubeck and Jianfeng Gao   19:00-21:00 (Metropolitan Centre)
Mixture-of-Expert (MoE) models have obtained state-of-the-art performance in Neural Machine Translation (NMT) tasks. Existing works in MoE mostly consider a homogeneous design where the same number of experts of the same size are placed uniformly throughout the network. Furthermore, existing MoE works do not consider computational constraints (e.g., FLOPs, latency) to guide their design. To this end, we develop AutoMoE – a framework for designing heterogeneous MoE's under computational constraints. AutoMoE leverages Neural Architecture Search (NAS) to obtain efficient sparse MoE sub-transformers with 4x inference speedup (CPU) and FLOPs reduction over manually designed Transformers, with parity in BLEU score over dense Transformer and within 1 BLEU point of MoE SwitchTransformer, on aggregate over benchmark datasets for NMT. Heterogeneous search space with dense and sparsely activated Transformer modules (e.g., how many experts? where to place them? what should be their sizes?) allows for adaptive compute – where different amounts of computations are used for different tokens in the input. Adaptivity comes naturally from routing decisions which send tokens to experts of different sizes. AutoMoE code, data, and trained models are available at https://aka.ms/AutoMoE.

Neural Architecture Search for Parameter-Efficient Fine-tuning of Large Pre-trained Language Models

Neal G Lawton, Anoop Kumar, Govind Thattai, Aram Galstyan and Greg Ver Steeg   19:00-21:00 (Metropolitan Centre)
Parameter-efficient tuning (PET) methods fit pre-trained language models (PLMs) to downstream tasks by either computing a small compressed update for a subset of model parameters, or appending and fine-tuning a small number of new model parameters to the pre-trained network. Hand-designed PET architectures from the literature perform well in practice, but have the potential to be improved via automated neural architecture search (NAS). We propose an efficient NAS method for learning PET architectures via structured and unstructured pruning. We present experiments on GLUE demonstrating the effectiveness of our algorithm and discuss how PET architectural design choices affect performance in practice.

Which Examples Should be Multiply Annotated? Active Learning When Annotators May Disagree
Connor T Baumler, Anna Sotnikova and Hal Daumé III   19:00-21:00 (Metropolitan Centre)
Linguistic annotations, especially for controversial topics like hate speech detection, are frequently contested due to annotator backgrounds and positionalities. In such situations, preserving this disagreement through the machine learning pipeline can be important for downstream use cases. However, capturing disagreement can increase annotation time and expense. Fortunately, for many tasks, not all examples are equally controversial; we develop an active learning approach, Disagreement Aware Active Learning (DAAL) that concentrates annotations on examples where model entropy and annotator entropy are the most different. Because we cannot know the true entropy of annotations on unlabeled examples, we estimate a model that predicts annotator entropy trained using very few multiply-labeled examples. We find that traditional uncertainty-based active learning underperforms simple passive learning on tasks with high levels of disagreement, but that our active learning approach is able to successfully improve on passive and active baselines, reducing the number of annotations required by at least 24% on average across several datasets.

B2T Connection: Serving Stability and Performance in Deep Transformers
Sho Takase, Shun Kiyono, Sosuke Kobayashi and Jun Suzuki   19:00-21:00 (Metropolitan Centre)
In the perspective of a layer normalization (LN) position, the architecture of Transformers can be categorized into two types: Post-LN and Pre-LN. Recent Transformers prefer to select Pre-LN because the training in Post-LN with deep Transformers, e.g., ten or more layers, often becomes unstable, resulting in useless models. However, in contrast, Post-LN has also consistently achieved better performance than Pre-LN in relatively shallow Transformers, e.g., six or fewer layers. This study first investigates the reason for these discrepant observations empirically and theoretically and discovers 1, the LN in Post-LN is the source of the vanishing gradient problem that mainly leads the unstable training whereas Pre-LN prevents it, and 2, Post-LN tends to preserve larger gradient norms in higher layers during the back-propagation that may lead an effective training. Exploiting the new findings, we propose a method that can equip both higher stability and effective training by a simple modification from Post-LN. We conduct experiments on a wide range of text generation tasks and demonstrate that our method outperforms Pre-LN, and stable training regardless of the shallow or deep layer settings.

Reinforced Active Learning for Low-Resource, Domain-Specific, Multi-Label Text Classification
Lukas Wertz, Jasmina Bogojeska, Katsiaryna Mirylenka and Jonas Kuhn   19:00-21:00 (Metropolitan Centre)
Text classification datasets from specialised or technical domains are in high demand, especially in industrial applications. However, due to the high cost of annotation such datasets are usually expensive to create. While Active Learning (AL) can reduce the labeling cost, required AL strategies are often only tested on general knowledge domains and tend to use information sources that are not consistent across tasks. We propose Reinforced Active Learning (RAL) to train a Reinforcement Learning policy that utilizes many different aspects of the data and the task in order to select the most informative unlabeled subset dynamically over the course of the AL procedure. We demonstrate the superior performance of the proposed RAL framework compared to strong AL baselines across four intricate multi-class, multi-label text classification datasets taken from specialised domains. In addition, we experiment with a unique data augmentation approach to further reduce the number of samples RAL needs to annotate.

PreQuant: A Task-agnostic Quantization Approach for Pre-trained Language Models
Zhuocheng Gong, Jiahao Liu, Qifan Wang, Yang Yang, Jingang Wang, Wei Wu, Yunsen Xian, Dongyan Zhao and Rui Yan   19:00-21:00 (Metropolitan Centre)
While transformer-based pre-trained language models (PLMs) have dominated a number of NLP applications, these models are heavy to deploy and expensive to use. Therefore, effectively compressing large-scale PLMs becomes an increasingly important problem. Quantization, which represents high-precision tensors with low-bit fix-point format, is a viable solution. However, most existing quantization methods are task-specific, requiring customized training and quantization with a large number of trainable parameters on each individual task. Inspired by the observation that the over-parameterization nature of PLMs makes it possible to freeze most of the parameters during the fine-tuning stage, in this work, we propose a novel "quantize before fine-tuning" framework, PreQuant, that differs from both quantization-aware training and post-training quantization. {pasted macro 'OUR'} is compatible with various quantization strategies, with outlier-aware parameter-efficient fine-tuning incorporated to correct the induced quantization error. We demonstrate the effectiveness of PreQuant on the GLUE benchmark using BERT, RoBERTa, and T5. We also provide an empirical investigation into the workflow of PreQuant, which sheds light on its efficacy.

Know Where You're Going: Meta-Learning for Parameter-Efficient Fine-Tuning
Mozhdeh Gheini, Xuezhe Ma and Jonathan May   19:00-21:00 (Metropolitan Centre)
A recent family of techniques, dubbed lightweight fine-tuning methods, facilitates parameter-efficient transfer by updating only a small set of additional parameters while keeping the parameters of the original model frozen. While proven to be an effective approach, there are no existing studies on if and how such knowledge of the downstream fine-tuning approach calls for complementary measures after pre-training and before fine-tuning. In this work, we show that taking the unique choice of fine-tuning into consideration boosts the performance of parameter-efficient fine-tuning. By relying on optimization-based meta-learning using MAML with certain modifications for our distinct purpose, we prime the pre-trained model specifically for parameter-efficient fine-tuning, resulting in gains of up to 4.96 points on cross-lingual NER fine-tuning. Our ablation settings and analyses further reveal that the specific approach we take to meta-learning is crucial for the attained gains.

Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes
Cheng-Yu Hsieh, Chun-Liang Li, CHIH-KUAN YEH, Hootan Nakhost, Yasuhisa Fujii, Alex Jason Ratner, Ranjay Krishna, Chen-Yu Lee and Tomas Pfister   19:00-21:00 (Metropolitan Centre)
Deploying large language models (LLMs) is challenging because they are memory inefficient and compute-intensive for practical applications. In reaction, researchers train smaller task-specific models by either finetuning with human labels or distilling using LLM-generated labels. However, finetuning and distillation require large amounts of training data to achieve comparable performance to LLMs. We introduce Distilling step-by-step, a new mechanism that (a) trains smaller models that outperform LLMs, and (b) achieves so by leveraging less training data needed by finetuning or distillation. Our method extracts LLM rationales as additional supervision for training small models within a multi-task framework. We present three findings across 4 NLP benchmarks: First, compared to both finetuning and distillation, our mechanism achieves better performance with much fewer labeled/unlabeled training examples. Second, compared to few-shot prompted LLMs, we achieve better performance using substantially smaller model sizes. Third, we reduce both the model size and the amount of data required to outperform LLMs; our finetuned 770M T5 model outperforms the few-shot prompted 540B PaLM model using only 80% of available data

on a benchmark, whereas standard finetuning the same T5 model struggles to match even by using 100% of the dataset.

LABO: Towards Learning Optimal Label Regularization via Bi-level Optimization
Peng Lu, Ahmad Rashid, Ivan Kobyzev, Mehdi Rezagholizadeh and Phillippe Langlais   19:00-21:00 (Metropolitan Centre)
Regularization techniques are crucial to improving the generalization performance and training efficiency of deep neural networks. Many deep learning algorithms rely on weight decay, dropout, batch/layer normalization to converge faster and generalize. Label Smoothing (LS) is another simple, versatile and efficient regularization which can be applied to various supervised classification tasks. Conventional LS, however, regardless of the training instance assumes that each non-target class is equally likely. In this work, we present a general framework for training with label regularization, which includes conventional LS but can also model instance-specific variants. Based on this formulation, we propose an efficient way of learning LAbel regularization by devising a Bi-level Optimization (LABO) problem. We derive a deterministic and interpretable solution of the inner loop as the optimal label smoothing without the need to store the parameters or the output of a trained model. Finally, we conduct extensive experiments and demonstrate our LABO consistently yields improvement over conventional label regularization on various fields, including seven machine translation and three image classification tasks across various neural network architectures while maintaining training efficiency.

Not Enough Data to Pre-train Your Language Model? MT to the Rescue!
Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi and Ander Corral   19:00-21:00 (Metropolitan Centre)
In recent years, pre-trained transformer-based language models (LM) have become a key resource for implementing most NLP tasks. However, pre-training such models demands large text collections not available in most languages. In this paper, we study the use of machine-translated corpora for pre-training LMs. We answer the following research questions: RQ1: Is MT-based data an alternative to real data for learning a LM?; RQ2: Can real data be complemented with translated data and improve the resulting LM? In order to validate these two questions, several BERT models for Basque have been trained, combining real data and synthetic data translated from Spanish. The evaluation carried out on 9 NLU tasks indicates that models trained exclusively on translated data offer competitive results. Furthermore, models trained with real data can be improved with synthetic data, although further research is needed on the matter.

Exclusive Supermask Subnetwork Training for Continual Learning
Prateek Yadav and Mohit Bansal   19:00-21:00 (Metropolitan Centre)
Continual Learning (CL) methods focus on accumulating knowledge over time while avoiding catastrophic forgetting. Recently, Wortsman et al. (2020) proposed a CL method, SupSup, which uses a randomly initialized, fixed base network (model) and finds a supermask for each new task that selectively keeps or removes weights to produce a subnetwork. They prevent forgetting as the network weights are not being updated. Although there is no forgetting, the performance of SupSup is sub-optimal because fixed weights restrict its representational power. Furthermore, there is no accumulation or transfer of knowledge inside the model when new tasks are learned. Hence, we propose ExSSNeT (Exclusive Supermask SubNetwork Training), that performs exclusive and non-overlapping subnetwork weight training. This avoids conflicting updates to the shared weights by subsequent tasks to improve performance while still preventing forgetting. Furthermore, we propose a novel KNN-based Knowledge Transfer (KKT) module that utilizes previously acquired knowledge to learn new tasks better and faster. We demonstrate that ExSSNeT outperforms strong previous methods on both NLP and Vision domains while preventing forgetting. Moreover, ExSSNeT is particularly advantageous for sparse masks that activate 2-10% of the model parameters, resulting in an average improvement of 8.3% over SupSup. Furthermore, ExSSNeT scales to a large number of tasks (100).

History repeats: Overcoming catastrophic forgetting for event-centric temporal knowledge graph completion
Mehrnoosh Mirtaheri, Mohammad Rostami and Aram Galstyan   19:00-21:00 (Metropolitan Centre)
Temporal knowledge graph (TKG) completion models typically rely on having access to the entire graph during training. However, in real-world scenarios, TKG data is often received incrementally as events unfold, leading to a dynamic non-stationary data distribution over time. While one could incorporate fine-tuning to existing methods to allow them to adapt to evolving TKG data, this can lead to forgetting previously learned patterns. Alternatively, retraining the model with the entire updated TKG can mitigate forgetting but is computationally burdensome. To address these challenges, we propose a general continual training framework that is applicable to any TKG completion method, and leverages two key ideas: (i) a temporal regularization that encourages repurposing of less important model parameters for learning new knowledge, and (ii) a clustering-based experience replay that reinforces the past knowledge by selectively preserving only a small portion of the past data. Our experimental results on widely used event-centric TKG datasets demonstrate the effectiveness of our proposed continual training framework in adapting to new events while reducing catastrophic forgetting. Further, we perform ablation studies to show the effectiveness of each component of our proposed framework. Finally, we investigate the relation between the memory dedicated to experience replay and the benefit gained from our clustering-based sampling strategy.

Label Agnostic Pre-training for Zero-shot Text Classification
Christopher Clarke, Yuzhao Heng, Yiping Kang, Krisztian Flautner, Lingjia Tang and Jason Mars   19:00-21:00 (Metropolitan Centre)
Conventional approaches to text classification typically assume the existence of a fixed set of predefined labels to which a given text can be classified. However, in real-world applications, there exists an infinite label space for describing a given text. In addition, depending on the aspect (sentiment, topic, etc.) and domain of the text (finance, legal, etc.), the interpretation of the label can vary greatly. This makes the task of text classification, particularly in the zero-shot scenario, extremely challenging. In this paper, we investigate the task of zero-shot text classification with the aim of improving the ability of pre-trained language models (PLMs) to generalize to both seen and unseen data across varying aspects and domains. To solve this we introduce two new simple yet effective pre-training strategies, Implicit and Explicit pre-training. These methods inject aspect-level understanding into the model at train time with the goal of conditioning the model to build task-level understanding. To evaluate this, we construct and release UTCD, a new benchmark dataset for evaluating text classification in zero-shot settings. Experimental results on UTCD show that our approach achieves improved zero-shot generalization on a suite of challenging datasets across an array of zero-shot formalizations.

ECOLA: Enhancing Temporal Knowledge Embeddings with Contextualized Language Representations
Zhen Han, Ruotong Liao, Jindong Gu, Yao Zhang, Zifeng Ding, Yujia Gu, Heinz Koeppl, Hinrich Schütze and Volker Tresp   19:00-21:00 (Metropolitan Centre)
Since conventional knowledge embedding models cannot take full advantage of the abundant textual information, there have been extensive research efforts in enhancing knowledge embedding using texts. However, existing enhancement approaches cannot apply to *temporal knowledge graphs* (tKGs), which contain time-dependent event knowledge with complex temporal dynamics. Specifically, existing enhancement approaches often assume knowledge embedding is time-independent. In contrast, the entity embedding in tKG models usually evolves, which poses the challenge of aligning *temporally relevant* texts with entities. To this end, we propose to study enhancing temporal knowledge embedding with textual data in this paper. As an approach to this task, we propose Enhanced Temporal Knowledge Embeddings with Contextualized Language Representations (ECOLA), which takes the temporal aspect into account and injects textual information into temporal knowledge embedding. To evaluate ECOLA, we introduce three new datasets for training and evaluating ECOLA. Extensive experiments show that ECOLA significantly enhances temporal KG embedding models with up to 287% relative improvements regarding Hits@1 on the link prediction task. The code and models are publicly available on https://github.com/mayhugotong/ECOLA.

Recyclable Tuning for Continual Pre-training

Yujia Qin, Cheng Qian, Xu Han, Yankai Lin, Huadong Wang, Ruobing Xie, Zhiyuan Liu, Maosong Sun and Jie Zhou 19:00-21:00 (Metropolitan Centre)

Continual pre-training is the paradigm where pre-trained language models (PLMs) continually acquire fresh knowledge from growing data and gradually get upgraded. Before an upgraded PLM is released, we may have tuned the original PLM for various tasks and stored the adapted weights. However, when tuning the upgraded PLM, these outdated adapted weights will typically be ignored and discarded, causing a potential waste of resources. We bring this issue to the forefront and contend that proper algorithms for recycling outdated adapted weights should be developed. To this end, we formulate the task of recyclable tuning for continual pre-training. In pilot studies, we find that after continual pre-training, the upgraded PLM remains compatible with the outdated adapted weights to some extent. Motivated by this finding, we analyze the connection between continually pre-trained PLMs from two novel aspects, i.e., mode connectivity, and functional similarity. Based on the corresponding findings, we propose both an initialization-based method and a distillation-based method for our task. We demonstrate their feasibility in improving the convergence and performance for tuning the upgraded PLM. We also show that both methods can be combined to achieve better performance.

The Larger they are, the Harder they Fail: Language Models do not Recognize Identifier Swaps in Python

Antonio Valerio Miceli Barone, Fazl Barez, Shay B. Cohen and Ioannis Konstas 19:00-21:00 (Metropolitan Centre)

Large Language Models (LLMs) have successfully been applied to code generation tasks, raising the question of how well these models understand programming. Typical programming languages have invariances and equivariances in their semantics that human programmers intuitively understand and exploit, such as the (near) invariance to the renaming of identifiers. We show that LLMs not only fail to properly generate correct Python code when default function names are swapped, but some of them even become more confident in their incorrect predictions as the model size increases, an instance of the recently discovered phenomenon of Inverse Scaling, which runs contrary to the commonly observed trend of increasing prediction quality with increasing model size. Our findings indicate that, despite their astonishing typical-case performance, LLMs still lack a deep, abstract understanding of the content they manipulate, making them unsuitable for tasks that statistically deviate from their training data, and that mere scaling is not enough to achieve such capability.

Evaluating the Factual Consistency of Large Language Models Through News Summarization

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal and Colin Raffel 19:00-21:00 (Metropolitan Centre)

While large language models (LLMs) have proven to be effective on a large variety of tasks, they are also known to hallucinate information. To measure whether an LLM prefers factually consistent continuations of its input, we propose a new benchmark called FIB (Factual Inconsistency Benchmark) that focuses on the task of summarization. Specifically, our benchmark involves comparing the scores an LLM assigns to a factually consistent versus a factually inconsistent summary for an input news article. For factually consistent summaries, we use human-written reference summaries that we manually verify as factually consistent. To generate summaries that are factually inconsistent, we generate summaries from a suite of summarization models that we have manually annotated as factually inconsistent. A model's factual consistency is then measured according to its accuracy, i.e. the proportion of documents where it assigns a higher score to the factually consistent summary. To validate the usefulness of {pasted macro 'BENCHMARK'}, we evaluate 23 large language models ranging from 1B to 176B parameters from six different model families including BLOOM and OPT. We find that existing LLMs generally assign a higher score to factually consistent summaries than to factually inconsistent summaries. However, if the factually inconsistent summaries occur verbatim in the document, then LLMs assign a higher score to these factually inconsistent summaries than factually consistent summaries. We validate design choices in our benchmark including the scoring method and source of distractor summaries.

HELP ME THINK: A Simple Prompting Strategy for Non-experts to Create Customized Content with Models

Swaroop Mishra and Elnaz Nouri 19:00-21:00 (Metropolitan Centre)

Controlling the text generated by language models and customizing the content has been a long-standing challenge. Existing prompting techniques proposed in pursuit of providing control are task-specific and lack generality; this provides overwhelming choices for non-expert users to find a suitable method for their task. The effort associated with those techniques, such as in writing examples, explanations, instructions, etc. further limits their adoption among non-expert users. In this paper, we propose a simple prompting strategy Help Me Think where we encourage large language models (such as GPT3 and ChatGPT) to help non-expert users by asking a set of relevant questions and leveraging user answers to execute the task. We demonstrate the efficacy of our technique Help Me Think on a variety of tasks. Specifically, we focus on tasks that are hard for average humans and require significant thinking to perform. We hope our work will encourage the development of unconventional ways to harness the power of large language models.

The Magic of IF: Investigating Causal Reasoning Abilities in Large Language Models of Code

Xiao Liu, Da Yin, Chen Zhang, Yansong Feng and Dongyan Zhao 19:00-21:00 (Metropolitan Centre)

Causal reasoning, the ability to identify cause-and-effect relationship, is crucial in human thinking. Although large language models (LLMs) succeed in many NLP tasks, it is still challenging for them to conduct complex causal reasoning like abductive reasoning and counterfactual reasoning. Given the fact that programming code may express causal relations more often and explicitly with conditional statements like "if", we want to explore whether Code-LLMs acquire better causal reasoning abilities. Our experiments show that compared to text-only LLMs, Code-LLMs with code prompts are better causal reasoners. We further intervene on the prompts from different aspects, and discover that the key point is the programming structure. Code and data are available at https://github.com/xxxiaol/magic-if.

Membership Inference Attacks against Language Models via Neighbourhood Comparison

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan and Taylor Berg-Kirkpatrick 19:00-21:00 (Metropolitan Centre)

Membership Inference attacks (MIAs) aim to predict whether a data sample was present in the training data of a machine learning model or not, and are widely used for assessing the privacy risks of language models. Most existing attacks rely on the observation that models tend to assign higher probabilities to their training samples than non-training points. However, simple thresholding of the model score in isolation tends to lead to high false-positive rates as it does not account for the intrinsic complexity of a sample. Recent work has demonstrated that reference-based attacks which compare model scores to those obtained from a reference model trained on similar data can substantially improve the performance of MIAs. However, in order to train reference models, attacks of this kind make the strong and arguably unrealistic assumption that an adversary has access to samples closely resembling the original training data. Therefore, we investigate their performance in more realistic scenarios and find that they are highly fragile in relation to the data distribution used to train reference models. To investigate whether this fragility provides a layer of safety, we propose and evaluate neighbourhood attacks, which compare model scores for a given sample to scores of synthetically generated neighbour texts and therefore eliminate the need for access to the training data distribution. We show that, in addition to being competitive with reference-based attacks that have perfect knowledge about the training data distribution, our attack clearly outperforms existing reference-free attacks as well as reference-based attacks with imperfect knowledge, which demonstrates the need for a reevaluation of the threat model of adversarial attacks.

Complementary Explanations for Effective In-Context Learning

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett and Ramakanth Pasunuru   19:00-21:00 (Metropolitan Centre)
Large language models (LLMs) have exhibited remarkable capabilities in learning from expla- nations in prompts, but there has been limited understanding of exactly how these explana- tions function or why they are effective. This work aims to better understand the mechanisms by which explanations are used for in-context learning. We first study the impact of two dif- ferent factors on the performance of prompts with explanations: the computation trace (the way the solution is decomposed) and the natural language used to express the prompt. By per- turbing explanations on three controlled tasks, we show that both factors contribute to the ef- fectiveness of explanations. We further study how to form maximally effective sets of expla- nations for solving a given test query. We find that LLMs can benefit from the complemen- tarity of the explanation set: diverse reasoning skills shown by different exemplars can lead to better performance. Therefore, we propose a maximal marginal relevance-based exemplar selection approach for constructing exemplar sets that are both relevant as well as comple- mentary, which successfully improves the in- context learning performance across three real- world tasks on multiple LLMs.

Nonparametric Masked Language Modeling
Sewon Min, Weijia Shi, Mike Lewis, Xilun Chen, Wen-tau Yih, Hannaneh Hajishirzi and Luke Zettlemoyer 19:00-21:00 (Metropolitan Centre)
Existing language models (LMs) predict tokens with a softmax over a finite vocabulary, which can make it difficult to predict rare tokens or phrases. We introduce NPM, the first nonparametric masked language model that replaces this softmax with a nonparametric distribution over every phrase in a reference corpus. NPM fills in the [MASK] solely from retrieving a token from a text corpus. We show that NPM can be ef- ficiently trained with a contrastive objective and an in-batch approximation to full corpus retrieval. Zero-shot evaluation on 16 tasks including classification, fact probing and question answering demonstrates that NPM outperforms significantly larger parametric models, either with or without a retrieve-and-generate approach. It is particularly better at dealing with rare patterns (word senses or facts) and predicting rare or nearly unseen words (e.g., non-Latin script). We release the model and code at github.com/facebookresearch/NPM.

Aligning Instruction Tasks Unlocks Large Language Models as Zero-Shot Relation Extractors
Kai Zhang, Bernal Jimenez Gutierrez and Yu Su   19:00-21:00 (Metropolitan Centre)
Recent work has shown that fine-tuning large language models (LLMs) on large-scale instruction-following datasets substantially improves their performance on a wide range of NLP tasks, especially in the zero-shot setting. However, even advanced instruction-tuned LLMs still fail to outperform small LMs on relation extraction (RE), a fundamental information extraction task. We hypothesize that instruction-tuning has been unable to elicit strong RE capabilities in LLMs due to RE's low incidence in instruction-tuning datasets, making up less than 1% of all tasks (Wang et al. 2022). To address this limitation, we propose QA4RE, a framework that aligns RE with question answering (QA), a predominant task in instruction-tuning datasets. Comprehensive zero-shot RE experiments over four datasets with two series of instruction- tuned LLMs (six LLMs in total) demonstrate that our QA4RE framework consistently improves LLM performance, strongly verifying our hypothesis and enabling LLMs to outperform strong zero-shot baselines by a large margin. Additionally, we provide thorough experiments and discussions to show the robustness, few-shot effectiveness, and strong transferability of our QA4RE framework. This work illustrates a promising way of adapting LLMs to challenging and underrepresented tasks by aligning these tasks with more common instruction-tuning tasks like QA.

Scaling Laws for BERT in Low-Resource Settings
Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri and Aitor Soroa   19:00-21:00 (Metropolitan Centre)
Large language models are very resource intensive, both financially and environmentally, and require an amount of training data which is simply unobtainable for the majority of NLP practitioners. Previous work has researched the scaling laws of such models, but optimal ratios of model parameters, dataset size, and computation costs focused on the large scale. In contrast, we analyze the effect those variables have on the performance of language models in constrained settings, by building three lightweight BERT models (16M/51M/124M parameters) trained over a set of small corpora (5M/25M/125M words). We experiment on four languages of different linguistic characteristics (Basque, Spanish, Swahili and Finnish), and evaluate the models on MLM and several NLU tasks. We conclude that the power laws for parameters, data and compute for low-resource settings differ from the optimal scaling laws previously inferred, and data requirements should be higher. Our insights are consistent across all the languages we study, as well as across the MLM and downstream tasks. Furthermore, we experimentally establish when the cost of using a Transformer-based approach is worth taking, instead of favouring other computationally lighter solutions.

Large Language Models with Controllable Working Memory
Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu and Sanjiv Kumar   19:00-21:00 (Metropolitan Centre)
Large language models (LLMs) have led to a series of breakthroughs in natural language processing (NLP), partly owing to the massive amounts of world knowledge they memorize during pretraining. While many downstream applications provide the model with an informa- tional context to aid its underlying task, how the model's world knowledge interacts with the factual information presented in the context remains under explored. As a desirable behavior, an LLM should give precedence to the context whenever it contains task-relevant infor- mation that conflicts with the model's memorized knowledge. This enables model predictions to be grounded in the context, which then facilitates updating specific model predictions without frequently retraining the model. By contrast, when the context is irrelevant to the task, the model should ignore it and fall back on its internal knowledge. In this paper, we undertake a first joint study of the aforementioned two properties, namely controllability and robustness, in the context of LLMs. We demonstrate that state-of-the-art T5 and PaLM models (both pretrained and finetuned) could exhibit low controllability and robustness that does not improve with increasing the model size. As a solution, we propose a simple yet effective method – knowledge aware finetuning (KAFT) – to strengthen both controllability and robustness by injecting counterfactual and irrelevant contexts to standard supervised datasets. Our comprehensive evaluation showcases the utility of KAFT across model architectures and sizes.

What In-Context Learning "Learns" In-Context: Disentangling Task Recognition and Task Learning
Jane Pan, Tianyu Gao, Howard Chen and Danqi Chen   19:00-21:00 (Metropolitan Centre)
Large language models (LLMs) exploit in-context learning (ICL) to solve tasks with only a few demonstrations, but its mechanisms are not yet well-understood. Some works suggest that LLMs only recall already learned concepts from pre-training, while others hint that ICL performs implicit learning over demonstrations. We characterize two ways through which ICL leverages demonstrations. Task recognition (TR) captures the extent to which LLMs can recognize a task through demonstrations – even without ground-truth labels – and apply their pre-trained priors, whereas task learning (TL) is the ability to capture new input-label mappings unseen in pre-training. Using a wide range of classification datasets and three LLM families (GPT-3, LLaMA and OPT), we design controlled experiments to disentangle the roles of TR and TL in ICL. We show that (1) models can achieve non-trivial performance with only TR, and TR does not further improve with larger models or more demonstrations; (2) LLMs acquire TL as the model scales, and TL's performance consistently improves with more demon- strations in context. Our findings unravel two different forces behind ICL and we advocate for discriminating them in future ICL research due to their distinct nature.

Recipes for Sequential Pre-training of Multilingual Encoder and Seq2Seq Models
Saleh Soltan, Andy Rosenbaum, Tobias Falke, Qin Lu, Anna Rumshisky and Wael Hamza   19:00-21:00 (Metropolitan Centre)
Pre-trained encoder-only and sequence-to-sequence (seq2seq) models each have advantages, however training both model types from scratch

is computationally expensive. We explore recipes to improve pre-training efficiency by initializing one model from the other. (1) Extracting the encoder from a seq2seq model, we show it under-performs a Masked Language Modeling (MLM) encoder, particularly on sequence labeling tasks. Variations of masking during seq2seq training, reducing the decoder size, and continuing with a small amount of MLM training do not close the gap. (2) Conversely, using an encoder to warm-start seq2seq training, we show that by unfreezing the encoder partway through training, we can match task performance of a from-scratch seq2seq model. Overall, this two-stage approach is an efficient recipe to obtain both a multilingual encoder and a seq2seq model, matching the performance of training each model from scratch while reducing the total compute cost by 27%.

Residual Prompt Tuning: improving prompt tuning with residual reparameterization
Anastasiia Razdaibiedina, Yuning Mao, Madian Khabsa, Mike Lewis, Rui Hou, Jimmy Ba and Amjad Almahairi   19:00-21:00 (Metropolitan Centre)
Prompt tuning is one of the successful approaches for parameter-efficient tuning of pre-trained language models. Despite being arguably the most parameter-efficient (tuned soft prompts constitute <0.1% of total parameters), it typically performs worse than other efficient tuning methods and is quite sensitive to hyper-parameters. In this work, we introduce Residual Prompt Tuning - a simple and efficient method that significantly improves the performance and stability of prompt tuning. We propose to reparameterize soft prompt embeddings using a shallow network with a residual connection. Our experiments show that Residual Prompt Tuning significantly outperforms prompt tuning across T5-Large, T5-Base and BERT-Base models. Notably, our method reaches +7 points improvement over prompt tuning on SuperGLUE benchmark with T5-Base model and allows to reduce the prompt length by 10 times without hurting performance. In addition, we show that our approach is robust to the choice of learning rate and prompt initialization, and is effective in few-shot settings.

Data-Efficient Finetuning Using Cross-Task Nearest Neighbors
Hamish Ivison, Noah A. Smith, Hannaneh Hajishirzi and Pradeep Dasigi   19:00-21:00 (Metropolitan Centre)
Obtaining labeled data to train a model for a task of interest is often expensive. Prior work shows training models on multitask data augmented with task descriptions (prompts) effectively transfers knowledge to new tasks. Towards efficiently building task-specific models, we assume access to a small number (32-1000) of unlabeled target-task examples and use those to retrieve the most similar labeled examples from a large pool of multitask data augmented with prompts. Compared to the current practice of finetuning models on uniformly sampled prompted multitask data (e.g.: FLAN, T0), our approach of finetuning on cross-task nearest neighbors is significantly more data-efficient. Using only 2% of the data from the P3 pool without any labeled target-task data, our models outperform strong baselines trained on all available data by 3-30% on 12 out of 14 datasets representing held-out tasks including legal and scientific document QA. Similarly, models trained on cross-task nearest neighbors from SuperNaturalInstructions, representing about 5% of that pool, obtain comparable performance to state-of-the-art models on 12 held-out tasks from that pool. Moreover, the models produced by our approach also provide a better initialization than single multitask finetuning for few-shot finetuning on target-task data, as shown by a 2-23

Honey, I Shrunk the Language: Language Model Behavior at Reduced Scale.
Vijeta Deshpande, Dan Pechi, Shree Thatte, Vladislav Lialin and Anna Rumshisky   19:00-21:00 (Metropolitan Centre)
In recent years, language models have drastically grown in size, and the abilities of these models have been shown to improve with scale. The majority of recent scaling laws studies focused on high-compute high-parameter count settings, leaving the question of when these abilities begin to emerge largely unanswered. In this paper, we investigate whether the effects of pre-training can be observed when the problem size is reduced, modeling a smaller, reduced-vocabulary language. We show the benefits of pre-training with masked language modeling (MLM) objective in models as small as 1.25M parameters, and establish a strong correlation between pre-training perplexity and downstream performance (GLUE benchmark). We examine downscaling effects, extending scaling laws to models as small as 1M parameters. At this scale, we observe a break of the power law for compute-optimal models and show that the MLM loss does not scale smoothly with compute-cost (FLOPs) below 2.2×10^{15} FLOPs. We also find that adding layers does not always benefit downstream performance.Our filtered pre-training data, reduced English vocabulary, and code are available at https://github.com/text-machine-lab/mini_bert $github.com/text-machine-lab/mini\_bert$

Masked Latent Semantic Modeling: an Efficient Pre-training Alternative to Masked Language Modeling
Gábor Berend   19:00-21:00 (Metropolitan Centre)
In this paper, we propose an alternative to the classic masked language modeling (MLM) pre-training paradigm, where the objective is altered from the reconstruction of the exact identity of randomly selected masked subwords to the prediction of their latent semantic properties. We coin the proposed pre-training technique masked latent semantic modeling (MLSM for short). In order to make the contextualized determination of the latent semantic properties of the masked subwords possible, we rely on an unsupervised technique which uses sparse coding. Our experimental results reveal that the fine-tuned performance of those models that we pre-trained via MLSM is consistently and significantly better compared to the use of vanilla MLM pretraining and other strong baselines.

Beyond Positive Scaling: How Negation Impacts Scaling Trends of Language Models
Yuhui Zhang, Michihiro Yasunaga, Zhengping Zhou, Jeff Z. HaoChen, James Zou, Percy Liang and Serena Yeung   19:00-21:00 (Metropolitan Centre)
Language models have been shown to exhibit positive scaling, where performance improves as models are scaled up in terms of size, compute, or data. In this work, we introduce NeQA, a dataset consisting of questions with negation in which language models do not exhibit straightforward positive scaling. We show that this task can exhibit inverse scaling, U-shaped scaling, or positive scaling, and the three scaling trends shift in this order as we use more powerful prompting methods or model families. We hypothesize that solving NeQA depends on two subtasks: question answering (task 1) and negation understanding (task 2). We find that task 1 has linear scaling, while task 2 has sigmoid-shaped scaling with an emergent transition point, and composing these two scaling trends yields the final scaling trend of NeQA. Our work reveals and provides a way to analyze the complex scaling trends of language models.

How does the task complexity of masked pretraining objectives affect downstream performance?
Atsuki Yamaguchi, Hiroaki Ozaki, Terufumi Morishita, Gaku Morio and Yasuhiro Sogawa   19:00-21:00 (Metropolitan Centre)
Masked language modeling (MLM) is a widely used self-supervised pretraining objective, where a model needs to predict an original token that is replaced with a mask given contexts. Although simpler and computationally efficient pretraining objectives, e.g., predicting the first character of a masked token, have recently shown comparable results to MLM, no objectives with a masking scheme actually outperform it in downstream tasks. Motivated by the assumption that their lack of complexity plays a vital role in the degradation, we validate whether more complex masked objectives can achieve better results and investigate how much complexity they should have to perform comparably to MLM. Our results using GLUE, SQuAD, and Universal Dependencies benchmarks demonstrate that more complicated objectives tend to show better downstream results with at least half of the MLM complexity needed to perform comparably to MLM. Finally, we discuss how we should pretrain a model using a masked objective from the task complexity perspective.

Critic-Guided Decoding for Controlled Text Generation
Minbeom Kim, Hwanhee Lee, Kang Min Yoo, Joonsuk Park, Hwaran Lee and Kyomin Jung   19:00-21:00 (Metropolitan Centre)
Steering language generation towards objectives or away from undesired content has been a long-standing goal in utilizing language models

(LM). Recent work has demonstrated reinforcement learning and weighted decoding as effective approaches to achieve a higher level of language control and quality with pros and cons. In this work, we propose a novel critic decoding method for controlled language generation (CriticControl) that combines the strengths of reinforcement learning and weighted decoding. Specifically, we adopt the actor-critic framework and train an LM-steering critic from reward models. Similar to weighted decoding, our method freezes the language model and manipulates the output token distribution using a critic to improve training efficiency and stability. Evaluation of our method on three controlled generation tasks, topic control, sentiment control, and detoxification, shows that our approach generates more coherent and well-controlled texts than previous methods. In addition, CriticControl demonstrates superior generalization ability in zero-shot settings. Human evaluation studies also corroborate our findings.

Revisiting the Architectures like Pointer Networks to Efficiently Improve the Next Word Distribution, Summarization Factuality, and Beyond
Haw-Shiuan Chang, Zonghai Yao, Alolika Gon, hong yu and Andrew McCallum   19:00-21:00 (Metropolitan Centre)
Is the output softmax layer, which is adopted by most language models (LMs), always the best way to compute the next word probability? Given so many attention layers in a modern transformer-based LM, are the pointer networks redundant nowadays? In this study, we discover that the answers to both questions are no. This is because the softmax bottleneck sometimes prevents the LMs from predicting the desired distribution and the pointer networks can be used to break the bottleneck efficiently. Based on the finding, we propose several softmax alternatives by simplifying the pointer networks and accelerating the word-by-word rerankers. In GPT-2, our proposals are significantly better and more efficient than mixture of softmax, a state-of-the-art softmax alternative. In summarization experiments, without very significantly decreasing its training/testing speed, our best method based on T5-Small improves factCC score by 2 points in CNN/DM and XSUM dataset, and improves MAUVE scores by 30% in BookSum paragraph-level dataset.

Focus-aware Response Generation in Inquiry Conversation
Yiquan Wu, Weiming Lu, Yating Zhang, Adam Jatowt, Jun Feng, Changlong Sun, Fei Wu and Kun Kuang 19:00-21:00 (Metropolitan Centre)
Inquiry conversation is a common form of conversation that aims to complete the investigation (e.g., court hearing, medical consultation and police interrogation) during which a series of focus shifts occurs. While many models have been proposed to generate a smooth response to a given conversation history, neglecting the focus can limit performance in inquiry conversation where the order of the focuses plays there a key role. In this paper, we investigate the problem of response generation in inquiry conversation by taking the focus into consideration. We propose a novel Focus-aware Response Generation (FRG) method by jointly optimizing a multi-level encoder and a set of focal decoders to generate several candidate responses that correspond to different focuses. Additionally, a focus ranking module is proposed to predict the next focus and rank the candidate responses. Experiments on two orthogonal inquiry conversation datasets (judicial, medical domain) demonstrate that our method generates results significantly better in automatic metrics and human evaluation compared to the state-of-the-art approaches.

Nano: Nested Human-in-the-Loop Reward Learning for Few-shot Language Model Control
Xiang Fan, Yiwei Lyu, Paul Pu Liang, Ruslan Salakhutdinov and Louis-Philippe Morency   19:00-21:00 (Metropolitan Centre)
Pretrained language models have demonstrated extraordinary capabilities in language generation. However, real-world tasks often require controlling the distribution of generated text in order to mitigate bias, promote fairness, and achieve personalization. Existing techniques for controlling the distribution of generated text only work with quantified distributions, which require pre-defined categories, proportions of the distribution, or an existing corpus following the desired distributions. However, many important distributions, such as personal preferences, are unquantified. In this work, we tackle the problem of generating text following arbitrary distributions (quantified and unquantified) by proposing NANO, a few-shot human-in-the-loop training algorithm that continuously learns from human feedback. NANO achieves state-of-the-art results on single topic/attribute as well as quantified distribution control compared to previous works. We also show that NANO is able to learn unquantified distributions, achieves personalization, and captures differences between different individuals' personal preferences with high sample efficiency.

Differentiable Instruction Optimization for Cross-Task Generalization
Masaru Isonuma, Junichiro Mori and Ichiro Sakata   19:00-21:00 (Metropolitan Centre)
Instruction tuning has been attracting much attention to achieve generalization ability across a wide variety of tasks. Although various types of instructions have been manually created for instruction tuning, it is still unclear what kind of instruction is optimal to obtain cross-task generalization ability. This work presents instruction optimization, which optimizes training instructions with respect to generalization ability. Rather than manually tuning instructions, we introduce learnable instructions and optimize them with gradient descent by leveraging bilevel optimization. Experimental results show that the learned instruction enhances the diversity of instructions and improves the generalization ability compared to using only manually created instructions.

Revisiting Sentence Union Generation as a Testbed for Text Consolidation
Eran Hirsch, Valentina Pyatkin, Ruben Wolhandler, Avi Caciularu, Asi Shefer and Ido Dagan   19:00-21:00 (Metropolitan Centre)
Tasks involving text generation based on multiple input texts, such as multi-document summarization, long-form question answering and contemporary dialogue applications, challenge models for their ability to properly consolidate partly-overlapping multi-text information. However, these tasks entangle the consolidation phase with the often subjective and ill-defined content selection requirement, impeding proper assessment of models' consolidation capabilities. In this paper, we suggest revisiting the sentence union generation task as an effective well-defined testbed for assessing text consolidation capabilities, decoupling the consolidation challenge from subjective content selection. To support research on this task, we present refined annotation methodology and tools for crowdsourcing sentence union, create the largest union dataset to date and provide an analysis of its rich coverage of various consolidation aspects. We then propose a comprehensive evaluation protocol for union generation, including both human and automatic evaluation. Finally, as baselines, we evaluate state-of-the-art language models on the task, along with a detailed analysis of their capacity to address multi-text consolidation challenges and their limitations.

PREADD: Prefix-Adaptive Decoding for Controlled Text Generation
Jonathan Pei, Kevin Yang and Dan Klein   19:00-21:00 (Metropolitan Centre)
We propose Prefix-Adaptive Decoding (PREADD), a flexible method for controlled text generation. Unlike existing methods that use auxiliary expert models to control for attributes, PREADD does not require an external model, instead relying on linearly combining output logits from multiple prompts. Specifically, PREADD contrasts the output logits generated using a raw prompt against those generated using a prefix-prepended prompt, enabling both positive and negative control with respect to any attribute encapsulated by the prefix. We evaluate PREADD on three tasks—toxic output mitigation, gender bias reduction, and sentiment control—and find that PREADD outperforms not only prompting baselines, but also an auxiliary-expert control method, by 12% or more in relative gain on our main metrics for each task.

Efficient Out-of-Domain Detection for Sequence to Sequence Models
Artem Vazhentsev, Akim Tsvigun, Roman Konstantinovich Vashurin, Sergey Petrakov, Daniil Vasilev, Maxim Panov, Alexander Panchenko and Artem Shelmanov   19:00-21:00 (Metropolitan Centre)
Sequence-to-sequence (seq2seq) models based on the Transformer architecture have become a ubiquitous tool applicable not only to classical text generation tasks such as machine translation and summarization but also to any other task where an answer can be represented in a form

of a finite text fragment (e.g., question answering). However, when deploying a model in practice, we need not only high performance but also an ability to determine cases where the model is not applicable. Uncertainty estimation (UE) techniques provide a tool for identifying out-of-domain (OOD) input where the model is susceptible to errors. State-of-the-art UE methods for seq2seq models rely on computationally heavyweight and impractical deep ensembles. In this work, we perform an empirical investigation of various novel UE methods for large pre-trained seq2seq models T5 and BART on three tasks: machine translation, text summarization, and question answering. We apply computationally lightweight density-based UE methods to seq2seq models and show that they often outperform heavyweight deep ensembles on the task of OOD detection.

Distilling Reasoning Capabilities into Smaller Language Models

Kumar Shridhar, Alessandro Stolfo and Mrinmaya Sachan   19:00-21:00 (Metropolitan Centre)

Step-by-step reasoning approaches like chain of thought (CoT) have proved to be very effective in inducing reasoning capabilities in large language models. However, the success of the CoT approach is fundamentally tied to the model size, and billion parameter-scale models are often needed to get CoT to work. In this paper, we propose a knowledge distillation approach that leverages the step-by-step CoT reasoning capabilities of larger models and distills these abilities into smaller models.

In this work, we propose an alternative reasoning scheme, Socratic CoT that learns a decomposition of the original problem into a sequence of subproblems and uses it to guide the intermediate reasoning steps. We use Socratic CoT to train a combination of two small distilled models: a problem decomposer and a subproblem solver. In practice, given a new problem, the two distilled models work in sync to decompose and solve complex problems. On multiple reasoning datasets (GSM8K, StrategyQA, and SVAMP), our proposed distillation strategies boosts the performance of smaller models over 70% compared to the baselines. Finally, we investigate when Socratic CoT is an effective alternative to CoT, demonstrating cases where a much smaller model (GPT-2 large) can outperform a 10X larger model (GPT-3 6B). Our code is available: https://github.com/kumar-shridhar/Distiiling-LM.

Language Modeling with Latent Situations

Belinda Z. Li, Maxwell Nye and Jacob Andreas   19:00-21:00 (Metropolitan Centre)

Language models (LMs) often generate incoherent outputs: they refer to events and entity states that are incompatible with the state of the world described in inputs. We introduce SITUATIONSUPERVISION, a family of approaches for improving coherence in LMs by training them to construct and condition on explicit representations of entities and their states. SITUATIONSUPERVISION has two components: an *auxiliary situation modeling* task that trains models to predict entity state representations in context, and a *latent state inference* procedure that imputes these states from partially annotated training data. SITUATIONSUPERVISION can be applied via fine-tuning (by supervising LMs to encode state variables in their hidden representations) and prompting (by inducing LMs to interleave textual descriptions of entity states with output text). In both cases, it requires only a small number of state annotations to produce substantial coherence improvements (up to an 16% reduction in errors), showing that standard LMs can be efficiently adapted to explicitly model language and aspects of its meaning.

Unsupervised Summarization Re-ranking

Mathieu Ravaut, Shafiq Joty and Nancy Chen   19:00-21:00 (Metropolitan Centre)

With the rise of task-specific pre-training objectives, abstractive summarization models like PEGASUS offer appealing zero-shot performance on downstream summarization tasks. However, the performance of such unsupervised models still lags significantly behind their supervised counterparts. Similarly to the supervised setup, we notice a very high variance in quality among summary candidates from these models while only one candidate is kept as the summary output. In this paper, we propose to re-rank summary candidates in an unsupervised manner, aiming to close the performance gap between unsupervised and supervised models. Our approach improves the unsupervised PEGASUS by up to 7.27% and ChatGPT by up to 6.86% relative mean ROUGE across four widely-adopted summarization benchmarks ; and achieves relative gains of 7.51% (up to 23.73% from XSum to WikiHow) averaged over 30 zero-shot transfer setups (finetuning on a dataset, evaluating on another).

RISE: Leveraging Retrieval Techniques for Summarization Evaluation

David Uthus and Jianmo Ni   19:00-21:00 (Metropolitan Centre)

Evaluating automatically-generated text summaries is a challenging task. While there have been many interesting approaches, they still fall short of human evaluations. We present RISE, a new approach for evaluating summaries by leveraging techniques from information retrieval. RISE is first trained as a retrieval task using a dual-encoder retrieval setup, and can then be subsequently utilized for evaluating a generated summary given an input document, without gold reference summaries. RISE is especially well suited when working on new datasets where one may not have reference summaries available for evaluation. We conduct comprehensive experiments on the SummEval benchmark (Fabbri et al., 2021) and a long document summarization benchmark. The results show that RISE consistently achieves higher correlation with human evaluations compared to many past approaches to summarization evaluation. Furthermore, RISE also demonstrates data-efficiency and generalizability across languages.

OpineSum: Entailment-based self-training for abstractive opinion summarization

Annie Louis and Joshua Maynez   19:00-21:00 (Metropolitan Centre)

A typical product or place often has hundreds of reviews, and summarization of these texts is an important and challenging problem. Recent progress on abstractive summarization in domains such as news has been driven by supervised systems trained on hundreds of thousands of news articles paired with human-written summaries. However for opinion texts, such large scale datasets are rarely available. Unsupervised methods, self-training, and few-shot learning approaches bridge that gap. In this work, we present a novel self-training approach, OpineSum for abstractive opinion summarization. The self-training summaries in this approach are built automatically using a novel application of textual entailment and capture the consensus of opinions across the various reviews for an item. This method can be used to obtain silver-standard summaries on a large scale and train both unsupervised and few-shot abstractive summarization systems. OpineSum outperforms strong peer systems in both settings.

Towards Argument-Aware Abstractive Summarization of Long Legal Opinions with Summary Reranking

Mohamed Elaraby, Yang Zhong and Diane Litman   19:00-21:00 (Metropolitan Centre)

We propose a simple approach for the abstractive summarization of long legal opinions that takes into account the argument structure of the document. Legal opinions often contain complex and nuanced argumentation, making it challenging to generate a concise summary that accurately captures the main points of the legal opinion. Our approach involves using argument role information to generate multiple candidate summaries, then reranking these candidates based on alignment with the document's argument structure. We demonstrate the effectiveness of our approach on a dataset of long legal opinions and show that it outperforms several strong baselines.

Multi-Dimensional Evaluation of Text Summarization with In-Context Learning

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig and Chunting Zhou 19:00-21:00 (Metropolitan Centre)

Evaluation of natural language generation (NLG) is complex and multi-dimensional. Generated text can be evaluated for fluency, coherence, factuality, or any other dimensions of interest. Most frameworks that perform such multi-dimensional evaluation require training on large

manually or synthetically generated datasets. In this paper, we study the efficacy of large language models as multi-dimensional evaluators using in-context learning, obviating the need for large training datasets. Our experiments show that in-context learning-based evaluators are competitive with learned evaluation frameworks for the task of text summarization, establishing state-of-the-art on dimensions such as relevance and factual consistency. We then analyze the effects of factors such as the selection and number of in-context examples on performance. Finally, we study the efficacy of in-context learning-based evaluators in evaluating zero-shot summaries written by large language models such as GPT-3.

Improving Long Dialogue Summarization with Semantic Graph Representation
Yilun Hua, Zhaoyuan Deng and Kathleen McKeown 19:00-21:00 (Metropolitan Centre)
Although Large Language Models (LLMs) are successful in abstractive summarization of short dialogues, summarization of long dialogues remains challenging. To address this challenge, we propose a novel algorithm that processes complete dialogues comprising thousands of tokens into topic-segment-level Abstract Meaning Representation (AMR) graphs, which explicitly capture the dialogue structure, highlight salient semantics, and preserve high-level information. We also develop a new text-graph attention to leverage both graph semantics and a pretrained LLM that exploits the text. Finally, we propose an AMR node selection loss used jointly with conventional cross-entropy loss, to create additional training signals that facilitate graph feature encoding and content selection. Experiments show that our system outperforms the state-of-the-art models on multiple long dialogue summarization datasets, especially in low-resource settings, and generalizes well to out-of-domain data.

Aspect-aware Unsupervised Extractive Opinion Summarization
Haoyuan Li, Somnath Basu Roy Chowdhury and Snigdha Chaturvedi 19:00-21:00 (Metropolitan Centre)
Extractive opinion summarization extracts sentences from users' reviews to represent the prevalent opinions about a product or service. However, the extracted sentences can be redundant and may miss some important aspects, especially for centroid-based extractive summarization models (Radev et al., 2004). To alleviate these issues, we introduce TokenCluster– a method for unsupervised extractive opinion summarization that automatically identifies the aspects described in the review sentences and then extracts sentences based on their aspects. It identifies the underlying aspects of the review sentences using roots of noun phrases and adjectives appearing in them. Empirical evaluation shows that TokenCluster improves aspect coverage in summaries and achieves strong performance on multiple opinion summarization datasets, for both general and aspect-specific summarization. We also perform extensive ablation and human evaluation studies to validate the design choices of our method. The implementation of our work is available at https://github.com/leehaoyuan/TokenCluster

An Investigation of Evaluation Methods in Automatic Medical Note Generation
Asma Ben Abacha, Wen-wai Yim, George Michalopoulos and Thomas Lin 19:00-21:00 (Metropolitan Centre)
Recent studies on automatic note generation have shown that doctors can save significant amounts of time when using automatic clinical note generation (Knoll et al., 2022). Summarization models have been used for this task to generate clinical notes as summaries of doctor-patient conversations (Krishna et al., 2021; Cai et al., 2022). However, assessing which model would best serve clinicians in their daily practice is still a challenging task due to the large set of possible correct summaries, and the potential limitations of automatic evaluation metrics. In this paper we study evaluation methods and metrics for the automatic generation of clinical notes from medical conversation. In particular, we propose new task-specific metrics and we compare them to SOTA evaluation metrics in text summarization and generation, including: (i) knowledge-graph embedding-based metrics, (ii) customized model-based metrics with domain-specific weights, (iii) domain-adapted/fine-tuned metrics, and (iv) ensemble metrics. To study the correlation between the automatic metrics and manual judgments, we evaluate automatic notes/summaries by comparing the system and reference facts and computing the factual correctness, and the hallucination and omission rates for critical medical facts. This study relied on seven datasets manually annotated by domain experts. Our experiments show that automatic evaluation metrics can have substantially different behaviors on different types of clinical notes datasets. However, the results highlight one stable subset of metrics as the most correlated with human judgments with a relevant aggregation of different evaluation criteria.

A Formal Perspective on Byte-Pair Encoding
Vilém Zouhar, Clara Meister, Juan Luis Gastaldi, Li Du, Tim Vieira, Mrinmaya Sachan and Ryan Cotterell 19:00-21:00 (Metropolitan Centre)
Byte-Pair Encoding (BPE) is a popular algorithm used for tokenizing data in NLP, despite being devised initially as a compression method. BPE appears to be a greedy algorithm at face value, but the underlying optimization problem that BPE seeks to solve has not yet been laid down. We formalize BPE as a combinatorial optimization problem. Via submodular functions, we prove that the iterative greedy version is a $1/\sigma^{*}(1-e^{(-\sigma)})$-approximation of an optimal merge sequence, where sigma is the total backward curvature with respect to the optimal merge sequence. Empirically, the lower bound of the approximation is approx0.37.
We provide a faster implementation of BPE which improves the runtime complexity from $O(NM)$ to $O(N \log M)$, where N is the sequence length and M is the merge count. Finally, we optimize the brute-force algorithm for optimal BPE using memoization.

Lego-MT: Learning Detachable Models for Massively Multilingual Machine Translation
Fei Yuan, Yinquan Lu, Wenhao Zhu, Lingpeng Kong, Lei Li, Yu Qiao and Jingjing Xu 19:00-21:00 (Metropolitan Centre)
Multilingual neural machine translation (MNMT) aims to build a unified model for many language directions. Existing monolithic models for MNMT encounter two challenges: parameter interference among languages and inefficient inference for large models. In this paper, we revisit the classic multi-way structures and develop a detachable model by assigning each language (or group of languages) to an individual branch that supports plug-and-play training and inference. To address the needs of learning representations for all languages in a unified space, we propose a novel efficient training recipe, upon which we build an effective detachable model, Lego-MT. For a fair comparison, we collect data from OPUS and build a translation benchmark covering 433 languages and 1.3B parallel data. Experiments show that Lego-MT with 1.2B parameters brings an average gain of 3.2 spBLEU. It even outperforms M2M-100 with 12B parameters. The proposed training recipe brings a 28.2× speedup over the conventional multi-way training method.code and data repo: https://github.com/CONE-MT/Lego-MT.git.

What Knowledge Is Needed? Towards Explainable Memory for kNN-MT Domain Adaptation
Wenhao Zhu, Shujian Huang, Yunzhe Lv, Xin Zheng and Jiajun CHEN 19:00-21:00 (Metropolitan Centre)
kNN-MT presents a new paradigm for domain adaptation by building an external datastore, which usually saves all target language token occurrences in the parallel corpus. As a result, the constructed datastore is usually large and possibly redundant. In this paper, we investigate the interpretability issue of this approach: what knowledge does the NMT model need? We propose the notion of local correctness (LAC) as a new angle, which describes the potential translation correctness for a single entry and for a given neighborhood. Empirical study shows that our investigation successfully finds the conditions where the NMT model could easily fail and need related knowledge. Experiments on six diverse target domains and two language-pairs show that pruning according to local correctness brings a light and more explainable memory for kNN-MT domain adaptation.

Towards Speech Dialogue Translation Mediating Speakers of Different Languages
Shuichiro Shimizu, Chenhui Chu, Sheng Li and Sadao Kurohashi 19:00-21:00 (Metropolitan Centre)

We present a new task, speech dialogue translation mediating speakers of different languages. We construct the SpeechBSD dataset for the task and conduct baseline experiments. Furthermore, we consider context to be an important aspect that needs to be addressed in this task and propose two ways of utilizing context, namely monolingual context and bilingual context. We conduct cascaded speech translation experiments using Whisper and mBART, and show that bilingual context performs better in our settings.

DUB: Discrete Unit Back-translation for Speech Translation
Dong Zhang, Rong Ye, Tom Ko, Mingxuan Wang and Yaqian Zhou   19:00-21:00 (Metropolitan Centre)
How can speech-to-text translation (ST) perform as well as machine translation (MT)? The key point is to bridge the modality gap between speech and text so that useful MT techniques can be applied to ST. Recently, the approach of representing speech with unsupervised discrete units yields a new way to ease the modality problem. This motivates us to propose Discrete Unit Back-translation(DUB) to answer two questions (1) Is it better to represent speech with discrete units than with continuous features in direct ST? (2) How much benefit can useful MT techniques bring to ST? With DUB, the back-translation technique can successfully be applied on direct ST and obtains an average boost of 5.5 BLEU on MuST-C En-De/Fr/Es. In the low-resource language scenario, our method achieves comparable performance to existing methods that rely on large-scale external data. Code and models are available at https://anonymous.4open.science/r/DUB/.

Synthetic Pre-Training Tasks for Neural Machine Translation
Zexue He, Graeme Blackwood, Rameswar Panda, Julian McAuley and Rogerio Feris   19:00-21:00 (Metropolitan Centre)
Pre-training models with large crawled corpora can lead to issues such as toxicity and bias, as well as copyright and privacy concerns. A promising way of alleviating such concerns is to conduct pre-training with synthetic tasks and data, since no real-world information is ingested by the model. Our goal in this paper is to understand the factors that contribute to the effectiveness of pre-training models when using synthetic resources, particularly in the context of neural machine translation. We propose several novel approaches to pre-training translation models that involve different levels of lexical and structural knowledge, including: 1) generating obfuscated data from a large parallel corpus 2) concatenating phrase pairs extracted from a small word-aligned corpus, and 3) generating synthetic parallel data without real human language corpora. Our experiments on multiple language pairs reveal that pre-training benefits can be realized even with high levels of obfuscation or purely synthetic parallel data. We hope the findings from our comprehensive empirical analysis will shed light on understanding what matters for NMT pre-training, as well as pave the way for the development of more efficient and less toxic models.

Towards Accurate Translation via Semantically Appropriate Application of Lexical Constraints
Yujin Baek, Koanho Lee, Dayeon Ki, Cheonbok Park, Hyoung-Gyu Lee and Jaegul Choo   19:00-21:00 (Metropolitan Centre)
Lexically-constrained NMT (LNMT) aims to incorporate user-provided terminology into translations. Despite its practical advantages, existing work has not evaluated LNMT models under challenging real-world conditions. In this paper, we focus on two important but understudied issues that lie in the current evaluation process of LNMT studies. The model needs to cope with challenging lexical constraints that are "homographs" or "unseen" during training. To this end, we first design a homograph disambiguation module to differentiate the meanings of homographs. Moreover, we propose PLUMCOT which integrates contextually rich information about unseen lexical constraints from pre-trained language models and strengthens a copy mechanism of the pointer network via direct supervision of a copying score. We also release HOLLY, an evaluation benchmark for assessing the ability of model to cope with "homographic" and "unseen" lexical constraints. Experiments on HOLLY and the previous test setup show the effectiveness of our method. The effects of PLUMCOT are shown to be remarkable in "unseen" constraints.Our dataset is available at https://github.com/papago-lab/HOLLY-benchmark.

Dual-Gated Fusion with Prefix-Tuning for Multi-Modal Relation Extraction
Qian Li, Shu Guo, Cheng Ji, Xutan Peng, Shiyao Cui, Jianxin Li and Lihong Wang   19:00-21:00 (Metropolitan Centre)
Multi-Modal Relation Extraction (MMRE) aims at identifying the relation between two entities in texts that contain visual clues. Rich visual content is valuable for the MMRE task, but existing works cannot well model finer associations among different modalities, failing to capture the truly helpful visual information and thus limiting relation extraction performance. In this paper, we propose a novel MMRE framework to better capture the deeper correlations of text, entity pair, and image/objects, so as to mine more helpful information for the task, termed as DGF-PT. We first propose a prompt-based autoregressive encoder, which builds the associations of intra-modal and inter-modal features related to the task, respectively by entity-oriented and object-oriented prefixes. To better integrate helpful visual information, we design a dual-gated fusion module to distinguish the importance of image/objects and further enrich text representations. In addition, a generative decoder is introduced with entity type restriction on relations, better filtering out candidates. Extensive experiments conducted on the benchmark dataset show that our approach achieves excellent performance compared to strong competitors, even in the few-shot situation.

Multi-hop Evidence Retrieval for Cross-document Relation Extraction
Keming Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma and Muhao Chen   19:00-21:00 (Metropolitan Centre)
Relation Extraction (RE) has been extended to cross-document scenarios because many relations are not simply described in a single document. This inevitably brings the challenge of efficient open-space evidence retrieval to support the inference of cross-document relations, along with the challenge of multi-hop reasoning on top of entities and evidence scattered in an open set of documents. To combat these challenges, we propose Mr.Cod (Multi-hop evidence retrieval for Cross-document relation extraction), which is a multi-hop evidence retrieval method based on evidence path mining and ranking. We explore multiple variants of retrievers to show evidence retrieval is essential in cross-document RE. We also propose a contextual dense retriever for this setting. Experiments on CodRED show that evidence retrieval with Mr.Cod effectively acquires cross-document evidence and boosts end-to-end RE performance in both closed and open settings.

A Diffusion Model for Event Skeleton Generation
Fangqi Zhu, Lin Zhang, Jun Gao, Bing Qin, Ruifeng Xu and Haiqin Yang   19:00-21:00 (Metropolitan Centre)
Event skeleton generation, aiming to induce an event schema skeleton graph with abstracted event nodes and their temporal relations from a set of event instance graphs, is a critical step in the temporal complex event schema induction task. Existing methods effectively address this task from a graph generation perspective but suffer from noise-sensitive and error accumulation, e.g., the inability to correct errors while generating schema. We, therefore, propose a novel Diffusion Event Graph Model (DEGM) to address these issues. Our DEGM is the first workable diffusion model for event skeleton generation, where the embedding and rounding techniques with a custom edge-based loss are introduced to transform a discrete event graph into learnable latent representations. Furthermore, we propose a denoising training process to maintain the model's robustness. Consequently, DEGM derives the final schema, where error correction is guaranteed by iteratively refining the latent representations during the schema generation process. Experimental results on three IED bombing datasets demonstrate that our DEGM achieves better results than other state-of-the-art baselines. Our code and data are available at https://github.com/zhufq00/EventSkeletonGeneration.

Data Augmentation for Low-Resource Keyphrase Generation
Krishna K Garg, Jishnu Ray Chowdhury and Cornelia Caragea   19:00-21:00 (Metropolitan Centre)
Keyphrase generation is the task of summarizing the contents of any given article into a few salient phrases (or keyphrases). Existing works for the task mostly rely on large-scale annotated datasets, which are not easy to acquire. Very few works address the problem of keyphrase generation in low-resource settings, but they still rely on a lot of additional unlabeled data for pretraining and on automatic methods for pseudo-annotations. In this paper, we present data augmentation strategies specifically to address keyphrase generation in purely resource-

constrained domains. We design techniques that use the full text of the articles to improve both present and absent keyphrase generation. We test our approach comprehensively on three datasets and show that the data augmentation strategies consistently improve the state-of-the-art performance. We release our source code at https://github.com/kgarg8/kpgen-lowres-data-aug.

Silver Syntax Pre-training for Cross-Domain Relation Extraction

Elisa Bassignana, Filip Ginter, Sampo Pyysalo, Rob van der Goot and Barbara Plank 19:00-21:00 (Metropolitan Centre)

Relation Extraction (RE) remains a challenging task, especially when considering realistic out-of-domain evaluations. One of the main reasons for this is the limited training size of current RE datasets: obtaining high-quality (manually annotated) data is extremely expensive and cannot realistically be repeated for each new domain. An intermediate training step on data from related tasks has shown to be beneficial across many NLP tasks.However, this setup still requires supplementary annotated data, which is often not available. In this paper, we investigate intermediate pre-training specifically for RE. We exploit the affinity between syntactic structure and semantic RE, and identify the syntactic relations which are closely related to RE by being on the shortest dependency path between two entities. We then take advantage of the high accuracy of current syntactic parsers in order to automatically obtain large amounts of low-cost pre-training data. By pre-training our RE model on the relevant syntactic relations, we are able to outperform the baseline in five out of six cross-domain setups, without any additional annotated data.

Teamwork Is Not Always Good: An Empirical Study of Classifier Drift in Class-incremental Information Extraction

Minqian Liu and Lifu Huang 19:00-21:00 (Metropolitan Centre)

Class-incremental learning (CIL) aims to develop a learning system that can continually learn new classes from a data stream without forgetting previously learned classes. When learning classes incrementally, the classifier must be constantly updated to incorporate new classes, and the drift in decision boundary may lead to severe forgetting. This fundamental challenge, however, has not yet been studied extensively, especially in the setting where no samples from old classes are stored for rehearsal. In this paper, we take a closer look at how the drift in the classifier leads to forgetting, and accordingly, design four simple yet (super-) effective solutions to alleviate the classifier drift: an Individual Classifiers with Frozen Feature Extractor (ICE) framework where we individually train a classifier for each learning session, and its three variants ICE-PL, ICE-O, and ICE-PL&O which further take the logits of previously learned classes from old sessions or a constant logit of an Other class as constraint to the learning of new classifiers. Extensive experiments and analysis on 6 class-incremental information extraction tasks demonstrate that our solutions, especially ICE-O, consistently show significant improvement over the previous state-of-the-art approaches with up to 44.7% absolute F-score gain, providing a strong baseline and insights for future research on class-incremental learning.

Text Augmented Open Knowledge Graph Completion via Pre-Trained Language Models

Pengcheng Jiang, Shivam Agarwal, Bowen Jin, Xuan Wang, Jimeng Sun and Jiawei Han 19:00-21:00 (Metropolitan Centre)

The mission of open knowledge graph (KG) completion is to draw new findings from known facts. Existing works that augment KG completion require either (1) factual triples to enlarge the graph reasoning space or (2) manually designed prompts to extract knowledge from a pre-trained language model (PLM), exhibiting limited performance and requiring expensive efforts from experts. To this end, we propose TagReal that automatically generates quality query prompts and retrieves support information from large text corpora to probe knowledge from PLM for KG completion. The results show that TagReal achieves state-of-the-art performance on two benchmark datasets. We find that TagReal has superb performance even with limited training data, outperforming existing embedding-based, graph-based, and PLM-based methods.

CoAug: Combining Augmentation of Labels and Labelling Rules

Rakesh R. Menon, Bingqing Wang, Jun Araki, Zhengyu Zhou, Zhe Feng and Liu Ren 19:00-21:00 (Metropolitan Centre)

Collecting labeled data for Named Entity Recognition (NER) tasks is challenging due to the high cost of manual annotations. Instead, researchers have proposed few-shot self-training and rule-augmentation techniques to minimize the reliance on large datasets. However, inductive biases and restricted logical language lexicon, respectively, can limit the ability of these models to perform well. In this work, we propose CoAug, a co-augmentation framework that allows us to improve few-shot models and rule-augmentation models by bootstrapping predictions from each model. By leveraging rules and neural model predictions to train our models, we complement the benefits of each and achieve the best of both worlds. In our experiments, we show that our best CoAug model can outperform strong weak-supervision-based NER models at least by 6.5 F1 points.

Generate then Select: Open-ended Visual Question Answering Guided by World Knowledge

Xingyu Fu, Sheng Zhang, Gukyeong Kwon, Pramuditha Perera, Henghui Zhu, Yuhao Zhang, Alexander Hanbo Li, William Yang Wang, Zhiguo Wang, Vittorio Castelli, Patrick Ng, Dan Roth and Bing Xiang 19:00-21:00 (Metropolitan Centre)

The open-ended Visual Question Answering (VQA) task requires AI models to jointly reason over visual and natural language inputs using world knowledge. Recently, pre-trained Language Models (PLM) such as GPT-3 have been applied to the task and shown to be powerful world knowledge sources. However, these methods suffer from low knowledge coverage caused by PLM bias – the tendency to generate certain tokens over other tokens regardless of prompt changes, and high dependency on the PLM quality – only models using GPT-3 can achieve the best result.

To address the aforementioned challenges, we propose RASO: a new VQA pipeline that deploys a generate-then-select strategy guided by world knowledge for the first time. Rather than following the de facto standard to train a multi-modal model that directly generates the VQA answer, {pasted macro 'MODEL'}name first adopts PLM to generate all the possible answers, and then trains a lightweight answer selection model for the correct answer. As proved in our analysis, RASO expands the knowledge coverage from in-domain training data by a large margin. We provide extensive experimentation and show the effectiveness of our pipeline by advancing the state-of-the-art by 4.1% on OK-VQA, without additional computation cost.

Modularized Zero-shot VQA with Pre-trained Models

RUI CAO and Jing Jiang 19:00-21:00 (Metropolitan Centre)

Large-scale pre-trained models (PTMs) show great zero-shot capabilities. In this paper, we study how to leverage them for zero-shot visual question answering (VQA). Our approach is motivated by a few observations. First, VQA questions often require multiple steps of reasoning, which is still a capability that most PTMs lack. Second, different steps in VQA reasoning chains require different skills such as object detection and relational reasoning, but a single PTM may not possess all these skills. Third, recent work on zero-shot VQA does not explicitly consider multi-step reasoning chains, which makes them less interpretable compared with a decomposition-based approach. We propose a modularized zero-shot network that explicitly decomposes questions into sub reasoning steps and is highly interpretable. We convert sub reasoning tasks to acceptable objectives of PTMs and assign tasks to proper PTMs without any adaptation. Our experiments on two VQA benchmarks under the zero-shot setting demonstrate the effectiveness of our method and better interpretability compared with several baselines.

Aerial Vision-and-Dialog Navigation

Yue Fan, Winson Chen, Tongzhou Jiang, Chun Zhou, Yi Zhang and Xin Eric Wang 19:00-21:00 (Metropolitan Centre)

The ability to converse with humans and follow natural language commands is crucial for intelligent unmanned aerial vehicles (a.k.a. drones).

It can relieve people's burden of holding a controller all the time, allow multitasking, and make drone control more accessible for people with disabilities or with their hands occupied. To this end, we introduce Aerial Vision-and-Dialog Navigation (AVDN), to navigate a drone via natural language conversation. We build a drone simulator with a continuous photorealistic environment and collect a new AVDN dataset of over 3k recorded navigation trajectories with asynchronous human-human dialogs between commanders and followers. The commander provides initial navigation instruction and further guidance by request, while the follower navigates the drone in the simulator and asks questions when needed. During data collection, followers' attention on the drone's visual observation is also recorded. Based on the AVDN dataset, we study the tasks of aerial navigation from (full) dialog history and propose an effective Human Attention Aided Transformer model (HAA-Transformer), which learns to predict both navigation waypoints and human attention.

Enhanced Chart Understanding via Visual Language Pre-training on Plot Table Pairs
Mingyang Zhou, Yi Fung, Long Chen, Christopher Thomas, Heng Ji and Shih-Fu Chang   19:00-21:00 (Metropolitan Centre)
Building cross-model intelligence that can understand charts and communicate the salient information hidden behind them is an appealing challenge in the vision and language (V+L) community. The capability to uncover the underlined table data of chart figures is a critical key to automatic chart understanding. We introduce ChartT5, a V+L model that learns how to interpret table information from chart images via cross-modal pre-training on plot table pairs. Specifically, we propose two novel pre-training objectives: Masked Header Prediction (MHP) and Masked Value Prediction (MVP) to facilitate the model with different skills to interpret the table information. We have conducted extensive experiments on chart question answering and chart summarization to verify the effectiveness of the proposed pre-training strategies. In particular, on the ChartQA benchmark, our ChartT5 outperforms the state-of-the-art non-pretraining methods by over 8% performance gains.

Towards Parameter-Efficient Integration of Pre-Trained Language Models In Temporal Video Grounding
Erica Kido Shimomoto, Edison Marrese-Taylor, Hiroya Takamura, Ichiro Kobayashi, Hideki Nakayama and Yusuke Miyao   19:00-21:00 (Metropolitan Centre)
This paper explores the task of Temporal Video Grounding (TVG) where, given an untrimmed video and a query sentence, the goal is to recognize and determine temporal boundaries of action instances in the video described by natural language queries. Recent works tackled this task by improving query inputs with large pre-trained language models (PLM), at the cost of more expensive training. However, the effects of this integration are unclear, as these works also propose improvements in the visual inputs. Therefore, this paper studies the role of query sentence representation with PLMs in TVG and assesses the applicability of parameter-efficient training with NLP adapters. We couple popular PLMs with a selection of existing approaches and test different adapters to reduce the impact of the additional parameters. Our results on three challenging datasets show that, with the same visual inputs, TVG models greatly benefited from the PLM integration and fine-tuning, stressing the importance of the text query representation in this task. Furthermore, adapters were an effective alternative to full fine-tuning, even though they are not tailored to our task, allowing PLM integration in larger TVG models and delivering results comparable to SOTA models. Finally, our results shed light on which adapters work best in different scenarios.

LMCap: Few-shot Multilingual Image Captioning by Retrieval Augmented Language Model Prompting
Rita Parada Ramos, Bruno Martins and Desmond Elliott   19:00-21:00 (Metropolitan Centre)
Multilingual image captioning has recently been tackled by training with large-scale machine translated data, which is an expensive, noisy, and time-consuming process. Without requiring any multilingual caption data, we propose LMCap, an image-blind few-shot multilingual captioning model that works by prompting a language model with retrieved captions. Specifically, instead of following the standard encoder-decoder paradigm, given an image, LMCap first retrieves the captions of similar images using a multilingual CLIP encoder. These captions are then combined into a prompt for an XGLM decoder, in order to generate captions in the desired language. In other words, the generation model does not directly process the image, instead it processes retrieved captions. Experiments on the XM3600 dataset of geographically diverse images show that our model is competitive with fully-supervised multilingual captioning models, without requiring any supervised training on any captioning data.

Speaking the Language of Your Listener: Audience-Aware Adaptation via Plug-and-Play Theory of Mind
Ece Takmaz, Nicolo' Brandizzi, Mario Giulianelli, Sandro Pezzelle and Raquel Fernandez   19:00-21:00 (Metropolitan Centre)
Dialogue participants may have varying levels of knowledge about the topic under discussion. In such cases, it is essential for speakers to adapt their utterances by taking their audience into account. Yet, it is an open question how such adaptation can be modelled in computational agents. In this paper, we model a visually grounded referential game between a knowledgeable speaker and a listener with more limited visual and linguistic experience. Inspired by psycholinguistic theories, we endow our speaker with the ability to adapt its referring expressions via a simulation module that monitors the effectiveness of planned utterances from the listener's perspective. We propose an adaptation mechanism building on plug-and-play approaches to controlled language generation, where utterance generation is steered on the fly by the simulator without finetuning the speaker's underlying language model. Our results and analyses show that our approach is effective: the speaker's utterances become closer to the listener's domain of expertise, which leads to higher communicative success.

Visual Coherence Loss for Coherent and Visually Grounded Story Generation
Xudong Hong, Vera Demberg, Asad Sayeed, Qiankun Zheng and Bernt Schiele   19:00-21:00 (Metropolitan Centre)
Local coherence is essential for long-form text generation models. We identify two important aspects of local coherence within the visual storytelling task: (1) the model needs to represent re-occurrences of characters within the image sequence in order to mention them correctly in the story; (2) character representations should enable us to find instances of the same characters and distinguish different characters. In this paper, we propose a loss function inspired by a linguistic theory of coherence for self-supervised learning for image sequence representations. We further propose combining features from an object and a face detector to construct stronger character features. To evaluate input-output relevance that current reference-based metrics don't measure, we propose a character matching metric to check whether the models generate referring expressions correctly for characters in input image sequences. Experiments on a visual story generation dataset show that our proposed features and loss function are effective for generating more coherent and visually grounded stories.

I Spy a Metaphor: Large Language Models and Diffusion Models Co-Create Visual Metaphors
Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki and Smaranda Muresan   19:00-21:00 (Metropolitan Centre)
Visual metaphors are powerful rhetorical devices used to persuade or communicate creative ideas through images. Similar to linguistic metaphors, they convey meaning implicitly through symbolism and juxtaposition of the symbols. We propose a new task of generating visual metaphors from linguistic metaphors. This is a challenging task for diffusion-based text-to-image models, such as DALL·E 2, since it requires the ability to model implicit meaning and compositionality. We propose to solve the task through the collaboration between Large Language Models (LLMs) and Diffusion Models: Instruct GPT-3 (davinci-002) with Chain-of-Thought prompting generates text that represents a visual elaboration of the linguistic metaphor containing the implicit meaning and relevant objects, which is then used as input to the diffusion-based text-to-image models. Using a human-AI collaboration framework, where humans interact both with the LLM and the top-performing diffusion model, we create a high-quality dataset containing 6,476 visual metaphors for 1,540 linguistic metaphors and their associated visual elaborations. Evaluation by professional illustrators shows the promise of LLM-Diffusion Model collaboration for this task.To evaluate the utility of our Human-AI collaboration framework and the quality of our dataset, we perform both an intrinsic human-based evaluation and an

extrinsic evaluation using visual entailment as a downstream task.

Multimedia Generative Script Learning for Task Planning

Qingyun Wang, Manling Li, Hou Pong Chan, Lifu Huang, Julia Hockenmaier, Girish Chowdhary and Heng Ji 19:00-21:00 (Metropolitan Centre)
Goal-oriented generative script learning aims to generate subsequent steps to reach a particular goal, which is an essential task to assist robots or humans in performing stereotypical activities. An important aspect of this process is the ability to capture historical states visually, which provides detailed information that is not covered by text and will guide subsequent steps. Therefore, we propose a new task, Multimedia Generative Script Learning, to generate subsequent steps by tracking historical states in both text and vision modalities, as well as presenting the first benchmark containing 5,652 tasks and 79,089 multimedia steps. This task is challenging in three aspects: the multimedia challenge of capturing the visual states in images, the induction challenge of performing unseen tasks, and the diversity challenge of covering different information in individual steps. We propose to encode visual state changes through a selective multimedia encoder to address the multimedia challenge, transfer knowledge from previously observed tasks using a retrieval-augmented decoder to overcome the induction challenge, and further present distinct information at each step by optimizing a diversity-oriented contrastive learning objective. We define metrics to evaluate both generation and inductive quality. Experiment results demonstrate that our approach significantly outperforms strong baselines.

Listen, Decipher and Sign: Toward Unsupervised Speech-to-Sign Language Recognition

Liming Wang, Junrui Ni, Heting Gao, Jialu Li, Kai Chieh Chang, Xulin Fan, Junkai Wu, Mark Hasegawa-Johnson and Chang D. Yoo 19:00-21:00 (Metropolitan Centre)
Existing supervised sign language recognition systems rely on an abundance of well-annotated data. Instead, an unsupervised speech-to-sign language recognition (SSR-U) system learns to translate between spoken and sign languages by observing only non-parallel speech and sign-language corpora. We propose speech2sign-U, a neural network-based approach capable of both character-level and word-level SSR-U. Our approach significantly outperforms baselines directly adapted from unsupervised speech recognition (ASR-U) models by as much as 50% recall@10 on several challenging American sign language corpora with various levels of sample sizes, vocabulary sizes, and audio and visual variability. The code is available at https://github.com/cactuswiththoughts/UnsupSpeech2Sign.gitcactuswiththoughts/UnsupSpeech2Sign.git.

Joint Speech Transcription and Translation: Pseudo-Labeling with Out-of-Distribution Data

Mozhdeh Gheini, Tatiana Likhomanenko, Matthias Sperber and Hendra Setiawan 19:00-21:00 (Metropolitan Centre)
Self-training has been shown to be helpful in addressing data scarcity for many domains, including vision, speech, and language. Specifically, self-training, or pseudo-labeling, labels unsupervised data and adds that to the training pool. In this work, we investigate and use pseudo-labeling for a recently proposed novel setup: joint transcription and translation of speech, which suffers from an absence of sufficient parallel data resources. We show that under such data-deficient circumstances, the unlabeled data can significantly vary in domain from the supervised data, which results in pseudo-label quality degradation. We investigate two categories of remedies that require no additional supervision and target the domain mismatch: pseudo-label filtering and data augmentation. We show that pseudo-label analysis and processing in this way results in additional gains on top of the vanilla pseudo-labeling setup providing a total improvement of up to 0.4% absolute WER and 2.1 BLEU points for En–De and 0.6% absolute WER and 2.2 BLEU points for En–Zh.

Visually-Enhanced Phrase Understanding

Tsu-Yuan Hsu, Chen-An Li, Chao-Wei Huang and Yun-Nung Chen 19:00-21:00 (Metropolitan Centre)
Large-scale vision-language pre-training has exhibited strong performance in various visual and textual understanding tasks. Recently, the textual encoders of multi-modal pre-trained models have been shown to generate high-quality textual representations, which often outperform models that are purely text-based, such as BERT. In this study, our objective is to utilize both textual and visual encoders of multi-modal pre-trained models to enhance language understanding tasks. We achieve this by generating an image associated with a textual prompt, thus enriching the representation of a phrase for downstream tasks. Results from experiments conducted on four benchmark datasets demonstrate that our proposed method, which leverages visually-enhanced text representations, significantly improves performance in the entity clustering task.

FastDiff 2: Revisiting and Incorporating GANs and Diffusion Models in High-Fidelity Speech Synthesis

Rongjie Huang, Yi Ren, Ziyue Jiang, Chenye Cui, Jinglin Liu and Zhou Zhao 19:00-21:00 (Metropolitan Centre)
Generative adversarial networks (GANs) and denoising diffusion probabilistic models (DDPMs) have recently achieved impressive performances in image and audio synthesis. After revisiting their success in conditional speech synthesis, we find that 1) GANs sacrifice sample diversity for quality and speed, 2) diffusion models exhibit outperformed sample quality and diversity at a high computational cost, where achieving high-quality, fast, and diverse speech synthesis challenges all neural synthesizers. In this work, we propose to converge advantages from GANs and diffusion models by incorporating both classes, introducing dual-empowered modeling perspectives: 1) FastDiff 2 (Diff-GAN), a diffusion model whose denoising process is parametrized by conditional GANs, and the non-Gaussian denoising distribution makes it much more stable to implement the reverse process with large steps sizes; and 2) FastDiff 2 (GANDiff), a generative adversarial network whose forward process is constructed by multiple denoising diffusion iterations, which exhibits better sample diversity than traditional GANs. Experimental results show that both variants enjoy an efficient 4-step sampling process and demonstrate superior sample quality and diversity. Audio samples are available at https://RevisitSpeech.github.io/

Prosody-TTS: Improving Prosody with Masked Autoencoder and Conditional Diffusion Model For Expressive Text-to-Speech

Rongjie Huang, Chunlei Zhang, Yi Ren, Zhou Zhao and Dong Yu 19:00-21:00 (Metropolitan Centre)
Expressive text-to-speech aims to generate high-quality samples with rich and diverse prosody, which is hampered by **dual challenges**: 1) prosodic attributes in highly dynamic voices are difficult to capture and model without intonation; and 2) highly multimodal prosodic representations cannot be well learned by simple regression (e.g., MSE) objectives, which causes blurry and over-smoothing predictions. This paper proposes Prosody-TTS, a two-stage pipeline that enhances **prosody modeling and sampling** by introducing several components: 1) a self-supervised masked autoencoder to model the prosodic representation without relying on text transcriptions or local prosody attributes, which ensures to cover diverse speaking voices with superior generalization; and 2) a diffusion model to sample diverse prosodic patterns within the latent space, which prevents TTS models from generating samples with dull prosodic performance. Experimental results show that Prosody-TTS achieves new state-of-the-art in text-to-speech with natural and expressive synthesis. Both subjective and objective evaluation demonstrate that it exhibits superior audio quality and prosody naturalness with rich and diverse prosodic attributes. Audio samples are available at https://improved_prosody.github.io

Findings Spotlights III

19:00-21:00 (Metropolitan West)

Subword Segmental Machine Translation: Unifying Segmentation and Target Sentence Generation
Francois Meyer and Jan Buys   19:00-21:00 (Metropolitan West)
Subword segmenters like BPE operate as a preprocessing step in neural machine translation and other (conditional) language models. They are applied to datasets before training, so translation or text generation quality relies on the quality of segmentations. We propose a departure from this paradigm, called subword segmental machine translation (SSMT). SSMT unifies subword segmentation and MT in a single trainable model. It learns to segment target sentence words while jointly learning to generate target sentences. To use SSMT during inference we propose dynamic decoding, a text generation algorithm that adapts segmentations as it generates translations. Experiments across 6 translation directions show that SSMT improves chrF scores for morphologically rich agglutinative languages. Gains are strongest in the very low-resource scenario. SSMT also learns subwords that are closer to morphemes compared to baselines and proves more robust on a test set constructed for evaluating morphological compositional generalisation.

An Investigation of Noise in Morphological Inflection
Adam Wiemerslage, Changbing Yang, Garrett Nicolai, Miikka Silfverberg and Katharina Kann   19:00-21:00 (Metropolitan West)
With a growing focus on morphological inflection systems for languages where high-quality data is scarce, training data noise is a serious but so far largely ignored concern. We aim at closing this gap by investigating the types of noise encountered within a pipeline for truly unsupervised morphological paradigm completion and its impact on morphological inflection systems: First, we propose an error taxonomy and annotation pipeline for inflection training data. Then, we compare the effect of different types of noise on multiple state-of-the- art inflection models. Finally, we propose a novel character-level masked language modeling (CMLM) pretraining objective and explore its impact on the models' resistance to noise. Our experiments show that various architectures are impacted differently by separate types of noise, but encoder-decoders tend to be more robust to noise than models trained with a copy bias. CMLM pretraining helps transformers, but has lower impact on LSTMs.

Pre-Trained Language-Meaning Models for Multilingual Parsing and Generation
Chunliu Wang, Huiyuan Lai, Malvina Nissim and Johan Bos   19:00-21:00 (Metropolitan West)
Pre-trained language models (PLMs) have achieved great success in NLP and have recently been used for tasks in computational semantics. However, these tasks do not fully benefit from PLMs since meaning representations are not explicitly included. We introduce multilingual pre-trained language-meaning models based on Discourse Representation Structures (DRSs), including meaning representations besides natural language texts in the same model, and design a new strategy to reduce the gap between the pre-training and fine-tuning objectives. Since DRSs are language neutral, cross-lingual transfer learning is adopted to further improve the performance of non-English tasks. Automatic evaluation results show that our approach achieves the best performance on both the multilingual DRS parsing and DRS-to-text generation tasks. Correlation analysis between automatic metrics and human judgements on the generation task further validates the effectiveness of our model. Human inspection reveals that out-of-vocabulary tokens are the main cause of erroneous results.

Unsupervised Mapping of Arguments of Deverbal Nouns to Their Corresponding Verbal Labels
Aviv Weinstein and Yoav Goldberg   19:00-21:00 (Metropolitan West)
Deverbal nouns are nominal forms of verbs commonly used in written English texts to describe events or actions, as well as their arguments. However, many NLP systems, and in particular pattern-based ones, neglect to handle such nominalized constructions. The solutions that do exist for handling arguments of nominalized constructions are based on semantic annotation and require semantic ontologies, making their applications restricted to a small set of nouns. We propose to adopt instead a more syntactic approach, which maps the arguments of deverbal nouns to the universal-dependency relations of the corresponding verbal construction. We present an unsupervised mechanism—based on contextualized word representations—which allows to enrich universal-dependency trees with dependency arcs denoting arguments of deverbal nouns, using the same labels as the corresponding verbal cases. By sharing the same label set as in the verbal case, patterns that were developed for verbs can be applied without modification but with high accuracy also to the nominal constructions.

DMLM: Descriptive Masked Language Modeling
Edoardo Barba, Niccolò Campolungo and Roberto Navigli   19:00-21:00 (Metropolitan West)
Over the last few years, Masked Language Modeling (MLM) pre-training has resulted in remarkable advancements in many Natural Language Understanding (NLU) tasks, which sparked an interest in researching alternatives and extensions to the MLM objective. In this paper, we tackle the absence of explicit semantic grounding in MLM and propose Descriptive Masked Language Modeling (DMLM), a knowledge-enhanced reading comprehension objective, where the model is required to predict the most likely word in a context, being provided with the word's definition. For instance, given the sentence "I was going to the _", if we provided as definition "financial institution", the model would have to predict the word "bank"; if, instead, we provided "sandy seashore", the model should predict "beach". Our evaluation highlights the effectiveness of DMLM in comparison with standard MLM, showing improvements on a number of well-established NLU benchmarks, as well as other semantics-focused tasks, e.g., Semantic Role Labeling. Furthermore, we demonstrate how it is possible to take full advantage of DMLM to embed explicit semantics in downstream tasks, explore several properties of DMLM-based contextual representations and suggest a number of future directions to investigate.

Ambiguity Meets Uncertainty: Investigating Uncertainty Estimation for Word Sense Disambiguation
Zhu Liu and Ying Liu   19:00-21:00 (Metropolitan West)
Word sense disambiguation (WSD), which aims to determine an appropriate sense for a target word given its context, is crucial for natural language understanding. Existing supervised methods treat WSD as a classification task and have achieved remarkable performance. However, they ignore uncertainty estimation (UE) in the real-world setting, where the data is always noisy and out of distribution. This paper extensively studies UE on the benchmark designed for WSD. Specifically, we first compare four uncertainty scores for a state-of-the-art WSD model and verify that the conventional predictive probabilities obtained at the end of the model are inadequate to quantify uncertainty. Then, we examine the capability of capturing data and model uncertainties by the model with the selected UE score on well-designed test scenarios and discover that the model reflects data uncertainty satisfactorily but underestimates model uncertainty. Furthermore, we explore numerous lexical properties that intrinsically affect data uncertainty and provide a detailed analysis of four critical aspects: the syntactic category, morphology, sense granularity, and semantic relations.

A Self-Supervised Integration Method of Pretrained Language Models and Word Definitions
Hwiyeol Jo   19:00-21:00 (Metropolitan West)
We investigate the representation of pretrained language models and humans, using the idea of word definition modeling–how well a word is represented by its definition, and vice versa. Our analysis shows that a word representation in pretrained language models does not successfully map its human-written definition and its usage in example sentences. We then present a simple method DefBERT that integrates pretrained models with word semantics in dictionaries. We show its benefits on newly-proposed tasks of definition ranking and definition sense disambiguation. Furthermore, we present the results on standard word similarity tasks and short text classification tasks where models are required to encode semantics with only a few words. The results demonstrate the effectiveness of integrating word definitions and pretrained language models.

Unsupervised Paraphrasing of Multiword Expressions

Takashi Wada, Yuji Matsumoto, Timothy Baldwin and Jey Han Lau   19:00-21:00 (Metropolitan West)
We propose an unsupervised approach to paraphrasing multiword expressions (MWEs) in context. Our model employs only monolingual corpus data and pre-trained language models (without fine-tuning), and does not make use of any external resources such as dictionaries. We evaluate our method on the SemEval 2022 idiomatic semantic text similarity task, and show that it outperforms all unsupervised systems and rivals supervised systems.

Together We Make Sense–Learning Meta-Sense Embeddings

Haochen Luo, Yi Zhou and Danushka Bollegala   19:00-21:00 (Metropolitan West)
Sense embedding learning methods learn multiple vectors for a given ambiguous word, corresponding to its different word senses. For this purpose, different methods have been proposed in prior work on sense embedding learning that use different sense inventories, sense-tagged corpora and learning methods. However, not all existing sense embeddings cover all senses of ambiguous words equally well due to the discrepancies in their training resources. To address this problem, we propose the first-ever meta-sense embedding method – Neighbour Preserving Meta-Sense Embeddings, which learns meta-sense embeddings by combining multiple independently trained source sense embeddings such that the sense neighbourhoods computed from the source embeddings are preserved in the meta-embedding space. Our proposed method can combine source sense embeddings that cover different sets of word senses. Experimental results on Word Sense Disambiguation (WSD) and Word-in-Context (WiC) tasks show that the proposed meta-sense embedding method consistently outperforms several competitive baselines.An anonymised version of the source code implementation for our proposed method is submitted to reviewing system. Both source code and the learnt meta-sense embeddings will be publicly released upon paper acceptance.

Solving Cosine Similarity Underestimation between High Frequency Words by ℓ_2 Norm Discounting

Saeth Wannasuphoprasit, Yi Zhou and Danushka Bollegala   19:00-21:00 (Metropolitan West)
Cosine similarity between two words, computed using their contextualised token embeddings obtained from masked language models (MLMs) such as BERT has shown to underestimate the actual similarity between those words CITATION. This similarity underestimation problem is particularly severe for high frequent words. Although this problem has been noted in prior work, no solution has been proposed thus far. We observe that the ℓ_2 norm of contextualised embeddings of a word correlates with its log-frequency in the pretraining corpus. Consequently, the larger ℓ_2 norms associated with the high frequent words reduce the cosine similarity values measured between them, thus underestimating the similarity scores. To solve this issue, we propose a method to *discount* the ℓ_2 norm of a contextualised word embedding by the frequency of that word in a corpus when measuring the cosine similarities between words. We show that the so called *stop* words behave differently from the rest of the words, which require special consideration during their discounting process. Experimental results on a contextualised word similarity dataset show that our proposed discounting method accurately solves the similarity underestimation problem.An anonymized version of the source code of our proposed method is submitted to the reviewing system.

Improving Diachronic Word Sense Induction with a Nonparametric Bayesian method

Ashjan Alsulaimani and Erwan Moreau   19:00-21:00 (Metropolitan West)
Diachronic Word Sense Induction (DWSI) is the task of inducing the temporal representations of a word meaning from the context, as a set of senses and their prevalence over time. We introduce two new models for DWSI, based on topic modelling techniques: one is based on Hierarchical Dirichlet Processes (HDP), a nonparametric model; the other is based on the Dynamic Embedded Topic Model (DETM), a recent dynamic neural model. We evaluate these models against two state of the art DWSI models, using a time-stamped labelled dataset from the biomedical domain. We demonstrate that the two proposed models perform better than the state of the art. In particular, the HDP-based model drastically outperforms all the other models, including the dynamic neural model.

On Isotropy, Contextualization and Learning Dynamics of Contrastive-based Sentence Representation Learning

Chenghao Xiao, Yang Long and Noura Al Moubayed   19:00-21:00 (Metropolitan West)
Incorporating contrastive learning objectives in sentence representation learning (SRL) has yielded significant improvements on many sentence-level NLP tasks. However, it is not well understood why contrastive learning works for learning sentence-level semantics. In this paper, we aim to help guide future designs of sentence representation learning methods by taking a closer look at contrastive SRL through the lens of isotropy, contextualization and learning dynamics. We interpret its successes through the geometry of the representation shifts and show that contrastive learning brings isotropy, and drives high intra-sentence similarity: when in the same sentence, tokens converge to similar positions in the semantic space. We also find that what we formalize as "spurious contextualization" is mitigated for semantically meaningful tokens, while augmented for functional ones. We find that the embedding space is directed towards the origin during training, with more areas now better defined. We ablate these findings by observing the learning dynamics with different training temperatures, batch sizes and pooling methods.

Exploring Non-Verbal Predicates in Semantic Role Labeling: Challenges and Opportunities

Riccardo Orlando, Simone Conia and Roberto Navigli   19:00-21:00 (Metropolitan West)
Although we have witnessed impressive progress in Semantic Role Labeling (SRL), most of the research in the area is carried out assuming that the majority of predicates are verbs. Conversely, predicates can also be expressed using other parts of speech, e.g., nouns and adjectives. However, non-verbal predicates appear in the benchmarks we commonly use to measure progress in SRL less frequently than in some real-world settings – newspaper headlines, dialogues, and tweets, among others. In this paper, we put forward a new PropBank dataset which boasts wide coverage of multiple predicate types. Thanks to it, we demonstrate empirically that standard benchmarks do not provide an accurate picture of the current situation in SRL and that state-of-the-art systems are still incapable of transferring knowledge across different predicate types. Having observed these issues, we also present a novel, manually-annotated challenge set designed to give equal importance to verbal, nominal, and adjectival predicate-argument structures. We use such dataset to investigate whether we can leverage different linguistic resources to promote knowledge transfer. In conclusion, we claim that SRL is far from "solved", and its integration with other semantic tasks might enable significant improvements in the future, especially for the long tail of non-verbal predicates, thereby facilitating further research on SRL for non-verbal predicates. We release our software and datasets at https://github.com/sapienzanlp/exploring-srl.

Incorporating Graph Information in Transformer-based AMR Parsing

Pavlo Vasylenko, Pere Lluís Huguet Cabot, Abelardo Carlos Martínez Lorenzo and Roberto Navigli   19:00-21:00 (Metropolitan West)
Abstract Meaning Representation (AMR) is a Semantic Parsing formalism that aims at providing a semantic graph abstraction representing a given text. Current approaches are based on autoregressive language models such as BART or T5, fine-tuned through Teacher Forcing to obtain a linearized version of the AMR graph from a sentence. In this paper, we present LeakDistill, a model and method that explores a modification to the Transformer architecture, using structural adapters to explicitly incorporate graph information into the learned representations and improve AMR parsing performance. Our experiments show how, by employing word-to-node alignment to embed graph structural information into the encoder at training time, we can obtain state-of-the-art AMR parsing through self-knowledge distillation, even without the use of additional data. We release the code at http://www.github.com/sapienzanlp/LeakDistill.

Align-then-Enhance: Multilingual Entailment Graph Enhancement with Soft Predicate Alignment
Yuting Wu, Yutong Hu, Yansong Feng, Tianyi Li, Mark Steedman and Dongyan Zhao   19:00-21:00 (Metropolitan West)
Entailment graphs (EGs) with predicates as nodes and entailment relations as edges are typically incomplete, while EGs in different languages are often complementary to each other. In this paper, we propose a new task, multilingual entailment graph enhancement, which aims to utilize the entailment information from one EG to enhance another EG in a different language. The ultimate goal is to obtain an enhanced EG containing richer and more accurate entailment information. We present an align-then-enhance framework (ATE) to achieve accurate multilingual entailment graph enhancement, which first exploits a cross-graph guided interaction mechanism to automatically discover potential equivalent predicates between different EGs and then constructs more accurate enhanced entailment graphs based on soft predicate alignments. Extensive experiments show that ATE achieves better and more robust predicate alignment results between different EGs, and the enhanced entailment graphs generated by ATE outperform the original graphs for entailment detection.

Unsupervised Semantic Variation Prediction using the Distribution of Sibling Embeddings
Taichi Aida and Danushka Bollegala   19:00-21:00 (Metropolitan West)
Languages are dynamic entities, where the meanings associated with words constantly change with time. Detecting the semantic variation of words is an important task for various NLP applications that must make time-sensitive predictions. Existing work on semantic variation prediction have predominantly focused on comparing some form of an averaged contextualised representation of a target word computed from a given corpus. However, some of the previously associated meanings of a target word can become obsolete over time (e.g. meaning of gay as happy), while novel usages of existing words are observed (e.g. meaning of cell as a mobile phone). We argue that mean representations alone cannot accurately capture such semantic variations and propose a method that uses the entire cohort of the contextualised embeddings of the target word, which we refer to as the sibling distribution. Experimental results on SemEval-2020 Task 1 benchmark dataset for semantic variation prediction show that our method outperforms prior work that consider only the mean embeddings, and is comparable to the current state-of-the-art. Moreover, a qualitative analysis shows that our method detects important semantic changes in words that are not captured by the existing methods.

Categorial grammar induction from raw data
Christian Clark and William Schuler   19:00-21:00 (Metropolitan West)
Grammar induction, the task of learning a set of grammatical rules from raw or minimally labeled text data, can provide clues about what kinds of syntactic structures are learnable without prior knowledge. Recent work (e.g., Kim et al., 2019; Zhu et al., 2020; Jin et al., 2021a) has achieved advances in unsupervised induction of probabilistic context-free grammars (PCFGs). However, categorial grammar induction has received less recent attention, despite allowing inducers to support a larger set of syntactic categories—due to restrictions on how categories can combine—and providing a transparent interface with compositional semantics, opening up possibilities for models that jointly learn form and meaning. Motivated by this, we propose a new model for inducing a basic (Ajdukiewicz, 1935; Bar-Hillel, 1953) categorial grammar. In contrast to earlier categorial grammar induction systems (e.g., Bisk and Hockenmaier, 2012), our model learns from raw data without any part-of-speech information. Experiments on child-directed speech show that our model attains a recall-homogeneity of 0.33 on average, which dramatically increases to 0.59 when a bias toward forward function application is added to the model.

Language acquisition: do children and language models follow similar learning stages?
Linnea Evanson, Yair Lakretz and Jean Rémi king   19:00-21:00 (Metropolitan West)
During language acquisition, children follow a typical sequence of learning stages, whereby they first learn to categorize phonemes before they develop their lexicon and eventually master increasingly complex syntactic structures. However, the computational principles that lead to this learning trajectory remain largely unknown. To investigate this, we here compare the learning trajectories of deep language models to those of human children. Specifically, we test whether, during its training, GPT-2 exhibits stages of language acquisition comparable to those observed in children aged between 18 months and 6 years. For this, we train 48 GPT-2 models from scratch and evaluate their syntactic and semantic abilities at each training step, using 96 probes curated from the BLiMP, Zorro and BIG-Bench benchmarks. We then compare these evaluations with the behavior of 54 children during language production. Our analyses reveal three main findings. First, similarly to children, the language models tend to learn linguistic skills in a systematic order. Second, this learning scheme is parallel: the language tasks that are learned last improve from the very first training steps. Third, some – but not all – learning stages are shared between children and these language models. Overall, these results shed new light on the principles of language acquisition, and highlight important divergences in how humans and modern algorithms learn to process natural language.

Automatic Readability Assessment for Closely Related Languages
Joseph Marvin Imperial and Ekaterina Kochmar   19:00-21:00 (Metropolitan West)
In recent years, the main focus of research on automatic readability assessment (ARA) has shifted towards using expensive deep learning-based methods with the primary goal of increasing models' accuracy. This, however, is rarely applicable for low-resource languages where traditional handcrafted features are still widely used due to the lack of existing NLP tools to extract deeper linguistic representations. In this work, we take a step back from the technical component and focus on how linguistic aspects such as mutual intelligibility or degree of language relatedness can improve ARA in a low-resource setting. We collect short stories written in three languages in the Philippines—Tagalog, Bikol, and Cebuano—to train readability assessment models and explore the interaction of data and features in various cross-lingual setups. Our results show that the inclusion of CrossNGO, a novel specialized feature exploiting n-gram overlap applied to languages with high mutual intelligibility, significantly improves the performance of ARA models compared to the use of off-the-shelf large multilingual language models alone. Consequently, when both linguistic representations are combined, we achieve state-of-the-art results for Tagalog and Cebuano, and baseline scores for ARA in Bikol.

$2 * n$ is better than n^2: Decomposing Event Coreference Resolution into Two Tractable Problems
Shafiuddin Rehan Ahmed, Abhijnan Nath, James H. Martin and Nikhil Krishnaswamy   19:00-21:00 (Metropolitan West)
Event Coreference Resolution (ECR) is the task of linking mentions of the same event either within or across documents. Most mention pairs are not coreferent, yet many that are coreferent can be identified through simple techniques such as lemma matching of the event triggers or the sentences in which they appear. Existing methods for training coreference systems sample from a largely skewed distribution, making it difficult for the algorithm to learn coreference beyond surface matching. Additionally, these methods are intractable because of the quadratic operations needed. To address these challenges, we break the problem of ECR into two parts: a) a heuristic to efficiently filter out a large number of non-coreferent pairs, and b) a training approach on a balanced set of coreferent and non-coreferent mention pairs. By following this approach, we show that we get comparable results to the state of the art on two popular ECR datasets while significantly reducing compute requirements. We also analyze the mention pairs that are "hard" to accurately classify as coreferent or non-coreferentcode repo: `github.com/ahmeshaf/lemma_ce_coref`.

$2 * n$ is better than n^2: Decomposing Event Coreference Resolution into Two Tractable Problems
Shafiuddin Rehan Ahmed, Abhijnan Nath, James H. Martin and Nikhil Krishnaswamy   19:00-21:00 (Metropolitan West)
Event Coreference Resolution (ECR) is the task of linking mentions of the same event either within or across documents. Most mention pairs are not coreferent, yet many that are coreferent can be identified through simple techniques such as lemma matching of the event triggers or

the sentences in which they appear. Existing methods for training coreference systems sample from a largely skewed distribution, making it difficult for the algorithm to learn coreference beyond surface matching. Additionally, these methods are intractable because of the quadratic operations needed. To address these challenges, we break the problem of ECR into two parts: a) a heuristic to efficiently filter out a large number of non-coreferent pairs, and b) a training approach on a balanced set of coreferent and non-coreferent mention pairs. By following this approach, we show that we get comparable results to the state of the art on two popular ECR datsets while significantly reducing compute requirements. We also analyze the mention pairs that are "hard" to accurately classify as coreferent or non-coreferentcode repo: github.com/ahmeshaf/lemma_ce_coref.

How Well Do Large Language Models Perform on Faux Pas Tests?
Natalie Shapira, Guy Zwirn and Yoav Goldberg   19:00-21:00 (Metropolitan West)
Motivated by the question of the extent to which large language models "understand" social intelligence, we investigate the ability of such models to generate correct responses to questions involving descriptions of faux pas situations. The faux pas test is a test used in clinical psychology, which is known to be more challenging for children than individual tests of theory-of-mind or social intelligence. Our results demonstrate that, while the models seem to sometimes offer correct responses, they in fact struggle with this task, and that many of the seemingly correct responses can be attributed to over-interpretation by the human reader ("the ELIZA effect"). An additional phenomenon observed is the failure of most models to generate a correct response to presupposition questions. Finally, in an experiment in which the models are tasked with generating original faux pas stories, we find that while some models are capable of generating novel faux pas stories, the stories are all explicit, as the models are limited in their abilities to describe situations in an implicit manner.

Distinguishing Address vs. Reference Mentions of Personal Names in Text
Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, Melissa Ferguson and Stav Atir   19:00-21:00 (Metropolitan West)
Detecting named entities in text has long been a core NLP task. However, not much work has gone into distinguishing whether an entity mention is addressing the entity vs. referring to the entity; e.g., *John, would you turn the light off?* vs. *John turned the light off*. While this distinction is marked by a *vocative case* marker in some languages, many modern Indo-European languages such as English do not use such explicit vocative markers, and the distinction is left to be interpreted in context. In this paper, we present a new annotated dataset that captures the *address* vs. *reference* distinction in English, an automatic tagger that performs at 85% accuracy in making this distinction, and demonstrate how this distinction is important in NLP and computational social science applications in English language.

PragmatiCQA: A Dataset for Pragmatic Question Answering in Conversations
Peng Qi, Nina Du, Christopher D. Manning and Jing Huang   19:00-21:00 (Metropolitan West)
Pragmatic reasoning about another speaker's unspoken intent and state of mind is crucial to efficient and effective human communication. It is virtually omnipresent in conversations between humans, e.g., when someone asks "do you have a minute?", instead of interpreting it literally as a query about your schedule, you understand that the speaker might have requests that take time, and respond accordingly. In this paper, we present PragmatiCQA, the first large-scale open-domain question answering (QA) dataset featuring 6873 QA pairs that explores pragmatic reasoning in conversations over a diverse set of topics. We designed innovative crowdsourcing mechanisms for interest-based and task-driven data collection to address the common issue of incentive misalignment between crowdworkers and potential users. To compare computational models' capability at pragmatic reasoning, we also propose several quantitative metrics to evaluate question answering systems on Pragmat-iCQA. We find that state-of-the-art systems still struggle to perform human-like pragmatic reasoning, and highlight their limitations for future research.

A Match Made in Heaven: A Multi-task Framework for Hyperbole and Metaphor Detection
Naveen Badathala, Abisek Rajakumar Kalarani, Tejpalsingh Siledar and Pushpak Bhattacharyya   19:00-21:00 (Metropolitan West)
Hyperbole and metaphor are common in day-to-day communication (e.g., "I am in deep trouble": how does trouble have depth?), which makes their detection important, especially in a conversational AI setting. Existing approaches to automatically detect metaphor and hyperbole have studied these language phenomena independently, but their relationship has hardly, if ever, been explored computationally. In this paper, we propose a multi-task deep learning framework to detect hyperbole and metaphor simultaneously. We hypothesize that metaphors help in hyperbole detection, and vice-versa. To test this hypothesis, we annotate two hyperbole datasets- HYPO and HYPO-L- with metaphor labels. Simultaneously, we annotate two metaphor datasets- TroFi and LCC- with hyperbole labels. Experiments using these datasets give an improvement of the state of the art of hyperbole detection by 12%. Additionally, our multi-task learning (MTL) approach shows an improvement of up to 17% over single-task learning (STL) for both hyperbole and metaphor detection, supporting our hypothesis. To the best of our knowledge, ours is the first demonstration of computational leveraging of linguistic intimacy between metaphor and hyperbole, leading to showing the superiority of MTL over STL for hyperbole and metaphor detection.

Towards Generative Event Factuality Prediction
John Murzaku, Tyler G Osborne, Amittai F Aviram and Owen Rambow   19:00-21:00 (Metropolitan West)
We present a novel end-to-end generative task and system for predicting event factuality holders, targets, and their associated factuality values. We perform the first experiments using all sources and targets of factuality statements from the FactBank corpus. We perform multi-task learning with other tasks and event-factuality corpora to improve on the FactBank source and target task. We argue that careful domain specific target text output format in generative systems is important and verify this with multiple experiments on target text output structure. We redo previous state-of-the-art author-only event factuality experiments and also offer insights towards a generative paradigm for the author-only event factuality prediction task.

Discourse Analysis via Questions and Answers: Parsing Dependency Structures of Questions Under Discussion
Wei-Jen Ko, Yating Wu, Cutter J Dalton, Dananjay T Srinivas, Greg Durrett and Junyi Jessy Li   19:00-21:00 (Metropolitan West)
Automatic discourse processing is bottlenecked by data: current discourse formalisms pose highly demanding annotation tasks involving large taxonomies of discourse relations, making them inaccessible to lay annotators. This work instead adopts the linguistic framework of Questions Under Discussion (QUD) for discourse analysis and seeks to derive QUD structures automatically. QUD views each sentence as an answer to a question triggered in prior context; thus, we characterize relationships between sentences as free-form questions, in contrast to exhaustive fine-grained taxonomies. We develop the first-of-its-kind QUD parser that derives a dependency structure of questions over full documents, trained using a large, crowdsourced question-answering dataset DCQA (Ko et al., 2022). Human evaluation results show that QUD dependency parsing is possible for language models trained with this crowdsourced, generalizable annotation scheme. We illustrate how our QUD structure is distinct from RST trees, and demonstrate the utility of QUD analysis in the context of document simplification. Our findings show that QUD parsing is an appealing alternative for automatic discourse processing.

AxomiyaBERTa: A Phonologically-aware Transformer Model for Assamese
Abhijnan Nath, Sheikh Abdul Mannan and Nikhil Krishnaswamy   19:00-21:00 (Metropolitan West)
Despite their successes in NLP, Transformer-based language models still require extensive computing resources and suffer in low-resource or low-compute settings. In this paper, we present AxomiyaBERTa, a novel BERT model for Assamese, a morphologically-rich low-resource language (LRL) of Eastern India. AxomiyaBERTa is trained only on the masked language modeling (MLM) task, without the typical addi-

tional next sentence prediction (NSP) objective, and our results show that in resource-scarce settings for very low-resource languages like Assamese, MLM alone can be successfully leveraged for a range of tasks. AxomiyaBERTa achieves SOTA on token-level tasks like Named Entity Recognition and also performs well on "longer-context" tasks like Cloze-style QA and Wiki Title Prediction, with the assistance of a novel embedding disperser and phonological signals respectively. Moreover, we show that AxomiyaBERTa can leverage phonological signals for even more challenging tasks, such as a novel cross-document coreference task on a translated version of the ECB+ corpus, where we present a new SOTA result for an LRL. Our source code and evaluation scripts may be found at https://github.com/csu-signal/axomiyaberta.

SERENGETI: Massively Multilingual Language Models for Africa
Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed and Alcides Alcoba Inciarte   19:00-21:00 (Metropolitan West)
Multilingual pretrained language models (mPLMs) acquire valuable, generalizable linguistic information during pretraining and have advanced the state of the art on task-specific finetuning. To date, only 31 out of 2,000 African languages are covered in existing language models. We ameliorate this limitation by developing SERENGETI, a set of massively multilingual language model that covers 517 African languages and language varieties. We evaluate our novel models on eight natural language understanding tasks across 20 datasets, comparing to 4 mPLMs that cover 4-23 African languages. SERENGETI outperforms other models on 11 datasets across the eights tasks, achieving 82.27 average F_1. We also perform analyses of errors from our models, which allows us to investigate the influence of language genealogy and linguistic similarity when the models are applied under zero-shot settings. We will publicly release our models for research.Anonymous link

Verifying Annotation Agreement without Multiple Experts: A Case Study with Gujarati SNACS
Maitrey Mehta and Vivek Srikumar   19:00-21:00 (Metropolitan West)
Good datasets are a foundation of NLP research, and form the basis for training and evaluating models of language use. While creating datasets, the standard practice is to verify the annotation consistency using a committee of human annotators. This norm assumes that multiple annotators are available, which is not the case for highly specialized tasks or low-resource languages. In this paper, we ask: Can we evaluate the quality of a dataset constructed by a single human annotator? To address this question, we propose four weak verifiers to help estimate dataset quality, and outline when each may be employed. We instantiate these strategies for the task of semantic analysis of adpositions in Gujarati, a low-resource language, and show that our weak verifiers concur with a double-annotation study. As an added contribution, we also release the first dataset with semantic annotations in Gujarati along with several model baselines.

X-RiSAWOZ: High-Quality End-to-End Multilingual Dialogue Datasets and Few-shot Agents
Mehrad Moradshahi, Tianhao Shen, Kalika Bali, Monojit Choudhury, Gael de Chalendar, Anmol Goel, Sungkyun Kim, Prashant Kodali, Ponnurangam Kumaraguru, Nasredine Semmar, Sina Semnani, Jiwon Seo, Vivek Seshadri, Manish Shrivastava, Michael Sun, Aditya Yadavalli, Chaobin You, Deyi Xiong and Monica S Lam   19:00-21:00 (Metropolitan West)
Task-oriented dialogue research has mainly focused on a few popular languages like English and Chinese, due to the high dataset creation cost for a new language. To reduce the cost, we apply manual editing to automatically translated data. We create a new multilingual benchmark, X-RiSAWOZ, by translating the Chinese RiSAWOZ to 4 languages: English, French, Hindi, Korean; and a code-mixed English-Hindi language. X-RiSAWOZ has more than 18,000 human-verified dialogue utterances for each language, and unlike most multilingual prior work, is an end-to-end dataset for building fully-functioning agents.
The many difficulties we encountered in creating X-RiSAWOZ led us to develop a toolset to accelerate the post-editing of a new language dataset after translation. This toolset improves machine translation with a hybrid entity alignment technique that combines neural with dictionary-based methods, along with many automated and semi-automated validation checks.
We establish strong baselines for X-RiSAWOZ by training dialogue agents in the zero- and few-shot settings where limited gold data is available in the target language. Our results suggest that our translation and post-editing methodology and toolset can be used to create new high-quality multilingual dialogue agents cost-effectively. Our dataset, code, and toolkit are released open-source.

Can Cross-Lingual Transferability of Multilingual Transformers Be Activated Without End-Task Data?
Zewen Chi, Heyan Huang and Xian-Ling Mao   19:00-21:00 (Metropolitan West)
Pretrained multilingual Transformers have achieved great success in cross-lingual transfer learning. Current methods typically activate the cross-lingual transferability of multilingual Transformers by fine-tuning them on end-task data. However, the methods cannot perform cross-lingual transfer when end-task data are unavailable. In this work, we explore whether the cross-lingual transferability can be activated without end-task data. We propose a cross-lingual transfer method, named PlugIn-X. PlugIn-X disassembles monolingual and multilingual Transformers into sub-modules, and reassembles them to be the multilingual end-task model. After representation adaptation, PlugIn-X finally performs cross-lingual transfer in a plug-and-play style. Experimental results show that PlugIn-X successfully activates the cross-lingual transferability of multilingual Transformers without accessing end-task data. Moreover, we analyze how the cross-model representation alignment affects the cross-lingual transferability.

Language Agnostic Multilingual Information Retrieval with Contrastive Learning
Xiyang Hu, Xinchi Chen, Peng Qi, Deguang Kong, Kunlun Liu, William Yang Wang and zhiheng huang   19:00-21:00 (Metropolitan West)
Multilingual information retrieval (IR) is challenging since annotated training data is costly to obtain in many languages. We present an effective method to train multilingual IR systems when only English IR training data and some parallel corpora between English and other languages are available. We leverage parallel and non-parallel corpora to improve the pretrained multilingual language models' cross-lingual transfer ability. We design a semantic contrastive loss to align representations of parallel sentences that share the same semantics in different languages, and a new language contrastive loss to leverage parallel sentence pairs to remove language-specific information in sentence representations from non-parallel corpora. When trained on English IR data with these losses and evaluated zero-shot on non-English data, our model demonstrates significant improvement to prior work on retrieval performance, while it requires much less computational effort. We also demonstrate the value of our model for a practical setting when a parallel corpus is only available for a few languages, but a lack of parallel corpora resources persists for many other low-resource languages. Our model can work well even with a small number of parallel sentences, and be used as an add-on module to any backbones and other tasks.

Why Does Zero-Shot Cross-Lingual Generation Fail? An Explanation and a Solution
Tianjian Li and Kenton Murray   19:00-21:00 (Metropolitan West)
Zero-shot cross-lingual transfer is when a multilingual model is trained to perform a task in one language and then is applied to another language. Although the zero-shot cross-lingual transfer approach has achieved success in various classification tasks, its performance on natural language generation tasks falls short in quality and sometimes outputs an incorrect language. In our study, we show that the fine-tuning process learns language invariant representations, which is beneficial for classification tasks but harmful for generation tasks. Motivated by this, we propose a simple method to regularize the model from learning language invariant representations and a method to select model checkpoints without a development set in the target language, both resulting in better generation quality. Experiments on three semantically diverse generation tasks show that our method reduces the accidental translation problem by 68% and improves the ROUGE-L score by 1.5 on average.

Cross-Lingual Retrieval Augmented Prompt for Low-Resource Languages
Ercong Nie, Sheng Liang, Helmut Schmid and Hinrich Schütze 19:00-21:00 (Metropolitan West)
Multilingual Pretrained Language Models (MPLMs) perform strongly in cross-lingual transfer. We propose Prompts Augmented by Retrieval Crosslingually (PARC) to improve zero-shot performance on low-resource languages (LRLs) by augmenting the context with prompts consisting of semantically similar sentences retrieved from a high-resource language (HRL). PARC improves zero-shot performance on three downstream tasks (sentiment classification, topic categorization, natural language inference) with multilingual parallel test sets across 10 LRLs covering 6 language families in unlabeled (+5.1%) and labeled settings (+16.3%). PARC also outperforms finetuning by 3.7%. We find a significant positive correlation between cross-lingual transfer performance on one side, and the similarity between high- and low-resource languages as well as the amount of low-resource pretraining data on the other side. A robustness analysis suggests that PARC has the potential to achieve even stronger performance with more powerful MPLMs.

Predicting Human Translation Difficulty Using Automatic Word Alignment
Zheng Wei Lim, Trevor Cohn, Charles Kemp and Ekaterina Vylomova 19:00-21:00 (Metropolitan West)
Translation difficulty arises when translators are required to resolve translation ambiguity from multiple possible translations. Translation difficulty can be measured by recording the diversity of responses provided by human translators and the time taken to provide these responses, but these behavioral measures are costly and do not scale. In this work, we use word alignments computed over large scale bilingual corpora to develop predictors of lexical translation difficulty. We evaluate our approach using behavioural data from translations provided both in and out of context, and report results that improve on a previous embedding-based approach (Thompson et al., 2020). Our work can therefore contribute to a deeper understanding of cross-lingual differences and of causes of translation difficulty.

Automatic Identification of Code-Switching Functions in Speech Transcripts
Ritu Madhura Belani and Jeffrey Flanigan 19:00-21:00 (Metropolitan West)
Code-switching, or switching between languages, occurs for many reasons and has important linguistic, sociological, and cultural implications. Multilingual speakers code-switch for a variety of communicative functions, such as expressing emotions, borrowing terms, making jokes, introducing a new topic, etc. The function of code-switching may be quite useful for the analysis of linguists, cognitive scientists, speech therapists, and others, but is not readily apparent. To remedy this situation, we annotate and release a new dataset of functions of code-switching in Spanish-English. We build the first system (to our knowledge) to automatically identify a wide range of functions for which speakers code-switch in everyday speech, achieving an accuracy of 75% across all functions.

Tokenization Impacts Multilingual Language Modeling: Assessing Vocabulary Allocation and Overlap Across Languages
Tomasz Limisiewicz, Jiří Balhar and David Mareček 19:00-21:00 (Metropolitan West)
Multilingual language models have recently gained attention as a promising solution for representing multiple languages in a single model. In this paper, we propose new criteria to evaluate the quality of lexical representation and vocabulary overlap observed in sub-word tokenizers. Our findings show that the overlap of vocabulary across languages can be actually detrimental to certain downstream tasks (POS, dependency tree labeling). In contrast, NER and sentence-level tasks (cross-lingual retrieval, NLI) benefit from sharing vocabulary. We also observe that the coverage of the language-specific tokens in the multilingual vocabulary significantly impacts the word-level tasks. Our study offers a deeper understanding of the role of tokenizers in multilingual language models and guidelines for future model developers to choose the most suitable tokenizer for their specific application before undertaking costly model pre-training.

Code-Switched Text Synthesis in Unseen Language Pairs
I-Hung Hsu, Avik Ray, Shubham Garg, Nanyun Peng and Jing Huang 19:00-21:00 (Metropolitan West)
Existing efforts on text synthesis for code-switching mostly require training on code-switched texts in the target language pairs, limiting the deployment of the models to cases lacking code-switched data. In this work, we study the problem of synthesizing code-switched texts for language pairs absent from the training data. We introduce GLOSS, a model built on top of a pre-trained multilingual machine translation model (PMMTM) with an additional code-switching module. This module, either an adapter or extra prefixes, learns code-switching patterns from code-switched data during training, while the primary component of GLOSS, i.e., the PMMTM, is frozen. The design of only adjusting the code-switching module prevents our model from overfitting to the constrained training data for code-switching. Hence, GLOSS exhibits the ability to generalize and synthesize code-switched texts across a broader spectrum of language pairs. Additionally, we develop a self-training algorithm on target language pairs further to enhance the reliability of GLOSS. Automatic evaluations on four language pairs show that GLOSS achieves at least 55% relative BLEU and METEOR scores improvements compared to strong baselines. Human evaluations on two language pairs further validate the success of GLOSS.

Exploring Anisotropy and Outliers in Multilingual Language Models for Cross-Lingual Semantic Sentence Similarity
Katharina Haemmerl, Alina Fastowski, Jindřich Libovický and Alexander Fraser 19:00-21:00 (Metropolitan West)
Previous work has shown that the representations output by contextual language models are more anisotropic than static type embeddings, and typically display outlier dimensions. This seems to be true for both monolingual and multilingual models, although much less work has been done on the multilingual context. Why these outliers occur and how they affect the representations is still an active area of research. We investigate outlier dimensions and their relationship to anisotropy in multiple pre-trained multilingual language models. We focus on cross-lingual semantic similarity tasks, as these are natural tasks for evaluating multilingual representations. Specifically, we examine sentence representations. Sentence transformers which are fine-tuned on parallel resources (that are not always available) perform better on this task, and we show that their representations are more isotropic. However, we aim to improve multilingual representations in general. We investigate how much of the performance difference can be made up by only transforming the embedding space without fine-tuning, and visualise the resulting spaces. We test different operations: Removing individual outlier dimensions, cluster-based isotropy enhancement, and ZCA whitening. We publish our code for reproducibility.

Frustratingly Easy Label Projection for Cross-lingual Transfer
Yang Chen, Chao Jiang, Alan Ritter and Wei Xu 19:00-21:00 (Metropolitan West)
Translating training data into many languages has emerged as a practical solution for improving cross-lingual transfer. For tasks that involve span-level annotations, such as information extraction or question answering, an additional label projection step is required to map annotated spans onto the translated texts. Recently, a few efforts have utilized a simple mark-then-translate method to jointly perform translation and projection by inserting special markers around the labeled spans in the original sentence. However, as far as we are aware, no empirical analysis has been conducted on how this approach compares to traditional annotation projection based on word alignment. In this paper, we present an extensive empirical study across 57 languages and three tasks (QA, NER, and Event Extraction) to evaluate the effectiveness and limitations of both methods, filling an important gap in the literature. Experimental results show that our optimized version of mark-then-translate, which we call EasyProject, is easily applied to many languages and works surprisingly well, outperforming the more complex word alignment-based methods. We analyze several key factors that affect the end-task performance, and show EasyProject works well because it can accurately preserve label span boundaries after translation. We will publicly release all our code and data.

TADA : Task Agnostic Dialect Adapters for English

William Held, Caleb Ziems and Diyi Yang 19:00-21:00 (Metropolitan West)
Large Language Models, the dominant starting point for Natural Language Processing (NLP) applications, fail at a higher rate for speakers of English dialects other than Standard American English (SAE). Prior work addresses this using task specific data or synthetic data augmentation, both of which require intervention for each dialect and task pair. This poses a scalability issue that prevents the broad adoption of robust dialectal English NLP. We introduce a simple yet effective method for task-agnostic dialect adaptation by aligning non-SAE dialects using adapters and composing them with task-specific adapters from SAE. Task-Agnostic Dialect Adapters (TADA) improve dialectal robustness on 4 dialectal variants of the GLUE benchmark without task-specific supervision.

Mini-Model Adaptation: Efficiently Extending Pretrained Models to New Languages via Aligned Shallow Training
Kelly Marchisio, Patrick Lewis, Yihong Chen and Mikel Artetxe 19:00-21:00 (Metropolitan West)
Prior work shows that it is possible to expand pretrained Masked Language Models (MLMs) to new languages by learning a new set of embeddings, while keeping the transformer body frozen. Despite learning a small subset of parameters, this approach is not compute-efficient, as training the new embeddings requires a full forward and backward pass over the entire model. We propose mini-model adaptation, a compute-efficient alternative that builds a shallow mini-model from a fraction of a large model's parameters. New language-specific embeddings can then be efficiently trained over the mini-model and plugged into the aligned large model for rapid cross-lingual transfer. We explore two approaches to learn mini-models: MINIJOINT, which jointly pretrains the primary model and the mini-model using a single transformer with a secondary MLM head at a middle layer; and MINIPOST, where we start from a regular pretrained model, build a mini-model by extracting and freezing a few layers, and learn a small number of parameters on top. Experiments on XNLI, MLQA and PAWS-X show that mini-model adaptation matches the performance of the standard approach using up to 2.3x less compute on average.

Explanation Regeneration via Information Bottleneck
Qintong Li, Zhiyong Wu, Lingpeng Kong and Wei Bi 19:00-21:00 (Metropolitan West)
Explaining the black-box predictions of NLP models naturally and accurately is an important open problem in natural language generation. These free-text explanations are expected to contain sufficient and carefully-selected evidence to form supportive arguments for predictions. Thanks to the superior generative capacity of large pretrained language models (PLM), recent work built on prompt engineering enables explanations generated without specific training. However, explanations generated through single-pass prompting often lack sufficiency and conciseness, due to the prompt complexity and hallucination issues. To discard the dross and take the essence of current PLM's results, we propose to produce sufficient and concise explanations via the information bottleneck (EIB) theory. EIB regenerates explanations by polishing the single-pass output of PLM but retaining the information that supports the contents being explained by balancing two information bottleneck objectives. Experiments on two different tasks verify the effectiveness of EIB through automatic evaluation and thoroughly-conducted human evaluation.

Characterizing the Impacts of Instances on Robustness
Rui Zheng, Zhiheng Xi, Qin Liu, Wenbin Lai, Tao Gui, Qi Zhang, Xuanjing Huang, Jin Ma, Ying Shan and Weifeng Ge 19:00-21:00 (Metropolitan West)
Building robust deep neural networks (DNNs) against adversarial attacks is an important but challenging task. Previous defense approaches mainly focus on developing new model structures or training algorithms, but they do little to tap the potential of training instances, especially instances with robust patterns carring innate robustness. In this paper, we show that robust and non-robust instances in the training dataset, though are both important for test performance, have contrary impacts on robustness, which makes it possible to build a highly robust model by leveraging the training dataset in a more effective way. We propose a new method that can distinguish between robust instances from non-robust ones according to the model's sensitivity to perturbations on individual instances during training. Surprisingly, we find that the model under standard training easily overfits the robust instances by relying on their simple patterns before the model completely learns their robust features. Finally, we propose a new mitigation algorithm to further release the potential of robust instances. Experimental results show that proper use of robust instances in the original dataset is a new line to achieve highly robust models.

Counterfactuals of Counterfactuals: a back-translation-inspired approach to analyse counterfactual editors
George Filandrianos, Edmund G Dervakos, Orfeas Menis Mastromichalakis, Chrysoula Zerva and Giorgos Stamou 19:00-21:00 (Metropolitan West)
In the wake of responsible AI, interpretability methods, which attempt to provide an explanation for the predictions of neural models have seen rapid progress. In this work, we are concerned with explanations that are applicable to natural language processing (NLP) models and tasks, and we focus specifically on the analysis of counterfactual, contrastive explanations. We note that while there have been several explainers proposed to produce counterfactual explanations, their behaviour can vary significantly and the lack of a universal ground truth for the counterfactual edits imposes an insuperable barrier on their evaluation. We propose a new back translation-inspired evaluation methodology that utilises earlier outputs of the explainer as ground truth proxies to investigate the consistency of explainers. We show that by iteratively feeding the counterfactual to the explainer we can obtain valuable insights into the behaviour of both the predictor and the explainer models, and infer patterns that would be otherwise obscured. Using this methodology, we conduct a thorough analysis and propose a novel metric to evaluate the consistency of counterfactual generation approaches with different characteristics across available performance indicators.

Conformal Nucleus Sampling
Shauli Rayfogel, Yoav Goldberg and Jacob Goldberger 19:00-21:00 (Metropolitan West)
Language models generate text based on successively sampling the next word. A decoding procedure based on nucleus (top-p) sampling chooses from the smallest possible set of words whose cumulative probability exceeds the probability p. In this work, we assess whether a top-p set is indeed aligned with its probabilistic meaning in various linguistic contexts. We employ conformal prediction, a calibration procedure that focuses on the construction of minimal prediction sets according to a desired confidence level, to calibrate the parameter p as a function of the entropy of the next word distribution. We find that OPT models are overconfident, and that calibration shows a moderate inverse scaling with model size.

Fighting Bias With Bias: Promoting Model Robustness by Amplifying Dataset Biases
Yuval Reif and Roy Schwartz 19:00-21:00 (Metropolitan West)
NLP models often rely on superficial cues known as dataset biases to achieve impressive performance, and can fail on examples where these biases do not hold. Recent work sought to develop robust, unbiased models by filtering biased examples from training sets. In this work, we argue that such filtering can obscure the true capabilities of models to overcome biases, which might never be removed in full from the dataset. We suggest that in order to drive the development of models robust to subtle biases, dataset biases should be amplified in the training set. We introduce an evaluation framework defined by a bias-amplified training set and an anti-biased test set, both automatically extracted from existing datasets. Experiments across three notions of bias, four datasets and two models show that our framework is substantially more challenging for models than the original data splits, and even more challenging than hand-crafted challenge sets. Our evaluation framework can use any existing dataset, even those considered obsolete, to test model robustness. We hope our work will guide the development of robust models that do not rely on superficial biases and correlations. To this end, we publicly release our code and data.

Robustness of Learning from Task Instructions

Jiasheng Gu, Hongyu Zhao, Hanzi Xu, Liangyu Nie, Hongyuan Mei and Wenpeng Yin 19:00-21:00 (Metropolitan West)

Traditional supervised learning mostly works on individual tasks and requires training on a large set of task-specific examples. This paradigm seriously hinders the development of task generalization since preparing a task-specific example set is costly. To build a system that can quickly and easily generalize to new tasks, task instructions have been adopted as an emerging trend of supervision recently. These instructions give the model the definition of the task and allow the model to output the appropriate answer based on the instructions and inputs. However, task instructions are often expressed in different forms, which can be interpreted from two threads: first, some instructions are short sentences and are pretrained language model (PLM) oriented, such as prompts, while other instructions are paragraphs and are human-oriented, such as those in Amazon MTurk; second, different end-users very likely explain the same task with instructions of different textual expressions. A robust system for task generalization should be able to handle any new tasks regardless of the variability of instructions.

However, the system robustness in dealing with instruction-driven task generalization is still unexplored. This work investigates the system robustness when the instructions of new tasks are (i) manipulated, (ii) paraphrased, or (iii) from different levels of conciseness. To our knowledge, this is the first work that systematically studies how robust a PLM is when it is supervised by instructions with different factors of variability.

COCKATIEL: COntinuous Concept ranKed ATtribution with Interpretable ELements for explaining neural net classifiers on NLP

Fanny Jourdan, Agustin Martin Picard, Thomas Fel, Laurent Risser, Jean-Michel Loubes and Nicholas Asher 19:00-21:00 (Metropolitan West)

Transformer architectures are complex and their use in NLP, while it has engendered many successes, makes their interpretability or explainability challenging. Recent debates have shown that attention maps and attribution methods are unreliable (Pruthi et al., 2019; Brunner et al., 2019). In this paper, we present some of their limitations and introduce COCKATIEL, which successfully addresses some of them. COCKATIEL is a novel, post-hoc, concept-based, model-agnostic XAI technique that generates meaningful explanations from the last layer of a neural net model trained on an NLP classification task by using Non-Negative Matrix Factorization (NMF) to discover the concepts the model leverages to make predictions and by exploiting a Sensitivity Analysis to estimate accurately the importance of each of these concepts for the model. It does so without compromising the accuracy of the underlying model or requiring a new one to be trained.

We conduct experiments in single and multi-aspect sentiment analysis tasks and we show COCKATIEL's superior ability to discover concepts that align with humans' on Transformer models without any supervision, we objectively verify the faithfulness of its explanations through fidelity metrics, and we showcase its ability to provide meaningful explanations in two different datasets.

Our code is freely available: https://github.com/fanny-jourdan/cockatiel

Figurative Language Processing: A Linguistically Informed Feature Analysis of the Behavior of Language Models and Humans

Hyewon Jang, Qi Yu and Diego Frassinelli 19:00-21:00 (Metropolitan West)

Recent years have witnessed a growing interest in investigating what Transformer-based language models (TLMs) actually learn from the training data. This is especially relevant for complex tasks such as the understanding of non-literal meaning. In this work, we probe the performance of three black-box TLMs and two intrinsically transparent white-box models on figurative language classification of sarcasm, similes, idioms, and metaphors. We conduct two studies on the classification results to provide insights into the inner workings of such models. With our first analysis on feature importance, we identify crucial differences in model behavior. With our second analysis using an online experiment with human annotators, we inspect different linguistic characteristics of the four figurative language types.

Model Interpretability and Rationale Extraction by Input Mask Optimization

Marc Felix Brinner and Sina Zarrieß 19:00-21:00 (Metropolitan West)

Concurrent with the rapid progress in neural network-based models in NLP, the need for creating explanations for the predictions of these black-box models has risen steadily. Yet, especially for complex inputs like texts or images, existing interpretability methods still struggle with deriving easily interpretable explanations that also accurately represent the basis for the model's decision. To this end, we propose a new, model-agnostic method to generate extractive explanations for predictions made by neural networks, that is based on masking parts of the input which the model does not consider to be indicative of the respective class. The masking is done using gradient-based optimization combined with a new regularization scheme that enforces sufficiency, comprehensiveness, and compactness of the generated explanation. Our method achieves state-of-the-art results in a challenging paragraph-level rationale extraction task, showing that this task can be performed without training a specialized model. We further apply our method to image inputs and obtain high-quality explanations for image classifications, which indicates that the objectives for optimizing explanation masks in text generalize to inputs of other modalities.

Layerwise universal adversarial attack on NLP models

Olga Tsymboi, Danil Malaev, Andrei Petrovskii and Ivan Oseledets 19:00-21:00 (Metropolitan West)

In this work, we examine the vulnerability of language models to universal adversarial triggers (UATs). We propose a new white-box approach to the construction of layerwise UATs (LUATs), which searches the triggers by perturbing hidden layers of a network. On the example of three transformer models and three datasets from the GLUE benchmark, we demonstrate that our method provides better transferability in a model-to-model setting with an average gain of 9.3% in the fooling rate over the baseline. Moreover, we investigate triggers transferability in the task-to-task setting. Using small subsets from the datasets similar to the target tasks for choosing a perturbed layer, we show that LUATs are more efficient than vanilla UATs by 7.1% in the fooling rate.

Inducing Character-level Structure in Subword-based Language Models with Type-level Interchange Intervention Training

Jing Huang, Zhengxuan Wu, Kyle Mahowald and Christopher Potts 19:00-21:00 (Metropolitan West)

Language tasks involving character-level manipulations (e.g., spelling corrections, arithmetic operations, word games) are challenging for models operating on subword units. To address this, we develop a causal intervention framework to learn robust and interpretable character representations inside subword-based language models. Our method treats each character as a typed variable in a causal model and learns such causal structures by adapting the interchange intervention training method of Geiger et al. (2021). We additionally introduce a suite of character-level tasks that systematically vary in their dependence on meaning and sequence-level context. While character-level models still perform best on purely form-based tasks like string reversal, our method outperforms character-level models on more complex tasks that blend form, meaning, and context, such as spelling correction in context and word search games. Compared with standard subword-based models, our approach also significantly improves robustness on unseen token sequences and leads to human-interpretable internal representations of characters.

SenteCon: Leveraging Lexicons to Learn Human-Interpretable Language Representations

Victoria Lin and Louis-Philippe Morency 19:00-21:00 (Metropolitan West)

Although deep language representations have become the dominant form of language featurization in recent years, in many settings it is important to understand a model's decision-making process. This necessitates not only an interpretable model but also interpretable features. In particular, language must be featurized in a way that is interpretable while still characterizing the original text well. We present SenteCon, a method for introducing human interpretability in deep language representations. Given a passage of text, SenteCon encodes the text as a layer of interpretable categories in which each dimension corresponds to the relevance of a specific category. Our empirical evaluations indicate

that encoding language with SenteCon provides high-level interpretability at little to no cost to predictive performance on downstream tasks. Moreover, we find that SenteCon outperforms existing interpretable language representations with respect to both its downstream performance and its agreement with human characterizations of the text.

HyHTM: Hyperbolic Geometry-based Hierarchical Topic Model
Simra Shahid, Tanay Anand, Nikitha Srikanth, Sumit Bhatia, Balaji Krishnamurthy and nikaash puri   19:00-21:00 (Metropolitan West)
Hierarchical Topic Models (HTMs) are useful for discovering topic hierarchies in a collection of documents. However, traditional HTMs often produce hierarchies where lower-level topics are unrelated and not specific enough to their higher-level topics. Additionally, these methods can be computationally expensive. We present HyHTM - a Hyperbolic geometry-based Hierarchical Topic Model - that addresses these limitations by incorporating hierarchical information from hyperbolic geometry to explicitly model hierarchies in topic models. Experimental results with four baselines show that HyHTM can better attend to parent-child relationships among topics. HyHTM produces coherent topic hierarchies that specialize in granularity from generic higher-level topics to specific lower-level topics. Further, our model is significantly faster and leaves a much smaller memory footprint than our best-performing baseline. We have made the source code for our algorithm publicly accessible.

Adversarial Robustness of Prompt-based Few-Shot Learning for Natural Language Understanding
Venkata Prabhakara Sarath Nookala, Gaurav Verma, Subhabrata Mukherjee and Srijan Kumar   19:00-21:00 (Metropolitan West)
State-of-the-art few-shot learning (FSL) methods leverage prompt-based fine-tuning to obtain remarkable results for natural language understanding (NLU) tasks. While much of the prior FSL methods focus on improving downstream task performance, there is a limited understanding of the adversarial robustness of such methods. In this work, we conduct an extensive study of several state-of-the-art FSL methods to assess their robustness to adversarial perturbations. To better understand the impact of various factors towards robustness (or the lack of it), we evaluate prompt-based FSL methods against fully fine-tuned models for aspects such as the use of unlabeled data, multiple prompts, number of few-shot examples, model size and type. Our results on six GLUE tasks indicate that compared to fully fine-tuned models, vanilla FSL methods lead to a notable relative drop in task performance (i.e., are less robust) in the face of adversarial perturbations. However, using (i) unlabeled data for prompt-based FSL and (ii) multiple prompts flip the trend – the few-shot learning approaches demonstrate a lesser drop in task performance than fully fine-tuned models. We further demonstrate that increasing the number of few-shot examples and model size lead to increased adversarial robustness of vanilla FSL methods. Broadly, our work sheds light on the adversarial robustness evaluation of prompt-based FSL methods for NLU tasks.

Robust Natural Language Understanding with Residual Attention Debiasing
Fei Wang, James Y. Huang, Tianyi Yan, Wenxuan Zhou and Muhao Chen   19:00-21:00 (Metropolitan West)
Natural language understanding (NLU) models often suffer from unintended dataset biases. Among bias mitigation methods, ensemble-based debiasing methods, especially product-of-experts (PoE), have stood out for their impressive empirical success. However, previous ensemble-based debiasing methods typically apply debiasing on top-level logits without directly addressing biased attention patterns. Attention serves as the main media of feature interaction and aggregation in PLMs and plays a crucial role in providing robust prediction. In this paper, we propose REsidual Attention Debiasing (READ), an end-to-end debiasing method that mitigates unintended biases from attention. Experiments on three NLU benchmarks show that READ significantly improves the OOD performance of BERT-based models, including +12.9% accuracy on HANS, +11.0% accuracy on FEVER-Symmetric, and +2.7% F1 on PAWS. Detailed analyses demonstrate the crucial role of unbiased attention in robust NLU models and that READ effectively mitigates biases in attention.

Transformer Language Models Handle Word Frequency in Prediction Head
Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi and Kentaro Inui   19:00-21:00 (Metropolitan West)
Prediction head is a crucial component of Transformer language models. Despite its direct impact on prediction, this component has often been overlooked in analyzing Transformers. In this study, we investigate the inner workings of the prediction head, specifically focusing on bias parameters. Our experiments with BERT and GPT-2 models reveal that the biases in their word prediction heads play a significant role in the models' ability to reflect word frequency in a corpus, aligning with the logit adjustment method commonly used in long-tailed learning. We also quantify the effect of controlling the biases in practical auto-regressive text generation scenarios; under a particular setting, more diverse text can be generated without compromising text quality.

Is Continuous Prompt a Combination of Discrete Prompts? Towards a Novel View for Interpreting Continuous Prompts
Tianjie Ju, Yubin Zheng, Hanyi Wang, Haodong Zhao and Gongshen Liu   19:00-21:00 (Metropolitan West)
The broad adoption of continuous prompts has brought state-of-the-art results on a diverse array of downstream natural language processing (NLP) tasks. Nonetheless, little attention has been paid to the interpretability and transferability of continuous prompts. Faced with the challenges, we investigate the feasibility of interpreting continuous prompts as the weighting of discrete prompts by jointly optimizing prompt fidelity and downstream fidelity. Our experiments show that: (1) one can always find a combination of discrete prompts as the replacement of continuous prompts that performs well on downstream tasks; (2) our interpretable framework faithfully reflects the reasoning process of source prompts; (3) our interpretations provide effective readability and plausibility, which is helpful to understand the decision-making of continuous prompts and discover potential shortcuts. Moreover, through the bridge constructed between continuous prompts and discrete prompts using our interpretations, it is promising to implement the cross-model transfer of continuous prompts without extra training signals. We hope this work will lead to a novel perspective on the interpretations of continuous prompts.

Claim-Dissector: An Interpretable Fact-Checking System with Joint Re-ranking and Veracity Prediction
Martin Fajcik, Petr Motlicek and Pavel Smrz   19:00-21:00 (Metropolitan West)
We present Claim-Dissector: a novel latent variable model for fact-checking and analysis, which given a claim and a set of retrieved evidence jointly learns to identify: (i) the relevant evidences to the given claim (ii) the veracity of the claim. We propose to disentangle the per-evidence relevance probability and its contribution to the final veracity probability in an interpretable way — the final veracity probability is proportional to a linear ensemble of per-evidence relevance probabilities. In this way, the individual contributions of evidences towards the final predicted probability can be identified. In per-evidence relevance probability, our model can further distinguish whether each relevant evidence is supporting (S) or refuting (R) the claim. This allows to quantify how much the S/R probability contributes to final verdict or to detect disagreeing evidence. Despite its interpretable nature, our system achieves results competetive with state-of-the-art on the FEVER dataset, as compared to typical two-stage system pipelines, while using significantly fewer parameters. Furthermore, our analysis shows that our model can learn fine-grained relevance cues while using coarse-grained supervision and we demonstrate it in 2 ways. (i) We show that our model can achieve competitive sentence recall while using only paragraph-level relevance supervision. (ii) Traversing towards the finest granularity of relevance, we show that our model is capable of identifying relevance at the token level. To do this, we present a new benchmark TLR-FEVER focusing on token-level interpretability — humans annotate tokens in relevant evidences they considered essential when making their judgment. Then we measure how similar are these annotations to the tokens our model is focusing on.

PromptAttack: Probing Dialogue State Trackers with Adversarial Prompts
Xiangjue Dong, Yun He, Ziwei Zhu and James Caverlee   19:00-21:00 (Metropolitan West)

A key component of modern conversational systems is the Dialogue State Tracker (or DST), which models a user's goals and needs. Toward building more robust and reliable DSTs, we introduce a prompt-based learning approach to automatically generate effective adversarial examples to probe DST models. Two key characteristics of this approach are: (i) it only needs the output of the DST with no need for model parameters, and (ii) it can learn to generate natural language utterances that can target any DST. Through experiments over state-of-the-art DSTs, the proposed framework leads to the greatest reduction in accuracy and the best attack success rate while maintaining good fluency and a low perturbation ratio. We also show how much the generated adversarial examples can bolster a DST through adversarial training. These results indicate the strength of prompt-based attacks on DSTs and leave open avenues for continued refinement.

Disagreement Matters: Preserving Label Diversity by Jointly Modeling Item and Annotator Label Distributions with DisCo
Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj B Bhensadadia, Ashiqur KhudaBukhsh and Christopher Homan   19:00-21:00
(Metropolitan West)
Annotator disagreement is common whenever human judgment is needed for supervised learning. It is conventional to assume that one label per item represents ground truth. However, this obscures minority opinions, if present. We regard "ground truth" as the distribution of all labels that a population of annotators could produce, if asked (and of which we only have a small sample). We next introduce DisCo (Distribution from Context), a simple neural model that learns to predict this distribution. The model takes annotator-item pairs, rather than items alone, as input, and performs inference by aggregating over all annotators. Despite its simplicity, our experiments show that, on six benchmark datasets, our model is competitive with, and frequently outperforms, other, more complex models that either do not model specific annotators or were not designed for label distribution learning.

Bias Beyond English: Counterfactual Tests for Bias in Sentiment Analysis in Four Languages
Seraphina Goldfarb-Tarrant, Adam Lopez, Roi Blanco and Diego Marcheggiani   19:00-21:00 (Metropolitan West)
Sentiment analysis (SA) systems are used in many products and hundreds of languages. Gender and racial biases are well-studied in English SA systems, but understudied in other languages, with few resources for such studies. To remedy this, we build a counterfactual evaluation corpus for gender and racial/migrant bias in four languages. We demonstrate its usefulness by answering a simple but important question that an engineer might need to answer when deploying a system: What biases do systems import from pre-trained models when compared to a baseline with no pre-training? Our evaluation corpus, by virtue of being counterfactual, not only reveals which models have less bias, but also pinpoints changes in model bias behaviour, which enables more targeted mitigation strategies. We release our code and evaluation corpora to facilitate future research.

Debiasing should be Good and Bad: Measuring the Consistency of Debiasing Techniques in Language Models
Robert A. Morabito, Jad Kabbara and Ali Emami   19:00-21:00 (Metropolitan West)
Debiasing methods that seek to mitigate the tendency of Language Models (LMs) to occasionally output toxic or inappropriate text have recently gained traction. In this paper, we propose a standardized protocol which distinguishes methods that yield not only desirable results, but are also consistent with their mechanisms and specifications. For example, we ask, given a debiasing method that is developed to reduce toxicity in LMs, if the definition of toxicity used by the debiasing method is reversed, would the debiasing results also be reversed? We used such considerations to devise three criteria for our new protocol: Specification Polarity, Specification Importance, and Domain Transferability. As a case study, we apply our protocol to a popular debiasing method, Self-Debiasing, and compare it to one we propose, called Instructive Debiasing, and demonstrate that consistency is as important an aspect to debiasing viability as is simply a desirable result. We show that our protocol provides essential insights into the generalizability and interpretability of debiasing methods that may otherwise go overlooked.

FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models
Shramay Palta and Rachel Rudinger   19:00-21:00 (Metropolitan West)
It is common sense that one should prefer to eat a salad with a fork rather than with a chainsaw. However, for eating a bowl of rice, the choice between a fork and a pair of chopsticks is culturally relative. We introduce FORK, a small, manually-curated set of CommonsenseQA-style questions for probing cultural biases and assumptions present in commonsense reasoning systems, with a specific focus on food-related customs. We test several CommonsenseQA systems on FORK, and while we see high performance on questions about the US culture, the poor performance of these systems on questions about non-US cultures highlights systematic cultural assumptions aligned with US over non-US cultures.

Foveate, Attribute, and Rationalize: Towards Physically Safe and Trustworthy AI
Alex Mei, Sharon Levy and William Yang Wang   19:00-21:00 (Metropolitan West)
Users' physical safety is an increasing concern as the market for intelligent systems continues to grow, where unconstrained systems may recommend users dangerous actions that can lead to serious injury. Covertly unsafe text is an area of particular interest, as such text may arise from everyday scenarios and are challenging to detect as harmful. We propose FARM, a novel framework leveraging external knowledge for trustworthy rationale generation in the context of safety. In particular, FARM foveates on missing knowledge to qualify the information required to reason in specific scenarios and retrieves this information with attribution to trustworthy sources. This knowledge is used to both classify the safety of the original text and generate human-interpretable rationales, shedding light on the risk of systems to specific user groups and helping both stakeholders manage the risks of their systems and policymakers to provide concrete safeguards for consumer safety. Our experiments show that FARM obtains state-of-the-art results on the SafeText dataset, showing absolute improvement in safety classification accuracy by 5.9%.

Shielded Representations: Protecting Sensitive Attributes Through Iterative Gradient-Based Projection
Shadi Iskander, Kira Radinsky and Yonatan Belinkov   19:00-21:00 (Metropolitan West)
Natural language processing models tend to learn and encode social biases present in the data. One popular approach for addressing such biases is to eliminate encoded information from the model's representations. However, current methods are restricted to removing only linearly encoded information. In this work, we propose Iterative Gradient-Based Projection (IGBP), a novel method for removing non-linear encoded concepts from neural representations. Our method consists of iteratively training neural classifiers to predict a particular attribute we seek to eliminate, followed by a projection of the representation on a hypersurface, such that the classifiers become oblivious to the target attribute. We evaluate the effectiveness of our method on the task of removing gender and race information as sensitive attributes. Our results demonstrate that IGBP is effective in mitigating bias through intrinsic and extrinsic evaluations, with minimal impact on downstream task accuracy.

T2IAT: Measuring Valence and Stereotypical Biases in Text-to-Image Generation
Jialu Wang, Xinyue Gabby Liu, Zonglin Di, Yang Liu and Xin Eric Wang   19:00-21:00 (Metropolitan West)
Warning: This paper contains several contents that may be toxic, harmful, or offensive.

In the last few years, text-to-image generative models have gained remarkable success in generating images with unprecedented quality accompanied by a breakthrough of inference speed. Despite their rapid progress, human biases that manifest in the training examples, particularly with regard to common stereotypical biases, like gender and skin tone, still have been found in these generative models. In this work, we seek to measure more complex human biases exist in the task of text-to-image generations. Inspired by the well-known Implicit

Association Test (IAT) from social psychology, we propose a novel Text-to-Image Association Test (T2IAT) framework that quantifies the implicit stereotypes between concepts and valence, and those in the images. We replicate the previously documented bias tests on generative models, including morally neutral tests on flowers and insects as well as demographic stereotypical tests on diverse social attributes. The results of these experiments demonstrate the presence of complex stereotypical behaviors in image generations.

An Exploratory Study on Model Compression for Text-to-SQL

Shuo Sun, Yuze Gao, Yuchen Zhang, Jian Su, Bin Chen, Yingzhan Lin and Shuqi Sun   19:00-21:00 (Metropolitan West)
Text-to-SQL translates user queries into SQL statements that can retrieve relevant answers from relational databases. Recent approaches to Text-to-SQL rely on pre-trained language models that are computationally expensive and technically challenging to deploy in real-world applications that require real-time or on-device processing capabilities. In this paper, we perform a focused study on the feasibility of applying recent model compression techniques to sketch-based and sequence-to-sequence Text-to-SQL models. Our results reveal that sketch-based Text-to-SQL models generally have higher inference efficiency and respond better to model compression than sequence-to-sequence models, making them ideal for real-world deployments, especially in use cases with simple SQL statements.

It Takes Two to Tango: Navigating Conceptualizations of NLP Tasks and Measurements of Performance

Arjun Subramonian, Xingdi Yuan, Hal Daumé III and Su Lin Blodgett   19:00-21:00 (Metropolitan West)
Progress in NLP is increasingly measured through benchmarks; hence, contextualizing progress requires understanding when and why practitioners may disagree about the validity of benchmarks. We develop a taxonomy of disagreement, drawing on tools from measurement modeling, and distinguish between two types of disagreement: 1) how tasks are conceptualized and 2) how measurements of model performance are operationalized. To provide evidence for our taxonomy, we conduct a meta-analysis of relevant literature to understand how NLP tasks are conceptualized, as well as a survey of practitioners about their impressions of different factors that affect benchmark validity. Our meta-analysis and survey across eight tasks, ranging from coreference resolution to question answering, uncover that tasks are generally not clearly and consistently conceptualized and benchmarks suffer from operationalization disagreements. These findings support our proposed taxonomy of disagreement. Finally, based on our taxonomy, we present a framework for constructing benchmarks and documenting their limitations.

Reimagining Retrieval Augmented Language Models for Answering Queries

Wang-Chiew Tan, Yuliang Li, Pedro Rodriguez, Richard James, Xi Victoria Lin, Alon Halevy and Wen-tau Yih   19:00-21:00 (Metropolitan West)
We present a reality check on large language models and inspect the promise of retrieval-augmented language models in comparison. Such language models are semi-parametric, where models integrate model parameters and knowledge from external data sources to make their predictions, as opposed to the parametric nature of vanilla large language models. We give initial experimental findings that semi-parametric architectures can be enhanced with views, a query analyzer/planner, and provenance to make a significantly more powerful system for question answering in terms of accuracy and efficiency, and potentially for other NLP tasks.

Follow the leader(board) with confidence: Estimating p-values from a single test set with item and response variance

Shira Wein, Christopher Homan, Lora Aroyo and Chris Welty   19:00-21:00 (Metropolitan West)
Among the problems with leaderboard culture in NLP has been the widespread lack of confidence estimation in reported results. In this work, we present a framework and simulator for estimating p-values for comparisons between the results of two systems, in order to understand the confidence that one is actually better (i.e. ranked higher) than the other. What has made this difficult in the past is that each system must itself be evaluated by comparison to a gold standard. We define a null hypothesis that each system's metric scores are drawn from the same distribution, using variance found naturally (though rarely reported) in test set items and individual labels on an item (responses) to produce the metric distributions. We create a test set that evenly mixes the responses of the two systems under the assumption the null hypothesis is true. Exploring how to best estimate the true p-value from a single test set under different metrics, tests, and sampling methods, we find that the presence of response variance (from multiple raters or multiple model versions) has a profound impact on p-value estimates for model comparison, and that choice of metric and sampling method is critical to providing statistical guarantees on model comparisons.

Can Language Models Be Specific? How?

Jie Huang, Kevin Chen-Chuan Chang, Jinjun Xiong and Wen-mei Hwu   19:00-21:00 (Metropolitan West)
"He is a person", "Paris is located on the earth". Both statements are correct but meaningless - due to lack of specificity. In this paper, we propose to measure how specific the language of pre-trained language models (PLMs) is. To achieve this, we introduce a novel approach to build a benchmark for specificity testing by forming masked token prediction tasks with prompts. For instance, given "Toronto is located in [MASK].", we want to test whether a more specific answer will be better filled in by PLMs, e.g., Ontario instead of Canada. From our evaluations, we show that existing PLMs have only a slight preference for more specific answers. We identify underlying factors affecting the specificity and design two prompt-based methods to improve the specificity. Results show that the specificity of the models can be improved by the proposed methods without additional training. We hope this work can bring to awareness the notion of specificity of language models and encourage the research community to further explore this important but understudied problem.

Are Layout-Infused Language Models Robust to Layout Distribution Shifts? A Case Study with Scientific Documents

Catherine Chen, Zejiang Shen, Dan Klein, Gabriel Stanovsky, Doug Downey and Kyle Lo   19:00-21:00 (Metropolitan West)
Recent work has shown that infusing layout features into language models (LMs) improves processing of visually-rich documents such as scientific papers. Layout-infused LMs are often evaluated on documents with familiar layout features (e.g., papers from the same publisher), but in practice models encounter documents with unfamiliar distributions of layout features, such as new combinations of text sizes and styles, or new spatial configurations of textual elements. In this work we test whether layout-infused LMs are robust to layout distribution shifts. As a case study we use the task of scientific document structure recovery, segmenting a scientific paper into its structural categories (e.g., "title", "caption", "reference"). To emulate distribution shifts that occur in practice we re-partition the GROTOAP2 dataset. We find that under layout distribution shifts model performance degrades by up to 20 F1. Simple training strategies, such as increasing training diversity, can reduce this degradation by over 35% relative F1; however, models fail to reach in-distribution performance in any tested out-of-distribution conditions. This work highlights the need to consider layout distribution shifts during model evaluation, and presents a methodology for conducting such evaluations.

A Comparative Analysis of the Effectiveness of Rare Tokens on Creative Expression using ramBERT

Youbin Lee, Deokgi Kim, Byung-Won On and Ingyu Lee   19:00-21:00 (Metropolitan West)
Until now, few studies have been explored on Automated Creative Essay Scoring (ACES), in which a pre-trained model automatically labels an essay as a creative or a non-creative. Since the creativity evaluation of essays is very subjective, each evaluator often has his or her own criteria for creativity. For this reason, quantifying creativity in essays is very challenging. In this work, as one of preliminary studies in developing a novel model for ACES, we deeply investigate the correlation between creative essays and expressiveness. Specifically, we explore how rare tokens affect the evaluation of creativity for essays. For such a journey, we present five distinct methods to extract rare tokens, and conduct a comparative study on the correlation between rare tokens and creative essay evaluation results using BERT. Our experimental

results showed clear correlation between rare tokens and creative essays. In all test sets, accuracies of our rare token masking-based BERT (ramBERT) model were improved over the existing BERT model up to 14%.

The Devil is in the Details: On the Pitfalls of Event Extraction Evaluation

Hao Peng, Xiaozhi Wang, Feng Yao, Kaisheng Zeng, Lei Hou, Juanzi Li, Zhiyuan Liu and Weixing Shen   19:00-21:00 (Metropolitan West)
Event extraction (EE) is a crucial task aiming at extracting events from texts, which includes two subtasks: event detection (ED) and event argument extraction (EAE). In this paper, we check the reliability of EE evaluations and identify three major pitfalls: (1) The data preprocessing discrepancy makes the evaluation results on the same dataset not directly comparable, but the data preprocessing details are not widely noted and specified in papers. (2) The output space discrepancy of different model paradigms makes different-paradigm EE models lack grounds for comparison and also leads to unclear mapping issues between predictions and annotations. (3) The absence of pipeline evaluation of many EAE-only works makes them hard to be directly compared with EE works and may not well reflect the model performance in real-world pipeline scenarios. We demonstrate the significant influence of these pitfalls through comprehensive meta-analyses of recent papers and empirical experiments. To avoid these pitfalls, we suggest a series of remedies, including specifying data preprocessing, standardizing outputs, and providing pipeline evaluation results. To help implement these remedies, we develop a consistent evaluation framework OmniEvent, which can be obtained from https://github.com/THU-KEG/OmniEvent.

A Call for Standardization and Validation of Text Style Transfer Evaluation

Phil Sidney Ostheimer, Mayank Kumar Nagda, Marius Kloft and Sophie Fellenz   19:00-21:00 (Metropolitan West)
Text Style Transfer (TST) evaluation is, in practice, inconsistent. Therefore, we conduct a meta-analysis on human and automated TST evaluation and experimentation that thoroughly examines existing literature in the field. The meta-analysis reveals a substantial standardization gap in human and automated evaluation. In addition, we also find a validation gap: only few automated metrics have been validated using human experiments. To this end, we thoroughly scrutinize both the standardization and validation gap and reveal the resulting pitfalls. This work also paves the way to close the standardization and validation gap in TST evaluation by calling out requirements to be met by future research.

This prompt is measuring <mask>: evaluating bias evaluation in language models

Seraphina Goldfarb-Tarrant, Eddie L. Ungless, Esma Balkir and Su Lin Blodgett   19:00-21:00 (Metropolitan West)
Bias research in NLP seeks to analyse models for social biases, thus helping NLP practitioners uncover, measure, and mitigate social harms. We analyse the body of work that uses prompts and templates to assess bias in language models. We draw on a measurement modelling framework to create a taxonomy of attributes that capture what a bias test aims to measure and how that measurement is carried out. By applying this taxonomy to 90 bias tests, we illustrate qualitatively and quantitatively that core aspects of bias test conceptualisations and operationalisations are frequently unstated or ambiguous, carry implicit assumptions, or be mismatched. Our analysis illuminates the scope of possible bias types the field is able to measure, and reveals types that are as yet under-researched. We offer guidance to enable the community to explore a wider section of the possible bias space, and to better close the gap between desired outcomes and experimental design, both for bias and for evaluating language models more broadly.

A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets

*Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty and Jimmy Xiangji Huang*19:00-21:00 (Metropolitan West)
The development of large language models (LLMs) such as ChatGPT has brought a lot of attention recently. However, their evaluation in the benchmark academic datasets remains under-explored due to the difficulty of evaluating the generative outputs produced by this model against the ground truth. In this paper, we aim to present a thorough evaluation of ChatGPT's performance on diverse academic datasets, covering tasks like question-answering, text summarization, code generation, commonsense reasoning, mathematical problem-solving, machine translation, bias detection, and ethical considerations. Specifically, we evaluate ChatGPT across 140 tasks and analyze 255K responses it generates in these datasets. This makes our work the largest evaluation of ChatGPT in NLP benchmarks. In short, our study aims to validate the strengths and weaknesses of ChatGPT in various tasks and provide insights for future research using LLMs. We also report a new emergent ability to follow multi-query instructions that we mostly found in ChatGPT and other instruction-tuned models. Our extensive evaluation shows that even though ChatGPT is capable of performing a wide variety of tasks, and may obtain impressive performance in several benchmark datasets, it is still far from achieving the ability to reliably solve many challenging tasks. By providing a thorough assessment of ChatGPT's performance across diverse NLP tasks, this paper sets the stage for a targeted deployment of ChatGPT-like LLMs in real-world applications.

Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow and Yanai Elazar   19:00-21:00 (Metropolitan West)
Few-shot fine-tuning and in-context learning are two alternative strategies for task adaptation of pre-trained language models. Recently, in-context learning has gained popularity over fine-tuning due to its simplicity and improved out-of-domain generalization, and because extensive evidence shows that fine-tuned models pick up on spurious correlations. Unfortunately, previous comparisons of the two approaches were done using models of different sizes. This raises the question of whether the observed weaker out-of-domain generalization of fine-tuned models is an inherent property of fine-tuning or a limitation of the experimental setup. In this paper, we compare the generalization of few-shot fine-tuning and in-context learning to challenge datasets, while controlling for the models used, the number of examples, and the number of parameters, ranging from 125M to 30B. Our results show that fine-tuned language models can in fact generalize well out-of-domain. We find that both approaches generalize similarly; they exhibit large variation and depend on properties such as model size and the number of examples, highlighting that robust task adaptation remains a challenge.

Reproducibility in NLP: What Have We Learned from the Checklist?

Ian Magnusson, Noah A. Smith and Jesse Dodge   19:00-21:00 (Metropolitan West)
Scientific progress in NLP rests on the reproducibility of researchers' claims. The *CL conferences created the NLP Reproducibility Checklist in 2020 to be completed by authors at submission to remind them of key information to include. We provide the first analysis of the Checklist by examining 10,405 anonymous responses to it. First, we find evidence of an increase in reporting of information on efficiency, validation performance, summary statistics, and hyperparameters after the Checklist's introduction. Further, we show acceptance rate grows for submissions with more Yes responses. We find that the 44% of submissions that gather new data are 5% less likely to be accepted than those that did not; the average reviewer-rated reproducibility of these submissions is also 2% lower relative to the rest. We find that only 46% of submissions claim to open-source their code, though submissions that do have 8% higher reproducibility score relative to those that do not, the most for any item. We discuss what can be inferred about the state of reproducibility in NLP, and provide a set of recommendations for future conferences, including: a) allowing submitting code and appendices one week after the deadline, and b) measuring dataset reproducibility by a checklist of data collection practices.

GUMSum: Multi-Genre Data and Evaluation for English Abstractive Summarization

Yang Janet Liu and Amir Zeldes   19:00-21:00 (Metropolitan West)

Automatic summarization with pre-trained language models has led to impressively fluent results, but is prone to 'hallucinations', low performance on non-news genres, and outputs which are not exactly summaries. Targeting ACL 2023's 'Reality Check' theme, we present GUMSum, a small but carefully crafted dataset of English summaries in 12 written and spoken genres for evaluation of abstractive summarization. Summaries are highly constrained, focusing on substitutive potential, factuality, and faithfulness. We present guidelines and evaluate human agreement as well as subjective judgments on recent system outputs, comparing general-domain untuned approaches, a fine-tuned one, and a prompt-based approach, to human performance. Results show that while GPT3 achieves impressive scores, it still underperforms humans, with varying quality across genres. Human judgments reveal different types of errors in supervised, prompted, and human-generated summaries, shedding light on the challenges of producing a good summary.

Towards Reasoning in Large Language Models: A Survey
Jie Huang and Kevin Chen-Chuan Chang   19:00-21:00 (Metropolitan West)
Reasoning is a fundamental aspect of human intelligence that plays a crucial role in activities such as problem solving, decision making, and critical thinking. In recent years, large language models (LLMs) have made significant progress in natural language processing, and there is observation that these models may exhibit reasoning abilities when they are sufficiently large. However, it is not yet clear to what extent LLMs are capable of reasoning. This paper provides a comprehensive overview of the current state of knowledge on reasoning in LLMs, including techniques for improving and eliciting reasoning in these models, methods and benchmarks for evaluating reasoning abilities, findings and implications of previous research in this field, and suggestions on future directions. Our aim is to provide a detailed and up-to-date review of this topic and stimulate meaningful discussion and future work.

Main Conference: Tuesday, July 11, 2023

Session 3 - 09:00-10:30

Interpretability and Analysis of Models for NLP

09:00-10:30 (Metropolitan East)

Incorporating Attribution Importance for Improving Faithfulness Metrics

Zhixue Zhao and Nikolaos Aletras   09:00-09:15 (Metropolitan East)

Feature attribution methods (FAs) are popular approaches for providing insights into the model reasoning process of making predictions. The more faithful a FA is, the more accurately it reflects which parts of the input are more important for the prediction. Widely used faithfulness metrics, such as sufficiency and comprehensiveness use a hard erasure criterion, i.e. entirely removing or retaining the top most important tokens ranked by a given FA and observing the changes in predictive likelihood. However, this hard criterion ignores the importance of each individual token, treating them all equally for computing sufficiency and comprehensiveness. In this paper, we propose a simple yet effective soft erasure criterion. Instead of entirely removing or retaining tokens from the input, we randomly mask parts of the token vector representations proportionately to their FA importance. Extensive experiments across various natural language processing tasks and different FAs show that our soft-sufficiency and soft-comprehensiveness metrics consistently prefer more faithful explanations compared to hard sufficiency and comprehensiveness.

Generalizing Backpropagation for Gradient-Based Interpretability

Kevin Du, Lucas Torroba Hennigen, Niklas Stoehr, Alex Warstadt and Ryan Cotterell   09:15-09:30 (Metropolitan East)

Many popular feature-attribution methods for interpreting deep neural networks rely on computing the gradients of a model's output with respect to its inputs. While these methods can indicate which input features may be important for the model's prediction, they reveal little about the inner workings of the model itself. In this paper, we observe that the gradient computation of a model is a special case of a more general formulation using semirings. This observation allows us to generalize the backpropagation algorithm to efficiently compute other interpretable statistics about the gradient graph of a neural network, such as the highest-weighted path and entropy. We implement this generalized algorithm, evaluate it on synthetic datasets to better understand the statistics it computes, and apply it to study BERT's behavior on the subject–verb number agreement task (SVA). With this method, we (a) validate that the amount of gradient flow through a component of a model reflects its importance to a prediction and (b) for SVA, identify which pathways of the self-attention mechanism are most important.

CREST: A Joint Framework for Rationalization and Counterfactual Text Generation

Marcos Treviso, Alexis Ross, Nuno M. Guerreiro and André Martins   09:30-09:45 (Metropolitan East)

Selective rationales and counterfactual examples have emerged as two effective, complementary classes of interpretability methods for analyzing and training NLP models. However, prior work has not explored how these methods can be integrated to combine their complementary advantages. We overcome this limitation by introducing CREST (ContRastive Edits with Sparse raTionalization), a joint framework for selective rationalization and counterfactual text generation, and show that this framework leads to improvements in counterfactual quality, model robustness, and interpretability. First, CREST generates valid counterfactuals that are more natural than those produced by previous methods, and subsequently can be used for data augmentation at scale, reducing the need for human-generated examples. Second, we introduce a new loss function that leverages CREST counterfactuals to regularize selective rationales and show that this regularization improves both model robustness and rationale quality, compared to methods that do not leverage CREST counterfactuals. Our results demonstrate that CREST successfully bridges the gap between selective rationales and counterfactual examples, addressing the limitations of existing methods and providing a more comprehensive view of a model's predictions.

SCOTT: Self-Consistent Chain-of-Thought Distillation

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin and Xiang Ren   09:45-10:00 (Metropolitan East)

Large language models (LMs) beyond a certain scale, demonstrate the emergent capability of generating free-text rationales for their predictions via chain-of-thought (CoT) prompting. While CoT can yield dramatically improved performance, such gains are only observed for sufficiently large LMs. Even more concerning, there is little guarantee that the generated rationales are consistent with LM's predictions or faithfully justify the decisions. In this work, we propose SCOTT, a faithful knowledge distillation method to learn a small, self-consistent CoT model from a teacher model that is orders of magnitude larger. To form better supervision, we elicit rationales supporting the gold answers from a large LM (teacher) by contrastive decoding, which encourages the teacher to generate tokens that become more plausible only when the answer is considered. To ensure faithful distillation, we use the teacher-generated rationales to learn a student LM with a counterfactual reasoning objective, which prevents the student from ignoring the rationales to make inconsistent predictions. Experiments show that while yielding comparable performance, our method leads to a more faithful model than baselines. Further analysis shows that such a model respects the rationales more when making decisions; thus, we can improve its performance more by refining its rationales.

Are Machine Rationales (Not) Useful to Humans? Measuring and Improving Human Utility of Free-text Rationales

Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi and Xiang Ren   10:00-10:15 (Metropolitan East)

Among the remarkable emergent capabilities of large language models (LMs) is free-text rationalization; beyond certain scale, large LMs are capable of generating seemingly useful rationalizations, which in turn, can dramatically enhance their performances on leaderboards. This phenomenon raises a question: can machine generated rationales also be useful for humans, especially when lay humans try to answer questions based on those machine rationales? We observe that human utility of existing rationales is far from satisfactory and expensive to estimate with human studies. Existing metrics like task performance of the LM generating the rationales or similarity between generated and gold rationales are not good indicators of their human utility. While we observe that certain properties of rationales like conciseness and novelty are correlated with their human utility, estimating them without human involvement is challenging. We show that, by estimating a rationale's helpfulness in answering similar unseen instances, we can measure its human utility to a better extent. We also translate this finding into an automated score, Gen-U, that we propose, which can help improve LMs' ability to generate rationales with better human utility, while maintaining most of its task performance. Lastly, we release all code and collected data with this project.

Faithfulness Tests for Natural Language Explanations

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen and Isabelle Augenstein   10:15-10:30 (Metropolitan East)

Explanations of neural models aim to reveal a model's decision-making process for its predictions. However, recent work shows that current methods explaining explanations such as saliency maps or counterfactuals can be misleading, as they are prone to present reasons that lead to counterfactual predictions but are not reflected by the NLEs. Second, we reconstruct inputs from the reasons stated in the generated NLEs and check how often they lead to the same predictions. Our tests can evaluate emerging NLE models, proving a fundamental tool in the development of faithful NLEs.

Large Language Models

09:00-10:30 (Metropolitan Centre)

Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages
Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon and Hinrich Schütze 09:00-09:15 (Metropolitan Centre)
The NLP community has mainly focused on scaling Large Language Models (LLMs) vertically, i.e., making them better for about 100 languages. We instead scale LLMs horizontally: we create, through continued pretraining, Glot500-m, an LLM that covers 511 predominantly low-resource languages. An important part of this effort is to collect and clean Glot500-c, a corpus that covers these 511 languages and allows us to train Glot500-m. We evaluate Glot500-m on five diverse tasks across these languages. We observe large improvements for both high-resource and low-resource languages compared to an XLM-R baseline. Our analysis shows that no single factor explains the quality of multilingual LLM representations. Rather, a combination of factors determines quality including corpus size, script, "help" from related languages and the total capacity of the model. Our work addresses an important goal of NLP research: we should not limit NLP to a small fraction of the world's languages and instead strive to support as many languages as possible to bring the benefits of NLP technology to all languages and cultures. Code, data and models are available at https://github.com/cisnlp/Glot500.

mCLIP: Multilingual CLIP via Cross-lingual Transfer
Guanhua Chen, Lu Hou, Yun Chen, Wenliang Dai, Lifeng Shang, Xin Jiang, Qun Liu, Jia Pan and Wenping Wang 09:15-09:30 (Metropolitan Centre)
Large-scale vision-language pretrained (VLP) models like CLIP have shown remarkable performance on various downstream cross-modal tasks. However, they are usually biased towards English due to the lack of sufficient non-English image-text pairs. Existing multilingual VLP methods often learn retrieval-inefficient single-stream models by translation-augmented non-English image-text pairs. In this paper, we introduce mCLIP, a retrieval-efficient dual-stream multilingual VLP model, trained by aligning the CLIP model and a Multilingual Text Encoder (MTE) through a novel Triangle Cross-modal Knowledge Distillation (TriKD) method. It is parameter-efficient as only two light projectors on the top of them are updated during distillation. Furthermore, to enhance the token- and sentence-level multilingual representation of the MTE, we propose to train it with machine translation and contrastive learning jointly before the TriKD to provide a better initialization. Empirical results show that mCLIP achieves new state-of-the-art performance for both zero-shot and finetuned multilingual image-text retrieval task.

Preserving Commonsense Knowledge from Pre-trained Language Models via Causal Inference
Junhao Zheng, Qianli Ma, Shengjie Qiu, Yue Wu, Peitian Ma, Junlong Liu, Huawen Feng, Xichen Shang and Haibin Chen 09:30-09:45 (Metropolitan Centre)
Fine-tuning has been proven to be a simple and effective technique to transfer the learned knowledge of Pre-trained Language Models (PLMs) to downstream tasks. However, vanilla fine-tuning easily overfits the target data and degrades the generalization ability. Most existing studies attribute it to catastrophic forgetting, and they retain the pre-trained knowledge indiscriminately without identifying what knowledge is transferable. Motivated by this, we frame fine-tuning into a causal graph and discover that the crux of catastrophic forgetting lies in the missing causal effects from the pre-trained data. Based on the causal view, we propose a unified objective for fine-tuning to retrieve the causality back. Intriguingly, the unified objective can be seen as the sum of the vanilla fine-tuning objective, which learns new knowledge from target data, and the causal objective, which preserves old knowledge from PLMs. Therefore, our method is flexible and can mitigate negative transfer while preserving knowledge. Since endowing models with commonsense is a long-standing challenge, we implement our method on commonsense QA with a proposed heuristic estimation to verify its effectiveness. In the experiments, our method outperforms state-of-the-art fine-tuning methods on all six commonsense QA datasets and can be implemented as a plug-in module to inflate the performance of existing QA models.

Self-Adaptive In-Context Learning: An Information Compression Perspective for In-Context Example Selection and Ordering
Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye and Lingpeng Kong 09:45-10:00 (Metropolitan Centre)
Despite the surprising few-shot performance of in-context learning (ICL), it is still a common practice to randomly sample examples to serve as context. This paper advocates a new principle for ICL: self-adaptive in-context learning. The self-adaption mechanism is introduced to help each sample find an in-context example organization (i.e., selection and permutation) that can derive the correct prediction, thus maximizing performance. To validate the effectiveness of self-adaptive ICL, we propose a general select-then-rank framework and instantiate it with new selection and ranking algorithms. Upon extensive evaluation on eight different NLP datasets, our self-adaptive ICL method achieves a 40% relative improvement over the common practice setting. Further analysis reveals the enormous potential of self-adaptive ICL that it might be able to close the gap between ICL and finetuning given more advanced algorithms. Our code will be released to facilitate future research.

Pre-Training to Learn in Context
Yuxian Gu, Li Dong, Furu Wei and Minlie Huang 10:00-10:15 (Metropolitan Centre)
In-context learning, where pre-trained language models learn to perform tasks from task examples and instructions in their contexts, has attracted much attention in the NLP community. However, the ability of in-context learning is not fully exploited because language models are not explicitly trained to learn in context. To this end, we propose PICL (Pre-training for In-Context Learning), a framework to enhance the language models' in-context learning ability by pre-training the model on a large collection of "intrinsic tasks" in the general plain-text corpus using the simple language modeling objective. PICL encourages the model to infer and perform tasks by conditioning on the contexts while maintaining task generalization of pre-trained models. We evaluate the in-context learning performance of the model trained with PICL on seven widely-used text classification datasets and the Super-NaturalInstrctions benchmark, which contains 100+ NLP tasks formulated to text generation. Our experiments show that PICL is more effective and task-generalizable than a range of baselines, outperforming larger language models with nearly 4x parameters. The code is publicly available at https://github.com/thu-coai/PICL.

ReAugKD: Retrieval-Augmented Knowledge Distillation For Pre-trained Language Models
Jianyi Zhang, Aashiq Muhamed, Aditya Anantharaman, Guoyin Wang, Changyou Chen, Kai Zhong, Qingjun Cui, Yi Xu, Belinda Zeng, Tr-

ishul Chilimbi and Yiran Chen 10:15-10:30 (Metropolitan Centre)
Knowledge Distillation (KD) is one of the most effective approaches to deploying large-scale pre-trained language models in low-latency environments by transferring the knowledge contained in the large-scale models to smaller student models. Prior KD approaches use the soft labels and intermediate activations generated by the teacher to transfer knowledge to the student model parameters alone. In this paper, we show that having access to non-parametric memory in the form of a knowledge base with the teacher's soft labels and predictions can further improve student generalization. To enable the student to retrieve from the knowledge base effectively, we propose a new framework and loss function that preserves the semantic similarities of teacher and student training examples. We show through extensive experiments that our retrieval mechanism can achieve state-of-the-art performance for task-specific knowledge distillation on the GLUE benchmark.

Dialogue and Interactive Systems

09:00-10:30 (Metropolitan West)

GIFT: Graph-Induced Fine-Tuning for Multi-Party Conversation Understanding
Jia-Chen Gu, Zhenhua Ling, Quan Liu, Cong Liu and Guoping Hu 09:00-09:15 (Metropolitan West)
Addressing the issues of who saying what to whom in multi-party conversations (MPCs) has recently attracted a lot of research attention. However, existing methods on MPC understanding typically embed interlocutors and utterances into sequential information flows, or utilize only the superficial of inherent graph structures in MPCs. To this end, we present a plug-and-play and lightweight method named graph-induced fine-tuning (GIFT) which can adapt various Transformer-based pre-trained language models (PLMs) for universal MPC understanding. In detail, the full and equivalent connections among utterances in regular Transformer ignore the sparse but distinctive dependency of an utterance on another in MPCs. To distinguish different relationships between utterances, four types of edges are designed to integrate graph-induced signals into attention mechanisms to refine PLMs originally designed for processing sequential texts. We evaluate GIFT by implementing it into three PLMs, and test the performance on three downstream tasks including addressee recognition, speaker identification and response selection. Experimental results show that GIFT can significantly improve the performance of three PLMs on three downstream tasks and two benchmarks with only 4 additional parameters per encoding layer, achieving new state-of-the-art performance on MPC understanding.

Envisioning Future from the Past: Hierarchical Duality Learning for Multi-Turn Dialogue Generation
Ang Lv, Jinpeng Li, Shufang Xie and Rui Yan 09:15-09:30 (Metropolitan West)
In this paper, we define a widely neglected property in dialogue text, duality, which is a hierarchical property that is reflected in human behaviours in daily conversations: Based on the logic in a conversation (or a sentence), people can infer follow-up utterances (or tokens) based on the previous text, and vice versa. We propose a hierarchical duality learning for dialogue (HDLD) to simulate this human cognitive ability, for generating high quality responses that connect both previous and follow-up dialogues. HDLD utilizes hierarchical dualities at token hierarchy and utterance hierarchy. HDLD maximizes the mutual information between past and future utterances. Thus, even if future text is invisible during inference, HDLD is capable of estimating future information implicitly based on dialogue history and generates both coherent and informative responses. In contrast to previous approaches that solely utilize future text as auxiliary information to encode during training, HDLD leverages duality to enable interaction between dialogue history and the future. This enhances the utilization of dialogue data, leading to the improvement in both automatic and human evaluation.

Schema-Guided User Satisfaction Modeling for Task-Oriented Dialogues
Yue Feng, Yunlong Jiao, Animesh Prasad, Nikolaos Aletras, Emine Yilmaz and Gabriella Kazai 09:30-09:45 (Metropolitan West)
User Satisfaction Modeling (USM) is one of the popular choices for task-oriented dialogue systems evaluation, where user satisfaction typically depends on whether the user's task goals were fulfilled by the system. Task-oriented dialogue systems use task schema, which is a set of task attributes, to encode the user's task goals. Existing studies on USM neglect explicitly modeling the user's task goals fulfillment using the task schema. In this paper, we propose SG-USM, a novel schema-guided user satisfaction modeling framework. It explicitly models the degree to which the user's preferences regarding the task attributes are fulfilled by the system for predicting the user's satisfaction level. SG-USM employs a pre-trained language model for encoding dialogue context and task attributes. Further, it employs a fulfillment representation layer for learning how many task attributes have been fulfilled in the dialogue, an importance predictor component for calculating the importance of task attributes. Finally, it predicts the user satisfaction based on task attribute fulfillment and task attribute importance. Experimental results on benchmark datasets (i.e. MWOZ, SGD, ReDial, and JDDC) show that SG-USM consistently outperforms competitive existing methods. Our extensive analysis demonstrates that SG-USM can improve the interpretability of user satisfaction modeling, has good scalability as it can effectively deal with unseen tasks and can also effectively work in low-resource settings by leveraging unlabeled data.Code is available at https://github.com/amzn/user-satisfaction-modeling.

Divide, Conquer, and Combine: Mixture of Semantic-Independent Experts for Zero-Shot Dialogue State Tracking
Qingyue Wang, Liang Ding, Yanan Cao, Yibing Zhan, Zheng Lin, Shi Wang, Dacheng Tao and Li Guo 09:45-10:00 (Metropolitan West)
Zero-shot transfer learning for Dialogue State Tracking (DST) helps to handle a variety of task-oriented dialogue domains without the cost of collecting in-domain data. Existing works mainly study common data- or model-level augmentation methods to enhance the generalization but fail to effectively decouple semantics of samples, limiting the zero-shot performance of DST. In this paper, we present a simple and effective "divide, conquer and combine" solution, which explicitly disentangles the semantics of seen data, and leverages the performance and robustness with the mixture-of-experts mechanism. Specifically, we divide the seen data into semantically independent subsets and train corresponding experts, the newly unseen samples are mapped and inferred with mixture-of-experts with our designed ensemble inference. Extensive experiments on MultiWOZ2.1 upon T5-Adapter show our scheme significantly and consistently improves the zero-shot performance, achieving the SOTA on settings without external knowledge, with only 10M trainable parameters.

Towards Boosting the Open-Domain Chatbot with Human Feedback
Hua Lu, Siqi Bao, Huang He, Fan Wang, Hua Wu and Haifeng Wang 10:00-10:15 (Metropolitan West)
Many open-domain dialogue models pre-trained with social media comments can generate coherent replies but have difficulties producing engaging responses. This phenomenon might mainly result from the deficiency of annotated human-human conversations and the misalignment with human preference. In this paper, we propose a novel and efficient framework Diamante to boost the open-domain chatbot, where two kinds of human feedback (including explicit demonstration and implicit preference) are collected and leveraged. By asking annotators to select or amend the model-generated candidate responses, Diamante efficiently collects the human demonstrated responses and constructs a Chinese chit-chat dataset. To enhance the alignment with human preference, Diamante leverages the implicit preference in the data collection process and introduces the generation-evaluation joint training. Comprehensive experiments indicate that the Diamante dataset and joint training paradigm can significantly boost the performance of pre-trained dialogue models. The overall engagingness of the previous state-of-the-art model has been improved remarkably by 50% in Chinese open-domain conversations.

DiffusEmp: A Diffusion Model-Based Framework with Multi-Grained Control for Empathetic Response Generation
Guanqun Bi, Lei Shen, Yanan Cao, Meng Chen, Yuqiang Xie, Zheng Lin and Xiaodong He   10:15-10:30 (Metropolitan West)
Empathy is a crucial factor in open-domain conversations, which naturally shows one's caring and understanding to others. Though several methods have been proposed to generate empathetic responses, existing works often lead to monotonous empathy that refers to generic and safe expressions. In this paper, we propose to use explicit control to guide the empathy expression and design a framework DiffusEmp based on conditional diffusion language model to unify the utilization of dialogue context and attribute-oriented control signals. Specifically, communication mechanism, intent, and semantic frame are imported as multi-grained signals that control the empathy realization from coarse to fine levels. We then design a specific masking strategy to reflect the relationship between multi-grained signals and response tokens, and integrate it into the diffusion model to influence the generative process. Experimental results on a benchmark dataset EmpatheticDialogue show that our framework outperforms competitive baselines in terms of controllability, informativeness, and diversity without the loss of context-relatedness.

Posters
09:00-10:30 (Frontenac Ballroom and Queen's Quay)

[TACL] Helpful Neighbors: Leveraging Neighbors in Geographic Feature Pronunciation
Richard Sproat, Llion Jones, Haruko Ishikawa and Alexander Gutkin   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
If one sees the place name Houston Mercer Dog Run in New York, how does one know how to pronounce it? Assuming one knows that Houston in New York is pronounced not like the Texas city, then one can probably guess that the same pronunciation is also used in the name of the dog park. We present a novel architecture that learns to use the pronunciations of neighboring names in order to guess the pronunciation of a given target feature. Applied to Japanese place names, we demonstrate the utility of the model to finding and proposing corrections for errors in Google Maps. To demonstrate the utility of this approach to structurally similar problems, we also report on an application to a totally different task: Cognate reflex prediction in comparative historical linguistics. A version of the code has been open-sourced.

[TACL] Unleashing the True Potential of Sequence-to-Sequence Models for Sequence Tagging and Structure Parsing
Han He and Jinho Choi   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Sequence-to-Sequence (S2S) models have achieved remarkable success on various text generation tasks. However, learning complex structures with S2S models remains challenging as external neural modules and additional lexicons are often supplemented to predict non-textual outputs. We present a systematic study of S2S modeling using contained decoding on four core tasks: part-of-speech tagging, named entity recognition, constituency and dependency parsing, to develop efficient exploitation methods costing zero extra parameters. In particular, 3 lexically diverse linearization schemas and corresponding constrained decoding methods are designed and evaluated. Experiments show that although more lexicalized schemas yield longer output sequences that require heavier training, their sequences being closer to natural language makes them easier to learn. Moreover, S2S models using our constrained decoding outperform other S2S approaches using external resources. Our best models perform better than or comparably to the state-of-the-art for all 4 tasks, lighting a promise for S2S models to generate non-sequential structures.

[TACL] Design Choices for Crowdsourcing Implicit Discourse Relations: Revealing the Biases introduced by Task Design
Valentina Pyatkin, Frances Yung, Merel Scholman, Ido Dagan, Reut Tsarfaty and Vera Demberg   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Disagreement in natural language annotation has been predominantly studied from a perspective of biases introduced by the annotators and the annotation frameworks. Here, wepropose to analyze another source of bias: task design bias,which has a particularly strong impact on crowdsourcing linguistic annotations where natural language is used to elicit the interpretation of laymen annotators. For this purpose we look at implicit discourse relation annotation, a task that has repeatedly been shown to be difficult due to the relations' ambiguity. We compare the annotations of 1,200 discourse relations obtained using two distinct discourse relation annotation tasks and quantify the biases of both methods across four different domains. Both methods are natural language annotation tasks designed for crowdsourcing: one reformulating discourse segments as questions and answers and one using discourse connective insertion. We show that the task design can push annotators towards certain annotations and that some discourse relations senses can be better elicited with the one or the other NL annotation approach.We also conclude that this type of bias should be taken into account when training and testing models.

[SRW] Constructing Multilingual Code Search Dataset Using Neural Machine Translation
Ryo Sekizawa, Nan Duan, Shuai Lu and Hitomi Yanaka   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We create a multilingual code search dataset by translating the existing English dataset with a machine translation model and conduct a baseline experiment on a code search task.

[SRW] Multimodal Neural Machine Translation Using Synthetic Images Transformed by Latent Diffusion Model
*Ryoya Yuasa, Akihiro Tamura, Tomoyuki Kajiwara, Takashi Ninomiya and Tsuneo Kato*09:00-10:30 (Frontenac Ballroom and Queen's Quay)
This study proposes a new multimodal neural machine translation model using synthetic images transformed by a latent diffusion model.

[SRW] Predicting Human Translation Difficulty Using Automatic Word Alignment
Zheng Wei Lim, Trevor Cohn, Charles Kemp and Ekaterina Vylomova   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
In this work, we use word alignments computed over large scale bilingual corpora to develop predictors of lexical translation difficulty.

[SRW] Is Anisotropy Inherent to Transformers?
Nathan Godey, Éric De La Clergerie and Benoît Sagot   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
This paper suggests that anisotropy may be an inherent property of Transformers-based models, and extends to other modalities beyond token-based NLP.

[SRW] Geometric Locality of Entity Embeddings in Masked Language Models
Masaki Sakata, Sho Yokoi, Benjamin Heinzerling and Kentaro Inui   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
The paper assesses whether masked language models can distinguish entities in their internal representations, finding they can to a certain degree, even when there was variation in the surrounding context and mentions.

[SRW] Native Language Prediction from Gaze: a Reproducibility Study
Lina Skerath, Paulina Toborek, Anita Zielińska, Maria Barrett and Rob Van Der Goot 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
A reproduction study of native language prediction from English as second language reading eye-tracking data.

[SRW] Sudden Semantic Shifts in Swedish NATO discourse
Brian Bonafilia, Bastiaan Bruinsma, Denitsa Saynova and Moa Johansson   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We look at short-term semantic shifts in the Swedish discussion about NATO membership.

[SRW] Choosing What to Mask: More Informed Masking for Multimodal Machine Translation
Julia Sato, Helena Caseli and Lucia Specia   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
More informed masking in cross-lingual visual pre-training for multimodal machine translation.

[SRW] Transformer Language Models Handle Word Frequency in Prediction Head
Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi and Kentaro Inui   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We reveal that the bias parameters in the word prediction head of Transformer LMs play a significant role in the model's ability to reflect corpus word frequency.

[SRW] Jamp: Controlled Japanese Temporal Inference Dataset for Evaluating Generalization Capacity of Language Models
Tomoki Sugimoto, Yasumasa Onoe and Hitomi Yanaka   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We construct Jamp, which is a Japanese NLI dataset for temporal inference, and evaluate the generalization capacity of several LMs on our dataset.

[Demo] YANMTT: Yet Another Neural Machine Translation Toolkit
Eiichiro Sumita, Chinmay Sawant, Diptesh Kanojia and Raj Dabre   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
In this paper, we present our open-source neural machine translation (NMT) toolkit called "Yet Another Neural Machine Translation Toolkit" abbreviated as YANMTT - https://github.com/prajdabre/yanmtt, which is built on top of the HuggingFace Transformers library. YANMTT focuses on transfer learning and enables easy pre-training and fine-tuning of sequence-to-sequence models at scale. It can be used for training parameter-heavy models with minimal parameter sharing and efficient, lightweight models via heavy parameter sharing. Additionally, it supports parameter-efficient fine-tuning (PEFT) through adapters and prompts. Our toolkit also comes with a user interface that can be used to demonstrate these models and visualize various parts of the model. Apart from these core features, our toolkit also provides other advanced functionalities such as but not limited to document/multi-source NMT, simultaneous NMT, mixtures-of-experts, model compression and continual learning.

[Demo] CARE: Collaborative AI-Assisted Reading Environment
Iryna Gurevych, Ilia Kuznetsov, Jan Buchmann, Nils Dycke and Dennis Zyska   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Recent years have seen impressive progress in AI-assisted writing, yet the developments in AI-assisted reading are lacking. We propose inline commentary as a natural vehicle for AI-based reading assistance, and present CARE: the first open integrated platform for the study of inline commentary and reading. CARE facilitates data collection for inline commentaries in a commonplace collaborative reading environment, and provides a framework for enhancing reading with NLP-based assistance, such as text classification, generation or question answering. The extensible behavioral logging allows unique insights into the reading and commenting behavior, and flexible configuration makes the platform easy to deploy in new scenarios. To evaluate CARE in action, we apply the platform in a user study dedicated to scholarly peer review. CARE facilitates the data collection and study of inline commentary in NLP, extrinsic evaluation of NLP assistance, and application prototyping. We invite the community to explore and build upon the open source implementation of CARE.

Github Repository: https://github.com/UKPLab/CARE Public Live Demo: https://care.ukp.informatik.tu-darmstadt.de

[Demo] LaTeX2Solver: a Hierarchical Semantic Parsing of LaTeX Document into Code for an Assistive Optimization Modeling Application
Yong Zhang, Kun Mao, Ren Li, Yuanzhe Chen, Xiongwei Han, Xiaojin Fu, Xiaorui Li, Mahdi Mostajabdaveh, Linzi Xing, Timothy Yu and Rindra Ramamonjison   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We demonstrate an interactive system to help operations research (OR) practitioners convert the mathematical formulation of optimization problems from TeX document format into the solver modeling language. In practice, a manual translation is cumbersome and time-consuming. Moreover, it requires an in-depth understanding of the problem description and a technical expertise to produce the modeling code. Thus, our proposed system TeX2Solver helps partially automate this conversion and help the users build optimization models more efficiently. In this paper, we describe its interface and the components of the hierarchical parsing system. A video demo walk-through is available online at http://bit.ly/3kuOm3x

[Demo] Disease Network Constructor: a Pathway Extraction and Visualization
Hiroya Takamura, Mari Itoh, Masakata Kuroda, Yayoi Natsume-Kitatani, Nozomi Nagano, Masami Ikeda, Goran Topić, Khoa Duong and Mohammad Golam Sohrab   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We present Disease Network Constructor (DNC), a system that extracts and visualizes a disease network, in which nodes are entities such as diseases, proteins, and genes, and edges represent regulation relation. We focused on the disease network derived through regulation events found in scientific articles on idiopathic pulmonary fibrosis (IPF). The front-end web-base user interface of DNC includes two-dimensional (2D) and 3D visualizations of the constructed disease network. The back-end system of DNC includes several natural language processing (NLP) techniques to process biomedical text including BERT-based tokenization on the basis of Bidirectional Encoder Representations from Transformers (BERT), flat and nested named entity recognition (NER), candidate generation and candidate ranking for entity linking (EL) or, relation extraction (RE), and event extraction (EE) tasks. We evaluated the end-to-end EL and end-to-end nested EE systems to determine the DNC's back-end implementation performance. To the best of our knowledge, this is the first attempt that addresses neural NER, EL, RE, and EE tasks in an end-to-end manner that constructs a path-way visualization from events, which we name Disease Network Constructor. The demonstration video can be accessed from https://youtu.be/rFhWwAgcXE8. We release an online system for end users and the source code is available at https://github.com/aistairc/PRISM-APIs/.

[Demo] GAIA Search: Hugging Face and Pyserini Interoperability for NLP Training Data Exploration
Jimmy Lin, Martin Potthast, Stella Biderman, Hailey Schoelkopf, Xinyu Zhang, Akintunde Oladipo, Christopher Akiki, Odunayo Ogundepo and Aleksandra Piktus   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Noticing the urgent need to provide tools for fast and user-friendly qualitative analysis of large-scale textual corpora of the modern NLP, we propose to turn to the mature and well-tested methods from the domain of Information Retrieval (IR) - a research field with a long history of tackling TB-scale document collections. We discuss how Pyserini - a widely used toolkit for reproducible IR research can be integrated with the Hugging Face ecosystem of open-source AI libraries and artifacts. We leverage the existing functionalities of both platforms while proposing novel features further facilitating their integration. Our goal is to give NLP researchers tools that will allow them to develop retrieval-based instrumentation for their data analytics needs with ease and agility. We include a Jupyter Notebook-based walk through the core interoperability features, available on GitHub: https://github.com/huggingface/gaia. We then demonstrate how the ideas we present can be operationalized to create a powerful tool for qualitative data analysis in NLP. We present GAIA Search - a search engine built following pre-

viously laid out principles, giving access to four popular large-scale text collections. GAIA serves a dual purpose of illustrating the potential of methodologies we discuss but also as a standalone qualitative analysis tool that can be leveraged by NLP researchers aiming to understand datasets prior to using them in training. GAIA is hosted live on Hugging Face Spaces: https://huggingface.co/spaces/spacerini/gaia.

[Demo] Lingxi: A Diversity-aware Chinese Modern Poetry Generation System
Xiaobing Li, Jiafeng Liu, Maosong Sun and Xinran Zhang   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Chinese modern poetry generation has been a challenging task. One issue is the Chinese word segmentation (CWS) which is critical to comprehend the Chinese language but was not always considered in common tokenization methods. Another is the decoding (sampling) method which may induce repetition and boredom and severely lower the diversity of the generated poetry. To address these issues, we present Lingxi, a diversity-aware Chinese modern poetry generation system. For the CWS issue, we propose a novel framework that incorporates CWS in the tokenization process. The proposed method can achieve a high vocabulary coverage rate with a reasonable vocabulary size. For the decoding method and the diversity issue, we propose a novel sampling algorithm that flattens the high likelihood part of the predicted distribution of the language model to emphasize the comparatively low-likelihood words and increase the diversity of generated poetry. Empirical results show that even when the top 60

[Demo] OpenSLU: A Unified, Modularized, and Extensible Toolkit for Spoken Language Understanding
Wanxiang Che, Yunlong Feng, Xiao Xu, Qiguang Chen and Libo Qin   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Spoken Language Understanding (SLU) is one of the core components of a task-oriented dialogue system, which aims to extract the semantic meaning of user queries (e.g., intents and slots). In this work, we introduce OpenSLU, an open-source toolkit to provide a unified, modularized, and extensible toolkit for spoken language understanding. Specifically, OpenSLU unifies 10 SLU models for both single-intent and multi-intent scenarios, which support both non-pretrained and pretrained models simultaneously. Additionally, OpenSLU is highly modularized and extensible by decomposing the model architecture, inference, and learning process into reusable modules, which allows researchers to quickly set up SLU experiments with highly flexible configurations. OpenSLU is implemented based on PyTorch, and released at https://github.com/LightChen233/OpenSLU.

[Demo] DIAGRAPH: An Open-Source Graphic Interface for Dialog Flow Design
Ngoc Thang Vu, Lindsey Vanderlyn and Dirk Väth   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
In this work, we present DIAGRAPH, an open-source graphical dialog flow editor built on the ADVISER toolkit. Our goal for this tool is threefold: 1) To support subject-experts to intuitively create complex and flexible dialog systems, 2) To support rapid prototyping of dialog system behavior, e.g., for research, and 3) To provide a hands-on test bed for students learning about dialog systems. To facilitate this, DIAGRAPH aims to provide a clean and intuitive graphical interface for creating dialog systems without requiring any coding knowledge. Once a dialog graph has been created, it is automatically turned into a dialog system using state of the art language models. This allows for rapid prototyping and testing. Dialog designers can then distribute a link to their finished dialog system or embed it into a website. Additionally, to support scientific experiments and data collection, dialog designers can access chat logs. Finally, to verify the usability of DIAGRAPH, we performed evaluation with subject-experts who extensively worked with the tool and users testing it for the first time, receiving above average System Usability Scale (SUS) scores from both (82 out 100 and 75 out of 100, respectively). In this way, we hope DIAGRAPH helps reduce the barrier to entry for creating dialog interactions.

[Demo] OpenRT: An Open-source Framework for Reasoning Over Tabular Data
Dragomir Radev, Arman Cohan, Minghao Guo, Linyong Nan, Zhenting Qi, Boyu Mi and Yilun Zhao   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
There are a growing number of table pre-training methods proposed for reasoning over tabular data (e.g., question answering, fact checking, and faithful text generation). However, most existing methods are benchmarked solely on a limited number of datasets, varying in configuration, which leads to a lack of unified, standardized, fair, and comprehensive comparison between methods. This paper presents OpenRT, the first open-source framework for reasoning over tabular data, to reproduce existing table pre-training models for performance comparison and develop new models quickly. We implemented and compared six table pre-training models on four question answering, one fact checking, and one faithful text generation datasets. Moreover, to enable the community to easily construct new table reasoning datasets, we developed TaRAT, an annotation tool which supports multi-person collaborative annotations for various kinds of table reasoning tasks. The researchers are able to deploy the newly-constructed dataset to OpenRT and compare the performances of different baseline systems.

[Demo] Effidit: An Assistant for Improving Writing Efficiency
Piji Li, Yan Wang, Guoping Huang, Chenyan Huang, Longyue Wang, Kaiqiang Song, Duyu Tang, Haiyun Jiang, Xinting Huang, Leyang Cui, Deng Cai, Wei Bi, Enbo Zhao and Shuming Shi   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Writing assistants are valuable tools that can help writers improve their writing skills. We introduce Effidit (**Eff**icient and **I**ntelligent **Edit**ing), a digital writing assistant that facilitates users to write higher-quality text more efficiently through the use of Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies. We significantly expand the capacities of a writing assistant by providing functions in three modules: text completion, hint recommendation, and writing refinement. Based on the above efforts, Effidit can efficiently assist users in creating their own text. Effidit has been deployed to several Tencent products and publicly released at https://effidit.qq.com/.

An Open Dataset and Model for Language Identification
Laurie V Burchell, Alexandra Birch, Nikolay Bogoychev and Kenneth Heafield   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Language identification (LID) is a fundamental step in many natural language processing pipelines. However, current LID systems are far from perfect, particularly on lower-resource languages. We present a LID model which achieves a macro-average F1 score of 0.93 and a false positive rate of 0.033% across 201 languages, outperforming previous work. We achieve this by training on a curated dataset of monolingual data, which we audit manually to ensure reliability. We make both the model and the dataset available to the research community. Finally, we carry out detailed analysis into our model's performance, both in comparison to existing open models and by language class.

Rethinking Annotation: Can Language Learners Contribute?
Haneul Yoo, Rifki Afina Putri, Changyoon Lee, Youngin Lee, So-Yeon Ahn, Dongyeop Kang and Alice Oh 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Researchers have traditionally recruited native speakers to provide annotations for the widely used benchmark datasets. But there are languages for which recruiting native speakers is difficult, and it would help to get learners of those languages to annotate the data. In this paper, we investigate whether language learners can contribute annotations to the benchmark datasets. In a carefully controlled annotation experiment, we recruit 36 language learners, provide two types of additional resources (dictionaries and machine-translated sentences), and perform mini-tests to measure their language proficiency. We target three languages, English, Korean, and Indonesian, and four NLP tasks, sentiment analysis, natural language inference, named entity recognition, and machine reading comprehension. We find that language learners, especially those with intermediate or advanced language proficiency, are able to provide fairly accurate labels with the help of additional resources. Moreover, we show that data annotation improves learners' language proficiency in terms of vocabulary and grammar. The implication of our findings is that broadening the annotation task to include language learners can open up the opportunity to build benchmark datasets for languages for which it is difficult to recruit native speakers.

Environmental Claim Detection

Dominik Stammbach, Nicolas Webersinke, Julia Anna Bingler, Mathias Kraus and Markus Leippold 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

To transition to a green economy, environmental claims made by companies must be reliable, comparable, and verifiable. To analyze such claims at scale, automated methods are needed to detect them in the first place. However, there exist no datasets or models for this. Thus, this paper introduces the task of environmental claim detection. To accompany the task, we release an expert-annotated dataset and models trained on this dataset. We preview one potential application of such models: We detect environmental claims made in quarterly earning calls and find that the number of environmental claims has steadily increased since the Paris Agreement in 2015.

Towards standardizing Korean Grammatical Error Correction: Datasets and Annotation

Soyoung Yoon, Sungjoon Park, Gyuwan Kim, Junhee Cho, Kihyo Park, Gyu Tae Kim, Minjoon Seo and Alice Oh 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Research on Korean grammatical error correction (GEC) is limited, compared to other major languages such as English. We attribute this problematic circumstance to the lack of a carefully designed evaluation benchmark for Korean GEC. In this work, we collect three datasets from different sources (Kor-Lang8, Kor-Native, and Kor-Learner) that covers a wide range of Korean grammatical errors. Considering the nature of Korean grammar, We then define 14 error types for Korean and provide KAGAS (Korean Automatic Grammatical error Annotation System), which can automatically annotate error types from parallel corpora. We use KAGAS on our datasets to make an evaluation benchmark for Korean, and present baseline models trained from our datasets. We show that the model trained with our datasets significantly outperforms the currently used statistical Korean GEC system (Hanspell) on a wider range of error types, demonstrating the diversity and usefulness of the datasets. The implementations and datasets are open-sourced.

TeCS: A Dataset and Benchmark for Tense Consistency of Machine Translation

Yiming Ai, Zhiwei He, Kai Yu and Rui Wang 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Tense inconsistency frequently occurs in machine translation. However, there are few criteria to assess the model's mastery of tense prediction from a linguistic perspective. In this paper, we present a parallel tense test set, containing French-English 552 utterances. We also introduce a corresponding benchmark, tense prediction accuracy. With the tense test set and the benchmark, researchers are able to measure the tense consistency performance of machine translation systems for the first time.

FERMAT: An Alternative to Accuracy for Numerical Reasoning

Jasivan Alex Sivakumar and Nafise Sadat Moosavi 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

While pre-trained language models achieve impressive performance on various NLP benchmarks, they still struggle with tasks that require numerical reasoning. Recent advances in improving numerical reasoning are mostly achieved using very large language models that contain billions of parameters and are not accessible to everyone. In addition, numerical reasoning is measured using a single score on existing datasets. As a result, we do not have a clear understanding of the strengths and shortcomings of existing models on different numerical reasoning aspects and therefore, potential ways to improve them apart from scaling them up. Inspired by CheckList (Ribeiro et al., 2020), we introduce a multi-view evaluation set for numerical reasoning in English, called FERMAT. Instead of reporting a single score on a whole dataset, FERMAT evaluates models on various key numerical reasoning aspects such as number understanding, mathematical operations, and training dependency. Apart from providing a comprehensive evaluation of models on different numerical reasoning aspects, FERMAT enables a systematic and automated generation of an arbitrarily large training or evaluation set for each aspect.The datasets and codes are publicly available to generate further multi-view data for ulterior tasks and languages.

Towards Leaving No Indic Language Behind: Building Monolingual Corpora, Benchmark and Models for Indic Languages

Sumanth Doddapaneni, Rahul Aralikatte, Gowthan Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan and Pratyush Kumar 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Building Natural Language Understanding (NLU) capabilities for Indic languages, which have a collective speaker base of more than one billion speakers is absolutely crucial. In this work, we aim to improve the NLU capabilities of Indic languages by making contributions along 3 important axes (i) monolingual corpora (ii) NLU testsets (iii) multilingual LLMs focusing on Indic languages. Specifically, we curate the largest monolingual corpora, IndicCorp, with 20.9B tokens covering 24 languages from 4 language families - a 2.3x increase over prior work, while supporting 12 additional languages. Next, we create a human-supervised benchmark, IndicXTREME, consisting of nine diverse NLU tasks covering 20 languages. Across languages and tasks, IndicXTREME contains a total of 105 evaluation sets, of which 52 are new contributions to the literature. To the best of our knowledge, this is the first effort towards creating a standard benchmark for Indic languages that aims to test the multilingual zero-shot capabilities of pretrained language models. Finally, we train IndicBERT v2, a state-of-the-art model supporting all the languages. Averaged across languages and tasks, the model achieves an absolute improvement of 2 points over a strong baseline. The data and models are available at https://github.com/AI4Bharat/IndicBERT.

ELQA: A Corpus of Metalinguistic Questions and Answers about English

Shabnam Behzad, Keisuke Sakaguchi, Nathan Schneider and Amir Zeldes 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

We present ELQA, a corpus of questions and answers in and about the English language. Collected from two online forums, the >70k questions (from English learners and others) cover wide-ranging topics including grammar, meaning, fluency, and etymology. The answers include descriptions of general properties of English vocabulary and grammar as well as explanations about specific (correct and incorrect) usage examples. Unlike most NLP datasets, this corpus is metalinguistic—it consists of language about language. As such, it can facilitate investigations of the metalinguistic capabilities of NLU models, as well as educational applications in the language learning domain. To study this, we define a free-form question answering task on our dataset and conduct evaluations on multiple LLMs (Large Language Models) to analyze their capacity to generate metalinguistic answers.

Transfer and Active Learning for Dissonance Detection: Addressing the Rare-Class Challenge

Vasudha Varadarajan, Swanie Juhng, Syeda Mahwish, Xiaoran Liu, Jonah Keith Luby, Christian C. Luhmann and H. Andrew Schwartz 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

While transformer-based systems have enabled greater accuracies with fewer training examples, data acquisition obstacles still persist for rare-class tasks – when the class label is very infrequent (e.g. < 5% of samples). Active learning has in general been proposed to alleviate such challenges, but choice of selection strategy, the criteria by which rare-class examples are chosen, has not been systematically evaluated. Further, transformers enable iterative transfer-learning approaches. We propose and investigate transfer- and active learning solutions to the rare class problem of dissonance detection through utilizing models trained on closely related tasks and the evaluation of acquisition strategies, including a proposed probability-of-rare-class (PRC) approach. We perform these experiments for a specific rare-class problem: collecting language samples of cognitive dissonance from social media. We find that PRC is a simple and effective strategy to guide annotations and ultimately improve model accuracy while transfer-learning in a specific order can improve the cold-start performance of the learner but does not benefit iterations of active learning.

VSTAR: A Video-grounded Dialogue Dataset for Situated Semantic Understanding with Scene and Topic Transitions

Yuxuan Wang, Zilong Zheng, Xueliang Zhao, Jinpeng Li, Yueqian Wang and Dongyan Zhao 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Video-grounded dialogue understanding is a challenging problem that requires machine to perceive, parse and reason over situated semantics extracted from weakly aligned video and dialogues. Most existing benchmarks treat both modalities the same as a frame-independent visual understanding task, while neglecting the intrinsic attributes in multimodal dialogues, such as scene and topic transitions. In this paper, we present **Video-grounded Scene&Topic AwaRe dialogue (VSTAR)** dataset, a large scale video-grounded dialogue understanding dataset based on 395 TV series. Based on VSTAR, we propose two benchmarks for video-grounded dialogue understanding: scene segmentation and topic segmentation, and one benchmark for video-grounded dialogue generation. Comprehensive experiments are performed on these benchmarks to demonstrate the importance of multimodal information and segments in video-grounded dialogue understanding and generation.

DarkBERT: A Language Model for the Dark Side of the Internet

Youngjin Jin, Eugene Jang, Jian Cui, Jin-Woo Chung, Yongjae Lee and Seungwon Shin 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Recent research has suggested that there are clear differences in the language used in the Dark Web compared to that of the Surface Web. As studies on the Dark Web commonly require textual analysis of the domain, language models specific to the Dark Web may provide valuable insights to researchers. In this work, we introduce DarkBERT, a language model pretrained on Dark Web data. We describe the steps taken to filter and compile the text data used to train DarkBERT to combat the extreme lexical and structural diversity of the Dark Web that may be detrimental to building a proper representation of the domain. We evaluate DarkBERT and its vanilla counterpart along with other widely used language models to validate the benefits that a Dark Web domain specific model offers in various use cases. Our evaluations show that DarkBERT outperforms current language models and may serve as a valuable resource for future research on the Dark Web.

READIN: A Chinese Multi-Task Benchmark with Realistic and Diverse Input Noises

Chenglei Si, Zhengyan Zhang, Yingfa Chen, Xiaozhi Wang, Zhiyuan Liu and Maosong Sun 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

For many real-world applications, the user-generated inputs usually contain various noises due to speech recognition errors caused by linguistic variations or typographical errors (typos). Thus, it is crucial to test model performance on data with realistic input noises to ensure robustness and fairness. However, little study has been done to construct such benchmarks for Chinese, where various language-specific input noises happen in the real world. In order to fill this important gap, we construct READIN: a Chinese multi-task benchmark with REalistic And Diverse Input Noises. READIN contains four diverse tasks and requests annotators to re-enter the original test data with two commonly used Chinese input methods: Pinyin input and speech input. We designed our annotation pipeline to maximize diversity, for example by instructing the annotators to use diverse input method editors (IMEs) for keyboard noises and recruiting speakers from diverse dialectical groups for speech noises. We experiment with a series of strong pretrained language models as well as robust training methods, we find that these models often suffer significant performance drops on READIN even with robustness methods like data augmentation. As the first large-scale attempt in creating a benchmark with noises geared towards user-generated inputs, we believe that READIN serves as an important complement to existing Chinese NLP benchmarks. The source code and dataset can be obtained from https://github.com/ thunlp/READIN.

A Study on the Efficiency and Generalization of Light Hybrid Retrievers

Man Luo, Shashank Jain, Anchit Gupta, Arash Einolghozati, Barlas Oguz, Debojeet Chatterjee, Xilun Chen, Chitta Baral and Peyman Heidari 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Hybrid retrievers can take advantage of both sparse and dense retrievers. Previous hybrid retrievers leverage indexing-heavy dense retrievers. In this work, we study "Is it possible to reduce the indexing memory of hybrid retrievers without sacrificing performance"? Driven by this question, we leverage an indexing-efficient dense retriever (i.e. DrBoost) and introduce a LITE retriever that further reduces the memory of DrBoost. LITE is jointly trained on contrastive learning and knowledge distillation from DrBoost. Then, we integrate BM25, a sparse retriever, with either LITE or DrBoost to form light hybrid retrievers. Our Hybrid-LITE retriever saves $13\times$ memory while maintaining 98.0% performance of the hybrid retriever of BM25 and DPR. In addition, we study the generalization capacity of our light hybrid retrievers on out-of-domain dataset and a set of adversarial attacks datasets. Experiments showcase that light hybrid retrievers achieve better generalization performance than individual sparse and dense retrievers. Nevertheless, our analysis shows that there is a large room to improve the robustness of retrievers, suggesting a new research direction.

Self-Edit: Fault-Aware Code Editor for Code Generation

Kechi Zhang, Zhuo Li, Jia Li, Ge Li and Zhi Jin 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Large language models (LLMs) have demonstrated an impressive ability to generate codes on competitive programming tasks. However, with limited sample numbers, LLMs still suffer from poor accuracy. Inspired by the process of human programming, we propose a generate-and-edit approach named Self-Edit that utilizes execution results of the generated code from LLMs to improve the code quality on the competitive programming task. We execute the generated code on the example test case provided in the question and wrap execution results into a supplementary comment. Utilizing this comment as guidance, our fault-aware code editor is employed to correct errors in the generated code. We perform extensive evaluations across two competitive programming datasets with nine different LLMs. Compared to directly generating from LLMs, our approach can improve the average of pass@1 by 89% on APPS-dev, 31% on APPS-test, and 48% on HumanEval over nine popular code generation LLMs with parameter sizes ranging from 110M to 175B. Compared to other post-processing methods, our method demonstrates superior accuracy and efficiency.

A Simple and Effective Framework for Strict Zero-Shot Hierarchical Classification

Rohan Bhambhoria, Lei Chen and Xiaodan Zhu 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

In recent years, large language models (LLMs) have achieved strong performance on benchmark tasks, especially in zero or few-shot settings. However, these benchmarks often do not adequately address the challenges posed in the real-world, such as that of hierarchical classification. In order to address this challenge, we propose refactoring conventional tasks on hierarchical datasets into a more indicative long-tail prediction task. We observe LLMs are more prone to failure in these cases. To address these limitations, we propose the use of entailment-contradiction prediction in conjunction with LLMs, which allows for strong performance in a strict zero-shot setting. Importantly, our method does not require any parameter updates, a resource-intensive process and achieves strong performance across multiple datasets.

A Causal Framework to Quantify the Robustness of Mathematical Reasoning with Language Models

Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schoelkopf and Mrinmaya Sachan 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

We have recently witnessed a number of impressive results on hard mathematical reasoning problems with language models. At the same time, the robustness of these models has also been called into question; recent works have shown that models can rely on shallow patterns in the problem description when generating a solution. Building on the idea of behavioral testing, we propose a novel framework, which pins down the causal effect of various factors in the input, e.g., the surface form of the problem text, the operands, and math operators on the output solution. By grounding the behavioral analysis in a causal graph describing an intuitive reasoning process, we study the behavior of language models in terms of robustness and sensitivity to direct interventions in the input space. We apply our framework on a test bed of math word

problems. Our analysis shows that robustness does not appear to continuously improve as a function of size, but the GPT-3 Davinci models (175B) achieve a dramatic improvement in both robustness and sensitivity compared to all other GPT variants.

MIReAD: Simple Method for Learning High-quality Representations from Scientific Documents

Anastasiia Razdaibiedina and Aleksandr V. Brechalov   09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Learning semantically meaningful representations from scientific documents can facilitate academic literature search and improve performance of recommendation systems. Pretrained language models have been shown to learn rich textual representations, yet they cannot provide powerful document-level representations for scientific articles. We propose MIReAD, a simple method that learns highquality representations of scientific papers by fine-tuning transformer model to predict the target journal class based on the abstract. We train MIReAD on more than 500,000 PubMed and arXiv abstracts across over 2,000 journal classes. We show that MIReAD produces representations that can be used for similar papers retrieval, topic categorization and literature search. Our proposed approach outperforms six existing models for representation learning on scientific documents across four evaluation standards.

MPCHAT: Towards Multimodal Persona-Grounded Conversation

Jaewoo Ahn, Yeda Song, Sangdoo Yun and Gunhee Kim   09:00-10:30 (Frontenac Ballroom and Queen's Quay)

In order to build self-consistent personalized dialogue agents, previous research has mostly focused on textual persona that delivers personal facts or personalities. However, to fully describe the multi-faceted nature of persona, image modality can help better reveal the speaker's personal characteristics and experiences in episodic memory (Rubin et al., 2003; Conway, 2009). In this work, we extend persona-based dialogue to the multimodal domain and make two main contributions. First, we present the first multimodal persona-based dialogue dataset named MPCHAT, which extends persona with both text and images to contain episodic memories. Second, we empirically show that incorporating multimodal persona, as measured by three proposed multimodal persona-grounded dialogue tasks (i.e., next response prediction, grounding persona prediction, and speaker identification), leads to statistically significant performance improvements across all tasks. Thus, our work highlights that multimodal persona is crucial for improving multimodal dialogue comprehension, and our MPCHAT serves as a high-quality resource for this research.

Span-Selective Linear Attention Transformers for Effective and Robust Schema-Guided Dialogue State Tracking

Björn Bebensee and Haejun Lee   09:00-10:30 (Frontenac Ballroom and Queen's Quay)

In schema-guided dialogue state tracking models estimate the current state of a conversation using natural language descriptions of the service schema for generalization to unseen services. Prior generative approaches which decode slot values sequentially do not generalize well to variations in schema, while discriminative approaches separately encode history and schema and fail to account for inter-slot and intent-slot dependencies. We introduce SPLAT, a novel architecture which achieves better generalization and efficiency than prior approaches by constraining outputs to a limited prediction space. At the same time, our model allows for rich attention among descriptions and history while keeping computation costs constrained by incorporating linear-time attention. We demonstrate the effectiveness of our model on the Schema-Guided Dialogue (SGD) and MultiWOZ datasets. Our approach significantly improves upon existing models achieving 85.3 JGA on the SGD dataset. Further, we show increased robustness on the SGD-X benchmark: our model outperforms the more than 30x larger D3ST-XXL model by 5.0 points.

Query-Efficient Black-Box Red Teaming via Bayesian Optimization

Deokjae Lee, JunYeong Lee, Jung-Woo Ha, Jin-Hwa Kim, Sang-Woo Lee, Hwaran Lee and Hyun Oh Song 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

The deployment of large-scale generative models is often restricted by their potential risk of causing harm to users in unpredictable ways. We focus on the problem of black-box red teaming, where a red team generates test cases and interacts with the victim model to discover a diverse set of failures with limited query access. Existing red teaming methods construct test cases based on human supervision or language model (LM) and query all test cases in a brute-force manner without incorporating any information from past evaluations, resulting in a prohibitively large number of queries. To this end, we propose *Bayesian red teaming* (BRT), novel query-efficient black-box red teaming methods based on Bayesian optimization, which iteratively identify diverse positive test cases leading to model failures by utilizing the pre-defined user input pool and the past evaluations. Experimental results on various user input pools demonstrate that our method consistently finds a significantly larger number of diverse positive test cases under the limited query budget than the baseline methods. The source code is available at https://github.com/snu-mllab/Bayesian-Red-Teaminghttps://github.com/snu-mllab/Bayesian-Red-Teaming.

SIMMC-VR: A Task-oriented Multimodal Dialog Dataset with Situated and Immersive VR Streams

*Te-Lin Wu, Satwik Kottur, Andrea Madotto, Mahmoud Azab, Pedro Rodriguez, Babak Damavandi, Nanyun Peng and Seungwhan Moon*09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Building an AI assistant that can seamlessly converse and instruct humans, in a user-centric situated scenario, requires several essential abilities: (1) spatial and temporal understanding of the situated and real-time user scenes, (2) capability of grounding the actively perceived visuals of users to conversation contexts, and (3) conversational reasoning over past utterances to perform just-in-time assistance. However, we currently lack a large-scale benchmark that captures user–assistant interactions with all of the aforementioned features. To this end, we propose SIMMC-VR, an extension of the SIMMC-2.0 dataset, to a video-grounded task-oriented dialog dataset that captures real-world AI-assisted user scenarios in VR. We propose a novel data collection paradigm that involves (1) generating object-centric multimodal dialog flows with egocentric visual streams and visually-grounded templates, and (2) manually paraphrasing the simulated dialogs for naturalness and diversity while preserving multimodal dependencies. To measure meaningful progress in the field, we propose four tasks to address the new challenges in SIMMC-VR, which require complex spatial-temporal dialog reasoning in active egocentric scenes. We benchmark the proposed tasks with strong multimodal models, and highlight the key capabilities that current models lack for future research directions.

Covering Uncommon Ground: Gap-Focused Question Generation for Answer Assessment

Roni Rabin, Alexandre Djerbetian, Roee Engelberg, Lidan Hackmon, Gal Elidan, Reut Tsarfaty and Amir Globerson 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Human communication often involves information gaps between the interlocutors. For example, in an educational dialogue a student often provides an answer that is incomplete, and there is a gap between this answer and the perfect one expected by the teacher. Successful dialogue then hinges on the teacher asking about this gap in an effective manner, thus creating a rich and interactive educational experience. We focus on the problem of generating such gap-focused questions (GFQs) automatically. We define the task, highlight key desired aspects of a good GFQ, and propose a model that satisfies these. Finally, we provide an evaluation by human annotators of our generated questions compared against human generated ones, demonstrating competitive performance.

Do I have the Knowledge to Answer? Investigating Answerability of Knowledge Base Questions

Mayur Patidar, Prayushi Faldu, Avinash Kumar Singh, Lovekesh Vig, Indrajit Bhattacharya and Mausam - 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

When answering natural language questions over knowledge bases, missing facts, incomplete schema and limited scope naturally lead to many questions being unanswerable. While answerability has been explored in other QA settings, it has not been studied for QA over knowledge

bases (KBQA). We create GrailQAbility, a new benchmark KBQA dataset with unanswerability, by first identifying various forms of KB incompleteness that make questions unanswerable, and then systematically adapting GrailQA (a popular KBQA dataset with only answerable questions). Experimenting with three state-of-the-art KBQA models, we find that all three models suffer a drop in performance even after suitable adaptation for unanswerable questions. In addition, these often detect unanswerability for wrong reasons and find specific forms of unanswerability particularly difficult to handle. This underscores the need for further research in making KBQA systems robust to unanswerability.

Query Structure Modeling for Inductive Logical Reasoning Over Knowledge Graphs

Siyuan Wang, Zhongyu Wei, meng han, Zhihao Fan, Haijun Shan, Qi Zhang and Xuanjing Huang 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Logical reasoning over incomplete knowledge graphs to answer complex logical queries is a challenging task. With the emergence of new entities and relations in constantly evolving KGs, inductive logical reasoning over KGs has become a crucial problem. However, previous PLMs-based methods struggle to model the logical structures of complex queries, which limits their ability to generalize within the same structure. In this paper, we propose a structure-modeled textual encoding framework for inductive logical reasoning over KGs. It encodes linearized query structures and entities using pre-trained language models to find answers. For structure modeling of complex queries, we design stepwise instructions that implicitly prompt PLMs on the execution order of geometric operations in each query. We further separately model different geometric operations (i.e., projection, intersection, and union) on the representation space using a pre-trained encoder with additional attention and maxout layers to enhance structured modeling. We conduct experiments on two inductive logical reasoning datasets and three transductive datasets. The results demonstrate the effectiveness of our method on logical reasoning over KGs in both inductive and transductive settings.

Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot and Ashish Sabharwal 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Prompting-based large language models (LLMs) are surprisingly powerful at generating natural language reasoning steps or Chains-of-Thoughts (CoT) for multi-step question answering (QA). They struggle, however, when the necessary knowledge is either unavailable to the LLM or not up-to-date within its parameters. While using the question to retrieve relevant text from an external knowledge source helps LLMs, we observe that this one-step retrieve-and-read approach is insufficient for multi-step QA. Here, *what to retrieve* depends on *what has already been derived*, which in turn may depend on *what was previously retrieved*. To address this, we propose IRCoT, a new approach for multi-step QA that interleaves retrieval with steps (sentences) in a CoT, guiding the retrieval with CoT and in turn using retrieved results to improve CoT. Using IRCoT with GPT3 substantially improves retrieval (up to 21 points) as well as downstream QA (up to 15 points) on four datasets: HotpotQA, 2WikiMultihopQA, MuSiQue, and IIRC. We observe similar substantial gains in out-of-distribution (OOD) settings as well as with much smaller models such as Flan-T5-large without additional training. IRCoT reduces model hallucination, resulting in factually more accurate CoT reasoning.

Single Sequence Prediction over Reasoning Graphs for Multi-hop QA

Gowtham Ramesh, Makesh Narsimhan Sreedhar and Junjie Hu 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Recent generative approaches for multi-hop question answering (QA) utilize the fusion-in-decoder method to generate a single sequence output which includes both a final answer and a reasoning path taken to arrive at that answer, such as passage titles and key facts from those passages. While such models can lead to better interpretability and high quantitative scores, they often have difficulty accurately identifying the passages corresponding to key entities in the context, resulting in incorrect passage hops and a lack of faithfulness in the reasoning path. To address this, we propose a single-sequence prediction method over a local reasoning graph that integrates a graph structure connecting key entities in each context passage to relevant subsequent passages for each question. We use a graph neural network to encode this graph structure and fuse the resulting representations into the entity representations of the model. Our experiments show significant improvements in answer exact-match/F1 scores and faithfulness of grounding in the reasoning path on the HotpotQA dataset and achieve state-of-the-art numbers on the Musique dataset with only up to a 4% increase in model parameters.

Few-shot Reranking for Multi-hop QA via Language Model Prompting

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee and Lu Wang 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

We study few-shot reranking for multi-hop QA (MQA) with open-domain questions. To alleviate the need for a large number of labeled question-document pairs for retriever training, we propose PromptRank, which relies on language model prompting for multi-hop path reranking. PromptRank first constructs an instruction-based prompt that includes a candidate document path and then computes the relevance score between a given question and the path based on the conditional likelihood of the question given the path prompt according to a language model. PromptRank yields strong retrieval performance on HotpotQA with only 128 training examples compared to state-of-the-art methods trained on thousands of examples — 73.6 recall@10 by PromptRank vs. 77.8 by PathRetriever and 77.5 by multi-hop dense retrieval.

Prefix Propagation: Parameter-Efficient Tuning for Long Sequences

Jonathan X Li, Will Aitken, Rohan Bhambhoria and Xiaodan Zhu 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Parameter-efficient tuning aims to mitigate the large memory requirements of adapting pretrained language models for downstream tasks. For example, one popular method, prefix-tuning, prepends trainable tokens to sequences while freezing the rest of the model's parameters. Although such models attain comparable performance with fine-tuning when applied to sequences with short to moderate lengths, we show their inferior performance when modelling long sequences. To bridge this gap, we propose prefix-propagation, a simple but effective approach that conditions prefixes on previous hidden states. We empirically demonstrate that prefix-propagation outperforms prefix-tuning across long-document tasks, while using 50% fewer parameters. To further investigate the proposed architecture, we also show its advantage in calibration, and perform additional study on its relationship with kernel attention. To the best of our knowledge, this work is the first to focus on parameter-efficient learning for long-sequence language tasks.

Multi-CLS BERT: An Efficient Alternative to Traditional Ensembling

Haw-Shiuan Chang, Ruei-Yao Sun, Kathryn Ricci and Andrew McCallum 09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Ensembling BERT models often significantly improves accuracy, but at the cost of significantly more computation and memory footprint. In this work, we propose Multi-CLS BERT, a novel ensembling method for CLS-based prediction tasks that is almost as efficient as a single BERT model. Multi-CLS BERT uses multiple CLS tokens with a parameterization and objective that encourages their diversity. Thus instead of fine-tuning each BERT model in an ensemble (and running them all at test time), we need only fine-tune our single Multi-CLS BERT model (and run the one model at test time, ensembling just the multiple final CLS embeddings). To test its effectiveness, we build Multi-CLS BERT on top of a state-of-the-art pretraining method for BERT (Aroca-Ouellette and Rudzicz, 2020). In experiments on GLUE and SuperGLUE we show that our Multi-CLS BERT reliably improves both overall accuracy and confidence estimation. When only 100 training samples are available in GLUE, the Multi-CLS BERT_Base model can even outperform the corresponding BERT_Large model. We analyze the behavior of our Multi-CLS BERT, showing that it has many of the same characteristics and behavior as a typical BERT 5-way ensemble, but with nearly 4-times less computation and memory.

PESCO: Prompt-enhanced Self Contrastive Learning for Zero-shot Text Classification
Yau-Shian Wang, Ta-Chung Chi, Ruohong Zhang and YIMING YANG   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We present PESCO, a novel contrastive learning framework that substantially improves the performance of zero-shot text classification. We formulate text classification as a neural text retrieval problem where each document is treated as a query, and the system learns the mapping from each query to the relevant class labels by (1) adding prompts to enhance label retrieval, and (2) using retrieved labels to enrich the training set in a self-training loop of contrastive learning. PESCO achieves state-of-the-art performance on four benchmark text classification datasets. On DBpedia, we achieve 98.5% accuracy without any labeled data, which is close to the fully-supervised result. Extensive experiments and analyses show all the components of PESCO are necessary for improving the performance of zero-shot text classification.

Parameter-Efficient Fine-Tuning without Introducing New Latency
Baohao Liao, Yan Meng and Christof Monz   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Parameter-efficient fine-tuning (PEFT) of pre-trained language models has recently demonstrated remarkable achievements, effectively matching the performance of full fine-tuning while utilizing significantly fewer trainable parameters, and consequently addressing the storage and communication constraints. Nonetheless, various PEFT methods are limited by their inherent characteristics. In the case of sparse fine-tuning, which involves modifying only a small subset of the existing parameters, the selection of fine-tuned parameters is task- and domain-specific, making it unsuitable for federated learning. On the other hand, PEFT methods with adding new parameters typically introduce additional inference latency. In this paper, we demonstrate the feasibility of generating a sparse mask in a task-agnostic manner, wherein all downstream tasks share a common mask. Our approach, which relies solely on the magnitude information of pre-trained parameters, surpasses existing methodologies by a significant margin when evaluated on the GLUE benchmark. Additionally, we introduce a novel adapter technique that directly applies the adapter to pre-trained parameters instead of the hidden representation, thereby achieving identical inference speed to that of full fine-tuning. Through extensive experiments, our proposed method attains a new state-of-the-art outcome in terms of both performance and storage efficiency, storing only 0.03% parameters of full fine-tuning.

Counterfactual Active Learning for Out-of-Distribution Generalization
Xun Deng, Wenjie Wang, Fuli Feng, Hanwang Zhang, Xiangnan He and Yong Liao   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We study the out-of-distribution generalization of active learning that adaptively selects samples for annotation in learning the decision boundary of classification. Our empirical study finds that increasingly annotating seen samples may hardly benefit the generalization. To address the problem, we propose Counterfactual Active Learning (CounterAL) that empowers active learning with counterfactual thinking to bridge the seen samples with unseen cases. In addition to annotating factual samples, CounterAL requires annotators to answer counterfactual questions to construct counterfactual samples. To achieve CounterAL, we design a new acquisition strategy that selects the informative factual-counterfactual pairs for annotation; and a new training strategy that pushes the model update to focus on the discrepancy between factual and counterfactual samples. We evaluate CounterAL on multiple public datasets of sentiment analysis and natural language inference. The experiment results show that CounterAL requires fewer acquisition rounds and outperforms existing active learning methods by a large margin in OOD tests with comparable IID performance.

Cold-Start Data Selection for Better Few-shot Language Model Fine-tuning: A Prompt-based Uncertainty Propagation Approach
Yue Yu, Rongzhi Zhang, Ran Xu, Jieyu Zhang, Jiaming Shen and Chao Zhang   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We present PATRON, a prompt-based data selection method for pre-trained language model fine-tuning under cold-start scenarios, i.e., no initial labeled data are available. In PATRON, we design (1) a prompt-based uncertainty propagation approach to estimate the importance of data points and (2) a partition-then-rewrite (PTR) strategy to promote sample diversity when querying for annotations. Experiments on six text classification datasets show that PATRON outperforms the strongest cold-start data selection baselines by up to 6.9%. Besides, with 128 labels only, PATRON achieves 91.0% and 92.1% of the fully supervised performance based on vanilla fine-tuning and prompt-based learning respectively. Our implementation of PATRON will be published upon acceptance.

PeerDA: Data Augmentation via Modeling Peer Relation for Span Identification Tasks
Weiwen Xu, Xin Li, Yang Deng, Wai Lam and Lidong Bing   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Span identification aims at identifying specific text spans from text input and classifying them into pre-defined categories. Different from previous works that merely leverage the Subordinate (SUB) relation (i.e. if a span is an instance of a certain category) to train models, this paper for the first time explores the Peer (PR) relation, which indicates that two spans are instances of the same category and share similar features. Specifically, a novel Peer Data Augmentation (PeerDA) approach is proposed which employs span pairs with the PR relation as the augmentation data for training. PeerDA has two unique advantages: (1) There are a large number of PR span pairs for augmenting the training data. (2) The augmented data can prevent the trained model from over-fitting the superficial span-category mapping by pushing the model to leverage the span semantics. Experimental results on ten datasets over four diverse tasks across seven domains demonstrate the effectiveness of PeerDA. Notably, PeerDA achieves state-of-the-art results on six of them.

Prompting Language Models for Linguistic Structure
Terra Blevins, Hila Gonen and Luke Zettlemoyer   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Although pretrained language models (PLMs) can be prompted to perform a wide range of language tasks, it remains an open question how much this ability comes from generalizable linguistic understanding versus surface-level lexical patterns. To test this, we present a structured prompting approach for linguistic structured prediction tasks, allowing us to perform zero- and few-shot sequence tagging with autoregressive PLMs. We evaluate this approach on part-of-speech tagging, named entity recognition, and sentence chunking, demonstrating strong few-shot performance in all cases. We also find that while PLMs contain significant prior knowledge of task labels due to task leakage into the pretraining corpus, structured prompting can also retrieve linguistic structure with arbitrary labels. These findings indicate that the in-context learning ability and linguistic knowledge of PLMs generalizes beyond memorization of their training data.

Large Language Models Are Reasoning Teachers
Namgyu Ho, Laura Schmid and Se-Young Yun   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Recent works have shown that chain-of-thought (CoT) prompting can elicit language models to solve complex reasoning tasks, step-by-step. However, prompt-based CoT methods are dependent on very large language models such as GPT-3 175B which are prohibitive to deploy at scale. In this paper, we use these large models as reasoning teachers to enable complex reasoning in smaller models and reduce model size requirements by several orders of magnitude. We propose Fine-tune-CoT, a method that generates reasoning samples from very large teacher models to fine-tune smaller models. We evaluate our method on a wide range of public models and complex tasks. We find that Fine-tune-CoT enables substantial reasoning capability in small models, far outperforming prompt-based baselines and even the teacher model in many tasks. Additionally, we extend our method by leveraging the teacher model's ability to generate multiple distinct rationales for each original sample. Enriching the fine-tuning data with such diverse reasoning results in a substantial performance boost across datasets, even for very small models. We conduct ablations and sample studies to understand the emergence of reasoning capabilities of student models. Our code implementation and data are available at https://github.com/itsnamgyu/reasoning-teacher.

Plug-and-Play Document Modules for Pre-trained Models

Chaojun Xiao, Zhengyan Zhang, Xu Han, Chi-Min Chan, Yankai Lin, Zhiyuan Liu, xiangyang li, Zhonghua Li, Zhao Cao and Maosong Sun
09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Large-scale pre-trained models (PTMs) have been widely used in document-oriented NLP tasks, such as question answering. However, the encoding-task coupling requirement results in the repeated encoding of the same documents for different tasks and queries, which is highly computationally inefficient. To this end, we target to decouple document encoding from downstream tasks, and propose to represent each document as a plug-and-play knowledge module, i.e., a document plugin, for PTMs (PlugD). By inserting document plugins into the backbone PTM for downstream tasks, we can encode a document one time to handle multiple tasks, which is more efficient than conventional encoding-task coupling methods that simultaneously encode documents and input queries using task-specific encoders. Extensive experiments on 8 datasets of 4 typical NLP tasks show that PlugD enables models to encode documents once and for all across different scenarios. Especially, PlugD can save 69% computational costs while achieving comparable performance to state-of-the-art encoding-task coupling methods. Additionally, we show that PlugD can serve as an effective post-processing way to inject knowledge into task-specific models, improving model performance without any additional model training. Our code and checkpoints can be found in https://github.com/thunlp/Document-Plugin.

Enhancing Dialogue Generation via Dynamic Graph Knowledge Aggregation
Chen Tang, Hongbo Zhang, Tyler Loakman, Chenghua Lin and Frank Guerin 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Incorporating external graph knowledge into neural chatbot models has been proven effective for enhancing dialogue generation. However, in conventional graph neural networks (GNNs), message passing on a graph is independent from text, resulting in the graph representation hidden space differing from that of the text. This training regime of existing models therefore leads to a semantic gap between graph knowledge and text. In this study, we propose a novel framework for knowledge graph enhanced dialogue generation. We dynamically construct a multi-hop knowledge graph with pseudo nodes to involve the language model in feature aggregation within the graph at all steps. To avoid the semantic biases caused by learning on vanilla subgraphs, the proposed framework applies hierarchical graph attention to aggregate graph features on pseudo nodes and then attains a global feature. Therefore, the framework can better utilise the heterogeneous features from both the post and external graph knowledge. Extensive experiments demonstrate that our framework outperforms state-of-the-art (SOTA) baselines on dialogue generation. Further analysis also shows that our representation learning framework can fill the semantic gap by coagulating representations of both text and graph knowledge. Moreover, the language model also learns how to better select knowledge triples for a more informative response via exploiting subgraph patterns within our feature aggregation process. Our code and resources are available at https://github.com/tangg555/SaBART.

Attractive Storyteller: Stylized Visual Storytelling with Unpaired Text
Dingyi Yang and Qin Jin 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Most research on stylized image captioning aims to generate style-specific captions using unpaired text, and has achieved impressive performance for simple styles like positive and negative. However, unlike previous single-sentence captions whose style is mostly embodied in distinctive words or phrases, real-world styles are likely to be implied at the syntactic and discourse levels. In this work, we introduce a new task of Stylized Visual Storytelling (SVST), which aims to describe a photo stream with stylized stories that are more expressive and attractive. We propose a multitasking memory-augmented framework called StyleVSG, which is jointly trained on factual visual storytelling data and unpaired style corpus, achieving a trade-off between style accuracy and visual relevance. Particularly for unpaired stylized text, StyleVSG learns to reconstruct the stylistic story from roughly parallel visual inputs mined with the CLIP model, avoiding problems caused by random mapping in previous methods. Furthermore, a memory module is designed to preserve the consistency and coherence of generated stories. Experiments show that our method can generate attractive and coherent stories with different styles such as fairy tale, romance, and humor. The overall performance of our StyleVSG surpasses state-of-the-art methods on both automatic and human evaluation metrics.

An Extensible Plug-and-Play Method for Multi-Aspect Controllable Text Generation
Xuancheng Huang, Zijun Liu, Peng Li, Tao Li, Maosong Sun and Yang Liu 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Recently, multi-aspect controllable text generation that controls the generated text in multiple aspects (e.g., sentiment, topic, and keywords) has attracted increasing attention. Although methods based on parameter efficient tuning like prefix-tuning could achieve multi-aspect controlling in a plug-and-play way, the mutual interference of multiple prefixes leads to significant degeneration of constraints and limits their extensibility to training-time unseen aspect combinations. In this work, we provide a theoretical lower bound for the interference and empirically found that the interference grows with the number of layers where prefixes are inserted. Based on these analyses, we propose using trainable gates to normalize the intervention of prefixes to restrain the growing interference. As a result, controlling training-time unseen combinations of aspects can be realized by simply concatenating corresponding plugins such that new constraints can be extended at a lower cost. In addition, we propose a unified way to process both categorical and free-form constraints. Experiments on text generation and machine translation demonstrate the superiority of our approach over baselines on constraint accuracy, text quality, and extensibility.

With a Little Push, NLI Models can Robustly and Efficiently Predict Faithfulness
Julius Steen, Juri Opitz, Anette Frank and Katja Markert 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Conditional language models still generate unfaithful output that is not supported by their input. These unfaithful generations jeopardize trust in real-world applications such as summarization or human-machine interaction, motivating a need for automatic faithfulness metrics. To implement such metrics, NLI models seem attractive, since they solve a strongly related task that comes with a wealth of prior research and data. But recent research suggests that NLI models require costly additional machinery to perform reliably across datasets, e.g., by running inference on a cartesian product of input and generated sentences, or supporting them with a question-generation/answering step.

In this work we show that pure NLI models _can_ outperform more complex metrics when combining task-adaptive data augmentation with robust inference procedures. We propose: (1) Augmenting NLI training data to adapt NL inferences to the specificities of faithfulness prediction in dialogue; (2) Making use of both entailment and contradiction probabilities in NLI, and (3) Using Monte-Carlo dropout during inference. Applied to the TRUE benchmark, which combines faithfulness datasets across diverse domains and tasks, our approach strongly improves a vanilla NLI model and significantly outperforms previous work, while showing favourable computational cost.

LENS: A Learnable Evaluation Metric for Text Simplification
Mounica Maddela, Yao Dou, David Heineman and Wei Xu 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Training learnable metrics using modern language models has recently emerged as a promising method for the automatic evaluation of machine translation. However, existing human evaluation datasets for text simplification have limited annotations that are based on unitary or outdated models, making them unsuitable for this approach. To address these issues, we introduce the SimpEval corpus that contains: SimpEval_past, comprising 12K human ratings on 2.4K simplifications of 24 past systems, and SimpEval_2022, a challenging simplification benchmark consisting of over 1K human ratings of 360 simplifications including GPT-3.5 generated text. Training on SimpEval, we present LENS, a Learnable Evaluation Metric for Text Simplification. Extensive empirical results show that LENS correlates much better with human judgment than existing metrics, paving the way for future progress in the evaluation of text simplification. We also introduce Rank & Rate, a human evaluation framework that rates simplifications from several models in a list-wise manner using an interactive interface, which ensures both consistency and accuracy in the evaluation process and is used to create the SimpEval datasets.

BOLT: Fast Energy-based Controlled Text Generation with Tunable Biases

Xin Liu, Muhammad Khalifa and Lu Wang   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Energy-based models (EBMs) have gained popularity for controlled text generation due to their high applicability to a wide range of constraints. However, sampling from EBMs is non-trivial, as it often requires a large number of iterations to converge to plausible text, which slows down the decoding process and makes it less practical for real-world applications. In this work, we propose BOLT, which relies on tunable biases to directly adjust the language model's output logits. Unlike prior work, BOLT maintains the generator's autoregressive nature to assert a strong control on token-level conditional dependencies and overall fluency, and thus converges faster. When compared with state-of-the-arts on controlled generation tasks using both soft constraints (e.g., sentiment control) and hard constraints (e.g., keyword-guided topic control), BOLT demonstrates significantly improved efficiency and fluency. On sentiment control, BOLT is 7x faster than competitive baselines, and more fluent in 74.4% of the evaluation samples according to human judges.

DuNST: Dual Noisy Self Training for Semi-Supervised Controllable Text Generation
Yuxi Feng, Xiaoyuan Yi, Xiting Wang, Laks Lakshmanan, V.S. and Xing Xie   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Self-training (ST) has prospered again in language understanding by augmenting the fine-tuning of big pre-trained models when labeled data is insufficient. However, it remains challenging to incorporate ST into attribute-controllable language generation. Augmented only by self-generated pseudo text, generation models *over-exploit* the previously learned text space and *fail to explore* a larger one, suffering from a restricted generalization boundary and limited controllability. In this work, we propose DuNST, a novel ST framework to tackle these problems. DuNST jointly models text generation and classification as a dual process and further perturbs and escapes from the collapsed space by adding two kinds of flexible noise. In this way, our model could construct and utilize both pseudo text generated from given labels and pseudo labels predicted from available unlabeled text, which are gradually refined during the ST phase. We theoretically demonstrate that DuNST can be regarded as enhancing the exploration of the potentially larger real text space while maintaining exploitation, guaranteeing improved performance. Experiments on three controllable generation tasks show that DuNST significantly boosts control accuracy with comparable generation fluency and diversity against several strong baselines.

NEUROSTRUCTURAL DECODING: Neural Text Generation with Structural Constraints
Mohaddeseh Bastan, Mihai Surdeanu and Niranjan Balasubramanian   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Text generation often involves producing coherent and grammatically correct texts that also satisfy a given set of semantic constraints. While most approaches for conditional text generation have primarily focused on lexical constraints, they often struggle to effectively incorporate syntactic constraints, which provide a richer language for approximating semantic constraints. We address this gap by introducing NeuroStructural Decoding, a new decoding algorithm that incorporates syntactic constraints to further improve the quality of the generated text. We build NeuroStructural Decoding on the NeuroLogic Decoding (Lu etal. 2021) algorithm, which enables language generation models to produce fluent text while satisfying complex lexical constraints. Our algorithm is powerful and scalable. It tracks lexico-syntactic constraints (e.g., we need to observe dog as subject and ball as object)during decoding by parsing the partial generations at each step. To this end, we adapt a dependency parser to generate parses for incomplete sentences. Our approach is evaluated on three different language generation tasks, and the results show improved performance in both lexical and syntactic metrics compared to previous methods. The results suggest this is a promising solution for integrating fine-grained controllable generation into the conventional beam search decoding.

Attention as a Guide for Simultaneous Speech Translation
Sara Papi, Matteo Negri and Marco Turchi   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
In simultaneous speech translation (SimulST), effective policies that determine when to write partial translations are crucial to reach high output quality with low latency. Towards this objective, we propose EDAtt (Encoder-Decoder Attention), an adaptive policy that exploits the attention patterns between audio source and target textual translation to guide an offline-trained ST model during simultaneous inference. EDAtt exploits the attention scores modeling the audio-translation relation to decide whether to emit a partial hypothesis or wait for more audio input. This is done under the assumption that, if attention is focused towards the most recently received speech segments, the information they provide can be insufficient to generate the hypothesis (indicating that the system has to wait for additional audio input). Results on en->de, es show that EDAtt yields better results compared to the SimulST state of the art, with gains respectively up to 7 and 4 BLEU points for the two languages, and with a reduction in computational-aware latency up to 1.4s and 0.7s compared to existing SimulST policies applied to offline-trained models.

Learning Language-Specific Layers for Multilingual Machine Translation
Telmo Pires, Robin M. Schmidt, Yi-Hsiu Liao and Stephan Peitz   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Multilingual Machine Translation promises to improve translation quality between non-English languages. This is advantageous for several reasons, namely lower latency (no need to translate twice), and reduced error cascades (e.g., avoiding losing gender and formality information when translating through English). On the downside, adding more languages reduces model capacity per language, which is usually countered by increasing the overall model size, making training harder and inference slower. In this work, we introduce Language-Specific Transformer Layers (LSLs), which allow us to increase model capacity, while keeping the amount of computation and the number of parameters used in the forward pass constant. The key idea is to have some layers of the encoder be source or target language-specific, while keeping the remaining layers shared. We study the best way to place these layers using a neural architecture search inspired approach, and achieve an improvement of 1.3 chrF (1.5 spBLEU) points over not using LSLs on a separate decoder architecture, and 1.9 chrF (2.2 spBLEU) on a shared decoder one.

Joint End-to-end Semantic Proto-role Labeling
Elizabeth Spaulding, Gary Kazantsev and Mark Dredze   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Semantic proto-role labeling (SPRL) assigns properties to arguments based on a series of binary labels. While multiple studies have evaluated various approaches to SPRL, it has only been studied in-depth as a standalone task using gold predicate/argument pairs. How do SPRL systems perform as part of an information extraction pipeline? We model SPRL jointly with predicate-argument extraction using a deep transformer model. We find that proto-role labeling is surprisingly robust in this setting, with only a small decrease when using predicted arguments. We include a detailed analysis of each component of the joint system, and an error analysis to understand correlations in errors between system stages. Finally, we study the effects of annotation errors on SPRL.

Jointprop: Joint Semi-supervised Learning for Entity and Relation Extraction with Heterogeneous Graph-based Propagation
yandan zheng, Anran Hao and Anh Tuan Luu   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Semi-supervised learning has been an important approach to address challenges in extracting entities and relations from limited data. However, current semi-supervised works handle the two tasks (i.e., Named Entity Recognition and Relation Extraction) separately and ignore the cross-correlation of entity and relation instances as well as the existence of similar instances across unlabeled data. To alleviate the issues, we propose Jointprop, a Heterogeneous Graph-based Propagation framework for joint semi-supervised entity and relation extraction, which captures the global structure information between individual tasks and exploits interactions within unlabeled data. Specifically, we construct a unified span-based heterogeneous graph from entity and relation candidates and propagate class labels based on confidence scores. We then employ a propagation learning scheme to leverage the affinities between labelled and unlabeled samples. Experiments on benchmark datasets show that our framework outperforms the state-of-the-art semi-supervised approaches on NER and RE tasks. We show that the joint semi-supervised learning of the two tasks benefits from their codependency and validates the importance of utilizing the shared information

between unlabeled data.

Learning "O" Helps for Learning More: Handling the Unlabeled Entity Problem for Class-incremental NER
Ruotian Ma, xuanting chen, zhang lin, Xin Zhou, Junzhe Wang, Tao Gui, Qi Zhang, Xiang Gao and Yun Wen Chen 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
As the categories of named entities rapidly increase, the deployed NER models are required to keep updating toward recognizing more entity types, creating a demand for class-incremental learning for NER. Considering the privacy concerns and storage constraints, the standard paradigm for class-incremental NER updates the models with training data only annotated with the new classes, yet the entities from other entity classes are regarded as "Non-entity" (or "O"). In this work, we conduct an empirical study on the "Unlabeled Entity Problem" and find that it leads to severe confusion between "O" and entities, decreasing class discrimination of old classes and declining the model's ability to learn new classes. To solve the Unlabeled Entity Problem, we propose a novel representation learning method to learn discriminative representations for the entity classes and "O". Specifically, we propose an entity-aware contrastive learning method that adaptively detects entity clusters in "O". Furthermore, we propose two effective distance-based relabeling strategies for better learning the old classes. We introduce a more realistic and challenging benchmark for class-incremental NER, and the proposed method achieves up to 10.62% improvement over the baseline methods.

WebIE: Faithful and Robust Information Extraction on the Web
Chenxi Whitehouse, Clara Vania, Alham Fikri Aji, Christos Christodoulopoulos and Andrea Pierleoni 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Extracting structured and grounded fact triples from raw text is a fundamental task in Information Extraction (IE). Existing IE datasets are typically collected from Wikipedia articles, using hyperlinks to link entities to the Wikidata knowledge base. However, models trained only on Wikipedia have limitations when applied to web domains, which often contain noisy text or text that does not have any factual information. We present WebIE, the first large-scale, entity-linked closed IE dataset consisting of 1.6M sentences automatically collected from the English Common Crawl corpus. WebIE also includes negative examples, i.e. sentences without fact triples, to better reflect the data on the web. We annotate 25K triples from WebIE through crowdsourcing and introduce mWebIE, a translation of the annotated set in four other languages: French, Spanish, Portuguese, and Hindi. We evaluate the in-domain, out-of-domain, and zero-shot cross-lingual performance of generative IE models and find models trained on WebIE show better generalisability. We also propose three training strategies that use entity linking as an auxiliary task. Our experiments show that adding Entity-Linking objectives improves the faithfulness of our generative IE models.

FormNetV2: Multimodal Graph Contrastive Learning for Form Document Information Extraction
Chen-Yu Lee, Chun-Liang Li, Hao Zhang, Timothy Dozat, Vincent Perot, Guolong Su, Xiang Zhang, Kihyuk Sohn, NIKOLAY GLUSHNEV, Renshen Wang, Joshua Ainslie, Shangbang Long, Siyang Qin, Yasuhisa Fujii, Nan Hua and Tomas Pfister 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
The recent advent of self-supervised pre-training techniques has led to a surge in the use of multimodal learning in form document understanding. However, existing approaches that extend the mask language modeling to other modalities require careful multi-task tuning, complex reconstruction target designs, or additional pre-training data. In FormNetV2, we introduce a centralized multimodal graph contrastive learning strategy to unify self-supervised pre-training for all modalities in one loss. The graph contrastive objective maximizes the agreement of multimodal representations, providing a natural interplay for all modalities without special customization. In addition, we extract image features within the bounding box that joins a pair of tokens connected by a graph edge, capturing more targeted visual cues without loading a sophisticated and separately pre-trained image embedder. FormNetV2 establishes new state-of-the-art performance on FUNSD, CORD, SROIE and Payment benchmarks with a more compact model size.

Automatic Creation of Named Entity Recognition Datasets by Querying Phrase Representations
Hyunjae Kim, jaehyo yoo, Seunghyun Yoon and Jaewoo Kang 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Most weakly supervised named entity recognition (NER) models rely on domain-specific dictionaries provided by experts. This approach is infeasible in many domains where dictionaries do not exist. While a phrase retrieval model was used to construct pseudo-dictionaries with entities retrieved from Wikipedia automatically in a recent study, these dictionaries often have limited coverage because the retriever is likely to retrieve popular entities rather than rare ones. In this study, we present a novel framework, HighGEN, that generates NER datasets with high-coverage pseudo-dictionaries. Specifically, we create entity-rich dictionaries with a novel search method, called phrase embedding search, which encourages the retriever to search a space densely populated with various entities. In addition, we use a new verification process based on the embedding distance between candidate entity mentions and entity types to reduce the false-positive noise in weak labels generated by high-coverage dictionaries. We demonstrate that HighGEN outperforms the previous best model by an average F1 score of 4.7 across five NER benchmark datasets.

S2ynRE: Two-stage Self-training with Synthetic data for Low-resource Relation Extraction
Benfeng Xu, Quan Wang, Yajuan Lyu, Dai Dai, Yongdong Zhang and Zhendong Mao 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Current relation extraction methods suffer from the inadequacy of large-scale annotated data. While distant supervision alleviates the problem of data quantities, there still exists domain disparity in data qualities due to its reliance on domain-restrained knowledge bases. In this work, we propose S2ynRE, a framework of two-stage Self-training with Synthetic data for Relation Extraction. We first leverage the capability of large language models to adapt to the target domain and automatically synthesize large quantities of coherent, realistic training data. We then propose an accompanied two-stage self-training algorithm that iteratively and alternately learns from synthetic and golden data together. We conduct comprehensive experiments and detailed ablations on popular relation extraction datasets to demonstrate the effectiveness of the proposed framework.

Character-Aware Models Improve Visual Text Rendering
Rosanne Liu, Dan Garrette, Chitwan Saharia, William Chan, Adam Roberts, Sharan Narang, Irina Blok, RJ Mical, Mohammad Norouzi and Noah Constant 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Current image generation models struggle to reliably produce well-formed visual text. In this paper, we investigate a key contributing factor: popular text-to-image models lack character-level input features, making it much harder to predict a word's visual makeup as a series of glyphs. To quantify this effect, we conduct a series of experiments comparing character-aware vs. character-blind text encoders. In the text-only domain, we find that character-aware models provide large gains on a novel spelling task (WikiSpell). Applying our learnings to the visual domain, we train a suite of image generation models, and show that character-aware variants outperform their character-blind counterparts across a range of novel text rendering tasks (our DrawText benchmark). Our models set a much higher state-of-the-art on visual spelling, with 30+ point accuracy gains over competitors on rare words, despite training on far fewer examples.

Generating Visual Spatial Description via Holistic 3D Scene Understanding
Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang and Tat-Seng Chua 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Visual spatial description (VSD) aims to generate texts that describe the spatial relations of the given objects within images. Existing VSD work merely models the 2D geometrical vision features, thus inevitably falling prey to the problem of skewed spatial understanding of target

objects. In this work, we investigate the incorporation of 3D scene features for VSD. With an external 3D scene extractor, we obtain the 3D objects and scene features for input images, based on which we construct a target object-centered 3D spatial scene graph (Go3D-S2G), such that we model the spatial semantics of target objects within the holistic 3D scenes. Besides, we propose a scene subgraph selecting mechanism, sampling topologically-diverse subgraphs from Go3D-S2G, where the diverse local structure features are navigated to yield spatially-diversified text generation. Experimental results on two VSD datasets demonstrate that our framework outperforms the baselines significantly, especially improving on the cases with complex visual spatial relations. Meanwhile, our method can produce more spatially-diversified generation.

BLASER: A Text-Free Speech-to-Speech Translation Evaluation Metric
Mingda Chen, Paul-Ambroise Augustin Duquenne, Pierre Y Andrews, Justine T Kao, Alexandre Mourachko, Holger Schwenk and Marta R. Costa-jussà 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
End-to-End speech-to-speech translation (S2ST) is generally evaluated with text-based metrics. This means that generated speech has to be automatically transcribed, making the evaluation dependent on the availability and quality of automatic speech recognition (ASR) systems. In this paper, we propose a text-free evaluation metric for end-to-end S2ST, named BLASER, to avoid the dependency on ASR systems. BLASER leverages a multilingual multimodal encoder to directly encode the speech segments for source input, translation output and reference into a shared embedding space and computes a score of the translation quality that can be used as a proxy to human evaluation. To evaluate our approach, we construct training and evaluation sets from more than 40k human annotations covering seven language directions. The best results of BLASER are achieved by training with supervision from human rating scores. We show that when evaluated at the sentence level, BLASER correlates significantly better with human judgment compared to ASR dependent metrics including ASR-SENTBLEU in all translation directions and ASR-COMET in five of them. Our analysis shows combining speech and text as inputs to BLASER does not increase the correlation with human scores, but best correlations are achieved when using speech, which motivates the goal of our research. Moreover, we show that using ASR for references is detrimental for text-based metrics.

Dynamic Regularization in UDA for Transformers in Multimodal Classification
Ivonne Monter-Aldana, Adrian Pastor Lopez Monroy and Fernando Sanchez-Vega 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Multimodal machine learning is a cutting-edge field that explores ways to incorporate information from multiple sources into models. As more multimodal data becomes available, this field has become increasingly relevant. This work focuses on two key challenges in multimodal machine learning. The first is finding efficient ways to combine information from different data types. The second is that often, one modality (e.g., text) is stronger and more relevant, making it difficult to identify meaningful patterns in the weaker modality (e.g., image). Our approach focuses on more effectively exploiting the weaker modality while dynamically regularizing the loss function. First, we introduce a new two-stream model called Multimodal BERT-ViT, which features a novel intra-CLS token fusion. Second, we utilize a dynamic adjustment that maintains a balance between specialization and generalization during the training to avoid overfitting, which we devised. We add this dynamic adjustment to the Unsupervised Data Augmentation (UDA) framework. We evaluate the effectiveness of these proposals on the task of multi-label movie genre classification using the Moviescope and MM-IMDb datasets. The evaluation revealed that our proposal offers substantial benefits, while simultaneously enabling us to harness the weaker modality without compromising the information provided by the stronger.

Layer-wise Fusion with Modality Independence Modeling for Multi-modal Emotion Recognition
Jun Sun, Shoukang Han, Yu-Ping Ruan, Xiaoning Zhang, Shu-Kai Zheng, Yulong Liu, Yuxin Huang and Taihao Li 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Multi-modal emotion recognition has gained increasing attention in recent years due to its widespread applications and the advances in multi-modal learning approaches. However, previous studies primarily focus on developing models that exploit the unification of multiple modalities. In this paper, we propose that maintaining modality independence is beneficial for the model performance. According to this principle, we construct a dataset, and devise a multi-modal transformer model. The new dataset, CHinese Emotion Recognition dataset with Modality-wise Annotations, abbreviated as CHERMA, provides uni-modal labels for each individual modality, and multi-modal labels for all modalities jointly observed. The model consists of uni-modal transformer modules that learn representations for each modality, and a multi-modal transformer module that fuses all modalities. All the modules are supervised by their corresponding labels separately, and the forward information flow is uni-directionally from the uni-modal modules to the multi-modal module. The supervision strategy and the model architecture guarantee each individual modality learns its representation independently, and meanwhile the multi-modal module aggregates all information. Extensive empirical results demonstrate that our proposed scheme outperforms state-of-the-art alternatives, corroborating the importance of modality independence in multi-modal emotion recognition. The dataset and codes are availabel at https://github.com/sunjunaimer/LFMIM

A Cautious Generalization Goes a Long Way: Learning Morphophonological Rules
Salam Khalifa, Sarah Payne, Jordan Kodner, Ellen Broselow and Owen Rambow 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Explicit linguistic knowledge, encoded by resources such as rule-based morphological analyzers, continues to prove useful in downstream NLP tasks, especially for low-resource languages and dialects. Rules are an important asset in descriptive linguistic grammars. However, creating such resources is usually expensive and non-trivial, especially for spoken varieties with no written standard. In this work, we present a novel approach for automatically learning morphophonological rules of Arabic from a corpus. Motivated by classic cognitive models for rule learning, rules are generalized cautiously. Rules that are memorized for individual items are only allowed to generalize to unseen forms if they are sufficiently reliable in the training data. The learned rules are further examined to ensure that they capture true linguistic phenomena described by domain experts. We also investigate the learnability of rules in low-resource settings across different experimental setups and dialects.

Decomposed scoring of CCG dependencies
Aditya Bhargava and Gerald Penn 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
In statistical parsing with CCG, the standard evaluation method is based on predicate-argument structure and evaluates dependencies labelled in part by lexical categories. When a predicate has multiple argument slots that can be filled, the same lexical category is used for the label of multiple dependencies. In this paper, we show that this evaluation can result in disproportionate penalization of supertagging errors and obfuscate the truly erroneous dependencies. Enabled by the compositional nature of CCG lexical categories, we propose *decomposed scoring* based on subcategorial labels to address this.
To evaluate our scoring method, we engage fellow categorial grammar researchers in two English-language judgement tasks: (1) directly ranking the outputs of the standard and experimental scoring methods; and (2) determining which of two sentences has the better parse in cases where the two scoring methods disagree on their ranks. Overall, the judges prefer decomposed scoring in each task; but there is substantial disagreement among the judges in 24% of the given cases, pointing to potential issues with parser evaluations in general.

Substitution-based Semantic Change Detection using Contextual Embeddings
Dallas Card 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Measuring semantic change has thus far remained a task where methods using contextual embeddings have struggled to improve upon simpler techniques relying only on static word vectors. Moreover, many of the previously proposed approaches suffer from downsides related to

scalability and ease of interpretation. We present a simplified approach to measuring semantic change using contextual embeddings, relying only on the most probable substitutes for masked terms. Not only is this approach directly interpretable, it is also far more efficient in terms of storage, achieves superior average performance across the most frequently cited datasets for this task, and allows for more nuanced investigation of change than is possible with static word vectors.

ParaAMR: A Large-Scale Syntactically Diverse Paraphrase Dataset by AMR Back-Translation
Kuan-Hao Huang, Varun Iyer, I-Hung Hsu, Anoop Kumar, Kai-Wei Chang and Aram Galstyan   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Paraphrase generation is a long-standing task in natural language processing (NLP). Supervised paraphrase generation models, which rely on human-annotated paraphrase pairs, are cost-inefficient and hard to scale up. On the other hand, automatically annotated paraphrase pairs (e.g., by machine back-translation), usually suffer from the lack of syntactic diversity – the generated paraphrase sentences are very similar to the source sentences in terms of syntax. In this work, we present ParaAMR, a large-scale syntactically diverse paraphrase dataset created by abstract meaning representation back-translation. Our quantitative analysis, qualitative examples, and human evaluation demonstrate that the paraphrases of ParaAMR are syntactically more diverse compared to existing large-scale paraphrase datasets while preserving good semantic similarity. In addition, we show that ParaAMR can be used to improve on three NLP tasks: learning sentence embeddings, syntactically controlled paraphrase generation, and data augmentation for few-shot learning. Our results thus showcase the potential of ParaAMR for improving various NLP applications.

Infusing Hierarchical Guidance into Prompt Tuning: A Parameter-Efficient Framework for Multi-level Implicit Discourse Relation Recognition
Haodong Zhao, Ruifang He, Mengnan Xiao and Jing Xu   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Multi-level implicit discourse relation recognition (MIDRR) aims at identifying hierarchical discourse relations among arguments. Previous methods achieve the promotion through fine-tuning PLMs. However, due to the data scarcity and the task gap, the pre-trained feature space cannot be accurately tuned to the task-specific space, which even aggravates the collapse of the vanilla space. Besides, the comprehension of hierarchical semantics for MIDRR makes the conversion much harder. In this paper, we propose a prompt-based Parameter-Efficient Multi-level IDRR (PEMI) framework to solve the above problems. First, we leverage parameter-efficient prompt tuning to drive the inputted arguments to match the pre-trained space and realize the approximation with few parameters. Furthermore, we propose a hierarchical label refining (HLR) method for the prompt verbalizer to deeply integrate hierarchical guidance into the prompt tuning. Finally, our model achieves comparable results on PDTB 2.0 and 3.0 using about 0.1% trainable parameters compared with baselines and the visualization demonstrates the effectiveness of our HLR method.

Ellipsis-Dependent Reasoning: a New Challenge for Large Language Models
Daniel Hardt   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We propose a novel challenge for large language models: ellipsis-dependent reasoning. We define several structures of paired examples, where an ellipsis example is matched to its non-ellipsis counterpart, and a question is posed which requires resolution of the ellipsis. Test results show that the best models perform well on non-elliptical examples but struggle with all but the simplest ellipsis structures.

Where's the Point? Self-Supervised Multilingual Punctuation-Agnostic Sentence Segmentation
Benjamin Minixhofer, Jonas Pfeiffer and Ivan Vulić   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Many NLP pipelines split text into sentences as one of the crucial preprocessing steps. Prior sentence segmentation tools either rely on punctuation or require a considerable amount of sentence-segmented training data: both central assumptions might fail when porting sentence segmenters to diverse languages on a massive scale. In this work, we thus introduce a multilingual punctuation-agnostic sentence segmentation method, currently covering 85 languages, trained in a self-supervised fashion on unsegmented text, by making use of newline characters which implicitly perform segmentation into paragraphs. We further propose an approach that adapts our method to the segmentation in a given corpus by using only a small number (64-256) of sentence-segmented examples. The main results indicate that our method outperforms all the prior best sentence-segmentation tools by an average of 6.1% F1 points. Furthermore, we demonstrate that proper sentence segmentation has a point: the use of a (powerful) sentence segmenter makes a considerable difference for a downstream application such as machine translation (MT). By using our method to match sentence segmentation to the segmentation used during training of MT models, we achieve an average improvement of 2.3 BLEU points over the best prior segmentation tool, as well as massive gains over a trivial segmenter that splits text into equally-sized blocks.

Analyzing and Reducing the Performance Gap in Cross-Lingual Transfer with Fine-tuning Slow and Fast
Yiduo Guo, Yaobo Liang, Dongyan Zhao, Bing Liu and Nan Duan   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Existing research has shown that a multilingual pre-trained language model fine-tuned with one (source) language also performs well on downstream tasks for non-source languages, even though no fine-tuning is done on these languages. However, there is a clear gap between the performance of the source language and that of the non-source languages. This paper analyzes the fine-tuning process, discovers when the performance gap changes and identifies which network weights affect the overall performance most. Additionally, the paper seeks to answer to what extent the gap can be reduced by reducing forgetting. Based on the analysis results, a method named Fine-tuning slow and fast with four training policies is proposed to address these issues. Experimental results show the proposed method outperforms baselines by a clear margin.

Towards Robust Low-Resource Fine-Tuning with Multi-View Compressed Representations
Linlin Liu, Xingxuan Li, Megh Thakkar, Xin Li, Shafiq Joty, Luo Si and Lidong Bing   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Due to the huge amount of parameters, finetuning of pretrained language models (PLMs) is prone to overfitting in the low resource scenarios. In this work, we present a novel method that operates on the hidden representations of a PLM to reduce overfitting. During fine-tuning, our method inserts random autoencoders between the hidden layers of a PLM, which transform activations from the previous layers into multi-view compressed representations before feeding them into the upper layers. The autoencoders are plugged out after fine-tuning, so our method does not add extra parameters or increase computation cost during inference. Our method demonstrates promising performance improvement across a wide range of sequence- and token-level lowresource NLP tasks.

Randomized Smoothing with Masked Inference for Adversarially Robust Text Classifications
Han Cheol Moon, Shafiq Joty, Ruochen Zhao, Megh Thakkar and Chi Xu   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Large-scale pre-trained language models have shown outstanding performance in a variety of NLP tasks. However, they are also known to be significantly brittle against specifically crafted adversarial examples, leading to increasing interest in probing the adversarial robustness of NLP systems. We introduce RSMI, a novel two-stage framework that combines randomized smoothing (RS) with masked inference (MI) to improve the adversarial robustness of NLP systems. RS transforms a classifier into a smoothed classifier to obtain robust representations, whereas MI forces a model to exploit the surrounding context of a masked token in an input sequence. RSMI improves adversarial robustness by 2 to 3 times over existing state-of-the-art methods on benchmark datasets. We also perform in-depth qualitative analysis to validate the effectiveness of the different stages of RSMI and probe the impact of its components through extensive ablations. By empirically proving the

stability of RSMI, we put it forward as a practical method to robustly train large-scale NLP models. Our code and datasets are available at https://github.com/Han8931/rsmi_nlp

Exploring Lottery Prompts for Pre-trained Language Models
Yulin Chen, Ning Ding, Xiaobin Wang, Shengding Hu, Haitao Zheng, Zhiyuan Liu and Pengjun Xie   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Consistently scaling pre-trained language models (PLMs) imposes substantial burdens on model adaptation, necessitating more efficient alternatives to conventional fine-tuning. Given the advantage of prompting in the zero-shot setting and the observed performance fluctuation among different prompts, we explore the instance-level prompt and their generalizability. By searching through the prompt space, we first validate the assumption that for every instance, there is almost always a lottery prompt that induces the correct prediction from the PLM, and such prompt can be obtained at a low cost thanks to the inherent ability of PLMs. Meanwhile, it is shown that some strong lottery prompts have high performance over the whole training set, and they are equipped with distinguishable linguistic features. Lastly, we attempt to generalize the searched strong lottery prompts to unseen data with prompt ensembling method. Experiments are conducted on various types of NLP classification tasks and demonstrate that the proposed method can achieve comparable results with other gradient-free and optimization-free baselines.

How to Plant Trees in Language Models: Data and Architectural Effects on the Emergence of Syntactic Inductive Biases
Aaron Mueller and Tal Linzen   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Accurate syntactic representations are essential for robust generalization in natural language. Recent work has found that pre-training can teach language models to rely on hierarchical syntactic features—as opposed to incorrect linear features—when performing tasks after fine-tuning. We test what aspects of pre-training are important for endowing encoder-decoder Transformers with an inductive bias that favors hierarchical syntactic generalizations. We focus on architectural features (depth, width, and number of parameters), as well as the genre and size of the pre-training corpus, diagnosing inductive biases using two syntactic transformation tasks: question formation and passivization, both in English. We find that the number of parameters alone does not explain hierarchical generalization: model depth plays greater role than model width. We also find that pre-training on simpler language, such as child-directed speech, induces a hierarchical bias using an order-of-magnitude less data than pre-training on more typical datasets based on web text or Wikipedia; this suggests that in cognitively plausible language acquisition settings, neural language models may be more data-efficient than previously thought.

Being Right for Whose Right Reasons?
Terne Sasha Thorn Jakobsen, Laura Cabello and Anders Søgaard   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Explainability methods are used to benchmark the extent to which model predictions align with human rationales i.e., are 'right for the right reasons'. Previous work has failed to acknowledge, however, that what counts as a rationale is sometimes subjective. This paper presents what we think is a first of its kind, a collection of human rationale annotations augmented with the annotators demographic information. We cover three datasets spanning sentiment analysis and common-sense reasoning, and six demographic groups (balanced across age and ethnicity). Such data enables us to ask both what demographics our predictions align with and whose reasoning patterns our models' rationales align with. We find systematic inter-group annotator disagreement and show how 16 Transformer-based models align better with rationales provided by certain demographic groups: We find that models are biased towards aligning best with older and/or white annotators. We zoom in on the effects of model size and model distillation, finding –contrary to our expectations– negative correlations between model size and rationale agreement as well as no evidence that either model size or model distillation improves fairness.

infoVerse: A Universal Framework for Dataset Characterization with Multidimensional Meta-information
Jaehyung Kim, Yekyung Kim, Karin Johanna Denton de Langis, Jinwoo Shin and Dongyeop Kang   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
The success of NLP systems often relies on the availability of large, high-quality datasets. However, not all samples in these datasets are equally valuable for learning, as some may be redundant or noisy. Several methods for characterizing datasets based on model-driven meta-information (e.g., model's confidence) have been developed, but the relationship and complementary effects of these methods have received less attention. In this paper, we introduce infoVerse, a universal framework for dataset characterization, which provides a new feature space that effectively captures multidimensional characteristics of datasets by incorporating various model-driven meta-information. infoVerse reveals distinctive regions of the dataset that are not apparent in the original semantic space, hence guiding users (or models) in identifying which samples to focus on for exploration, assessment, or annotation. Additionally, we propose a novel sampling method on infoVerse to select a set of data points that maximizes informativeness. In three real-world applications (data pruning, active learning, and data annotation), the samples chosen on infoVerse space consistently outperform strong baselines in all applications. Our code and demo are publicly available.

The Ecological Fallacy in Annotation: Modeling Human Label Variation goes beyond Sociodemographics
Matthias Orlikowski, Paul Röttger, Philipp Cimiano and Dirk Hovy   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Many NLP tasks exhibit human label variation, where different annotators give different labels to the same texts. This variation is known to depend, at least in part, on the sociodemographics of annotators. Recent research aims to model individual annotator behaviour rather than predicting aggregated labels, and we would expect that sociodemographic information is useful for these models. On the other hand, the ecological fallacy states that aggregate group behaviour, such as the behaviour of the average female annotator, does not necessarily explain individual behaviour. To account for sociodemographics of models of individual annotator behaviour, we introduce group-specific layers to multi-annotator models. In a series of experiments for toxic content detection, we find that explicitly accounting for sociodemographic attributes in this way does not significantly improve model performance. This result shows that individual annotation behaviour depends on much more than just sociodemographics.

Computational Social Science and Cultural Analytics

09:00-10:30 (Pier 2&3)

Conflicts, Villains, Resolutions: Towards models of Narrative Media Framing
Lea Frermann, Jiatong Li, Shima Khanehzar and Gosia Mikolajczak   09:00-09:15 (Pier 2&3)
Despite increasing interest in the automatic detection of media frames in NLP, the problem is typically simplified as single-label classification and adopts a topic-like view on frames, evading modelling the broader document-level narrative. In this work, we revisit a widely used conceptualization of framing from the communication sciences which explicitly captures elements of narratives, including conflict and its resolution, and integrate it with the narrative framing of key entities in the story as heroes, victims or villains. We adapt an effective annotation paradigm that breaks a complex annotation task into a series of simpler binary questions, and present an annotated data set of English news articles, and a case study on the framing of climate change in articles from news outlets across the political spectrum. Finally, we explore

automatic multi-label prediction of our frames with supervised and semi-supervised approaches, and present a novel retrieval-based method which is both effective and transparent in its predictions. We conclude with a discussion of opportunities and challenges for future work on document-level models of narrative framing.

My side, your side and the evidence: Discovering aligned actor groups and the narratives they weave
Pavan Holur, David Chong, Timothy R Tangherlini and Vwani Roychowdhury   09:15-09:30 (Pier 2&3)
News reports about emerging issues often include several conflicting story lines. Individual stories can be conceptualized as samples from an underlying mixture of competing narratives. The automated identification of these distinct narratives from unstructured text is a fundamental yet difficult task in Computational Linguistics since narratives are often intertwined and only implicitly conveyed in text. In this paper, we consider a more feasible proxy task: Identify the distinct sets of aligned story actors responsible for sustaining the issue-specific narratives. Discovering aligned actors, and the groups these alignments create, brings us closer to estimating the narrative that each group represents. With the help of Large Language Models (LLM), we address this task by: (i) Introducing a corpus of text segments rich in narrative content associated with six different current issues; (ii) Introducing a novel two-step graph-based framework that (a) identifies alignments between actors (INCANT) and (b) extracts aligned actor groups using the network structure (TAMPA). Amazon Mechanical Turk evaluations demonstrate the effectiveness of our framework. Across domains, alignment relationships from INCANT are accurate (macro F1 >= 0.75) and actor groups from TAMPA are preferred over 2 non-trivial baseline models (ACC >= 0.75).

Grounding Characters and Places in Narrative Text
Sandeep Soni, Amanpreet Sihra, Elizabeth F. Evans, Matthew Wilkens and David Bamman   09:30-09:45 (Pier 2&3)
Tracking characters and locations throughout a story can help improve the understanding of its plot structure. Prior research has analyzed characters and locations from text independently without grounding characters to their locations in narrative time. Here, we address this gap by proposing a new spatial relationship categorization task. The objective of the task is to assign a spatial relationship category for every character and location co-mention within a window of text, taking into consideration linguistic context, narrative tense, and temporal scope. To this end, we annotate spatial relationships in approximately 2500 book excerpts and train a model using contextual embeddings as features to predict these relationships. When applied to a set of books, this model allows us to test several hypotheses on mobility and domestic space, revealing that protagonists are more mobile than non-central characters and that women as characters tend to occupy more interior space than men. Overall, our work is the first step towards joint modeling and analysis of characters and places in narrative text.

Your spouse needs professional help: Determining the Contextual Appropriateness of Messages through Modeling Social Relationships
David Jurgens, Agrima Seth, Jackson Sargent, Athena Aghighi and Michael Geraci   09:45-10:00 (Pier 2&3)
Understanding interpersonal communication requires, in part, understanding the social context and norms in which a message is said. However, current methods for identifying offensive content in such communication largely operate independent of context, with only a few approaches considering community norms or prior conversation as context. Here, we introduce a new approach to identifying inappropriate communication by explicitly modeling the social relationship between the individuals. We introduce a new dataset of contextually-situated judgments of appropriateness and show that large language models can readily incorporate relationship information to accurately identify appropriateness in a given context. Using data from online conversations and movie dialogues, we provide insight into how the relationships themselves function as implicit norms and quantify the degree to which context-sensitivity is needed in different conversation settings. Further, we also demonstrate that contextual-appropriateness judgments are predictive of other social factors expressed in language such as condescension and politeness.

Understanding Client Reactions in Online Mental Health Counseling
Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu and Zhenzhong Lan   10:00-10:15 (Pier 2&3)
Communication success relies heavily on reading participants' reactions. Such feedback is especially important for mental health counselors, who must carefully consider the client's progress and adjust their approach accordingly. However, previous NLP research on counseling has mainly focused on studying counselors' intervention strategies rather than their clients' reactions to the intervention. This work aims to fill this gap by developing a theoretically grounded annotation framework that encompasses counselors' strategies and client reaction behaviors. The framework has been tested against a large-scale, high-quality test-based counseling dataset we collected over the past two years from an online welfare counseling platform. Our study show how clients react to counselors' strategies, how such reactions affect the final counseling outcomes, and how counselors can adjust their strategies in response to these reactions. We also demonstrate that this study can help counselors automatically predict their clients' states.

Industry track: Model efficiency, Information Extraction

09:00-10:30 (Pier 4&5)

[Industry] pNLP-Mixer: an Efficient all-MLP Architecture for Language
Francesco Fusco, Damian Pascual, Peter Staar and Diego Antognini   09:00-09:15 (Pier 4&5)
Large pre-trained language models based on transformer architecture*f*have drastically changed the natural language processing (NLP) landscape. However, deploying those models for on-device applications in constrained devices such as smart watches is considerably impractical due to their size and inference cost. As an alternative to transformer-based architectures, recent work on efficient NLP has shown that weight-efficient models can attain competitive performance for simple tasks, such as slot filling and intent classification, with model sizes in the order of the megabyte. This work introduces the pNLP-Mixer architecture, an embedding-free MLP-Mixer model for on-device NLP that achieves high weight-efficiency thanks to a novel projection layer. We evaluate a pNLP-Mixer model of only one megabyte in size on two multi-lingual semantic parsing datasets, MTOP and multiATIS. Our quantized model achieves 99.4% and 97.8% the performance of mBERT on MTOP and multiATIS, while using 170x less parameters. Our model consistently beats the state-of-the-art of tiny models (pQRNN), which is twice as large, by a margin up to 7.8% on MTOP.

[Industry] BADGE: Speeding Up BERT Inference after Deployment via Block-wise Bypasses and Divergence-based Early Exiting
Wei Zhu, Peng Wang, Yuan Ni, Guotong Xie and Xiaoling Wang   09:15-09:30 (Pier 4&5)
Early exiting can reduce the average latency of pre-trained language models (PLMs) via its adaptive inference mechanism and work with other inference speed-up methods like model pruning, thus drawing much attention from the industry. In this work, we propose a novel framework, BADGE, which consists of two off-the-shelf methods for improving PLMs' early exiting. We first address the issues of training a multi-exit PLM, the backbone model for early exiting. We propose the novel architecture of block-wise bypasses, which can alleviate the conflicts in jointly training multiple intermediate classifiers and thus improve the overall performances of multi-exit PLM while introducing negligible additional flops to the model. Second, we propose a novel divergence-based early exiting (DGE) mechanism, which obtains early

exiting signals by comparing the predicted distributions of two adjacent layers' exits. Extensive experiments on three proprietary datasets and three GLUE benchmark tasks demonstrate that our method can obtain a better speedup-performance trade-off than the existing baseline methods.\footnote{Code will be made publicly available to the research community upon acceptance.}

[Industry] K-pop and fake facts: from texts to smart alerting for maritime security
Maxime Prieur, Souhir Gahbiche, Guillaume Gadek, Sylvain Gatepaille, Kilian Vasnier and Valerian Justine　　09:30-09:45 (Pier 4&5)
Maritime security requires full-time monitoring of the situation, mainly based on technical data (radar, AIS) but also from OSINT-like inputs (e.g., newspapers). Some threats to the operational reliability of this maritime surveillance, such as malicious actors, introduce discrepancies between hard and soft data (sensors and texts), either by tweaking their AIS emitters or by emitting false information on pseudo-newspapers. Many techniques exist to identify these pieces of false information, including using knowledge base population techniques to build a structured view of the information. This paper presents a use case for suspect data identification in a maritime setting. The proposed system UMBAR ingests data from sensors and texts, processing them through an information extraction step, in order to feed a Knowledge Base and finally perform coherence checks between the extracted facts.

[Industry] Context-Aware Query Rewriting for Improving Users' Search Experience on E-commerce Websites
Simiao Zuo, Qingyu Yin, Haoming Jiang, Shaohui Xi, Bing Yin, Chao Zhang and Tuo Zhao　　09:45-10:00 (Pier 4&5)
E-commerce queries are often short and ambiguous. Consequently, query understanding often uses query rewriting to disambiguate user-input queries. While using e-commerce search tools, users tend to enter multiple searches, which we call context, before purchasing. These history searches contain contextual insights about users' true shopping intents. Therefore, modeling such contextual information is critical to a better query rewriting model. However, existing query rewriting models ignore users' history behaviors and consider only the instant search query, which is often a short string offering limited information about the true shopping intent. We propose an end-to-end context-aware query rewriting model to bridge this gap, which takes the search context into account. Specifically, our model builds a session graph using the history search queries and their contained words. We then employ a graph attention mechanism that models cross-query relations and computes contextual information of the session. The model subsequently calculates session representations by combining the contextual information with the instant search query using an aggregation network. The session representations are then decoded to generate rewritten queries. Empirically, we demonstrate the superiority of our method to state-of-the-art approaches under various metrics.

[Industry] Large Scale Generative Multimodal Attribute Extraction for E-commerce Attributes
Anant Khandelwal, Happy Mittal, Shreyas Kulkarni and Deepak Gupta　　10:00-10:15 (Pier 4&5)
E-commerce websites (e.g. Amazon, Alibaba) have a plethora of structured and unstructured information (text and images) present on the product pages. Sellers often don't label or mislabel values of the attributes (e.g. color, size etc.) for their products. Automatically identifying these attribute values from an eCommerce product page that contains both text and images is a challenging task, especially when the attribute value is not explicitly mentioned in the catalog. In this paper, we present a scalable solution for this problem where we pose attribute extraction problem as a question-answering task, which we solve using MXT, that consists of three key components: (i) MAG (Multimodal Adaptation Gate), (ii) Xception network, and (iii) T5 encoder-decoder. Our system consists of a generative model that generates attribute-values for a given product by using both textual and visual characteristics (e.g. images) of the product. We show that our system is capable of handling zero-shot attribute prediction (when attribute value is not seen in training data) and value-absent prediction (when attribute value is not mentioned in the text) which are missing in traditional classification-based and NER-based models respectively. We have trained our models using distant supervision, removing dependency on human labeling, thus making them practical for real-world applications. With this framework, we are able to train a single model for 1000s of (product-type, attribute) pairs, thus reducing the overhead of training and maintaining separate models. Extensive experiments on two real world datasets (total 57 attributes) show that our framework improves the absolute recall@90P by 10.16% and 6.9\ from the existing state of the art models. In a popular e-commerce store, we have productionized our models that cater to >12K (product-type, attribute) pairs, and have extracted >150MM attribute values.

[Industry] Annotating Research Infrastructure in Scientific Papers: An NLP-driven Approach
Seyed Amin Tabatabaei, Georgios Cheirmpos, Marius Doornenbal, Alberto Zigoni, Veronique Moore and Georgios Tsatsaronis 10:15-10:30 (Pier 4&5)
In this work, we present a natural language processing (NLP) pipeline for the identification, extraction and linking of Research Infrastructure (RI) used in scientific publications. Links between scientific equipment and publications where the equipment was used can support multiple use cases, such as evaluating the impact of RI investment, and supporting Open Science and research reproducibility. These links can also be used to establish a profile of the RI portfolio of each institution and associate each equipment with scientific output. The system we are describing here is already in production, and has been used to address real business use cases, some of which we discuss in this paper. The computational pipeline at the heart of the system comprises both supervised and unsupervised modules to detect the usage of research equipment by processing the full text of the articles. Additionally, we have created a knowledge graph of RI, which is utilized to annotate the articles with metadata. Finally, examples of the business value of the insights made possible by this NLP pipeline are illustrated.

Linguistic Diversity

09:00-10:30 (Pier 7&8)

Script Normalization for Unconventional Writing of Under-Resourced Languages in Bilingual Communities
Sina Ahmadi and Antonios Anastasopoulos　　09:00-09:15 (Pier 7&8)
The wide accessibility of social media has provided linguistically under-represented communities with an extraordinary opportunity to create content in their native languages. This, however, comes with certain challenges in script normalization, particularly where the speakers of a language in a bilingual community rely on another script or orthography to write their native language. This paper addresses the problem of script normalization for several such languages that are mainly written in a Perso-Arabic script. Using synthetic data with various levels of noise and a transformer-based model, we demonstrate that the problem can be effectively remediated. We conduct a small-scale evaluation of real data as well. Our experiments indicate that script normalization is also beneficial to improve the performance of downstream tasks such as machine translation and language identification.

Small Data, Big Impact: Leveraging Minimal Data for Effective Machine Translation
Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan and Francisco Guzman　　09:15-09:30 (Pier 7&8)
For many languages, machine translation progress is hindered by the lack of reliable training data. Models are trained on whatever pre-existing datasets may be available and then augmented with synthetic data, because it is often not economical to pay for the creation of large-scale datasets. But for the case of low-resource languages, would the creation of a few thousand professionally translated sentence pairs give any

benefit? In this paper, we show that it does.

We describe a broad data collection effort involving around 6k professionally translated sentence pairs for each of 39 low-resource languages, which we make publicly available. We analyse the gains of models trained on this small but high-quality data, showing that it has significant impact even when larger but lower quality pre-existing corpora are used, or when data is augmented with millions of sentences through back-translation.

Question-Answering in a Low-resourced Language: Benchmark Dataset and Models for Tigrinya
Fitsum Gaim, Wonsuk Yang, Hancheol Park and Jong Park   09:30-09:45 (Pier 7&8)
Question-Answering (QA) has seen significant advances recently, achieving near human-level performance over some benchmarks. However, these advances focus on high-resourced languages such as English, while the task remains unexplored for most other languages, mainly due to the lack of annotated datasets. This work presents a native QA dataset for an East African language, Tigrinya. The dataset contains 10.6K question-answer pairs spanning 572 paragraphs extracted from 290 news articles on various topics. The dataset construction method is discussed, which is applicable to constructing similar resources for related languages. We present comprehensive experiments and analyses of several resource-efficient approaches to QA, including monolingual, cross-lingual, and multilingual setups, along with comparisons against machine-translated silver data. Our strong baseline models reach 76% in the F1 score, while the estimated human performance is 92%, indicating that the benchmark presents a good challenge for future work. We make the dataset, models, and leaderboard publicly available.

MasakhaPOS: Part-of-Speech Tagging for Typologically Diverse African languages
Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Andrew Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing K Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson K KALIPE, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Ekene Onyenwe, Gratien G. Atindogbe, Tolulope Anu Adelani, Idris Akinade, Olanrewaju Samuel, Marien NAHIMANA, Théogène MUSABEYEZU, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete AGBOLO, SEYDOU TRAORE, Chinedu Uchechukwu, Aliyu Yakubu Yusuf, Muhammad Sulaiman Abdullahi and Dietrich Klakow 09:45-10:00 (Pier 7&8)
In this paper, we present AfricaPOS, the largest part-of-speech (POS) dataset for 20 typologically diverse African languages. We discuss the challenges in annotating POS for these languages using the universal dependencies (UD) guidelines. We conducted extensive POS baseline experiments using both conditional random field and several multilingual pre-trained language models. We applied various cross-lingual transfer models trained with data available in the UD. Evaluating on the AfricaPOS dataset, we show that choosing the best transfer language(s) in both single-source and multi-source setups greatly improves the POS tagging performance of the target languages, in particular when combined with parameter-fine-tuning methods. Crucially, transferring knowledge from a language that matches the language family and morphosyntactic properties seems to be more effective for POS tagging in unseen languages.

An (unhelpful) guide to selecting the best ASR architecture for your under-resourced language
Robert Jimerson, Zoey Liu and Emily Prud'hommeaux   10:00-10:15 (Pier 7&8)
Advances in deep neural models for automatic speech recognition (ASR) have yielded dramatic improvements in ASR quality for resource-rich languages, with English ASR now achieving word error rates comparable to that of human transcribers. The vast majority of the world's languages, however, lack the quantity of data necessary to approach this level of accuracy. In this paper we use four of the most popular ASR toolkits to train ASR models for eleven languages with limited ASR training resources: eleven widely spoken languages of Africa, Asia, and South America, one endangered language of Central America, and three critically endangered languages of North America. We find that no single architecture consistently outperforms any other. These differences in performance so far do not appear to be related to any particular feature of the datasets or characteristics of the languages. These findings have important implications for future research in ASR for under-resourced languages. ASR systems for languages with abundant existing media and available speakers may derive the most benefit simply by collecting large amounts of additional acoustic and textual training data. Communities using ASR to support endangered language documentation efforts, who cannot easily collect more data, might instead focus on exploring multiple architectures and hyperparameterizations to optimize performance within the constraints of their available data and resources.

NollySenti: Leveraging Transfer Learning and Machine Translation for Nigerian Movie Sentiment Classification
Iyanuoluwa Adeola Shode, David Ifeoluwa Adelani, JIng Peng and Anna Feldman   10:15-10:30 (Pier 7&8)
Africa has over 2000 indigenous languages but they are under-represented in NLP research due to lack of datasets. In recent years, there have been progress in developing labelled corpora for African languages. However, they are often available in a single domain and may not generalize to other domains.

In this paper, we focus on the task of sentiment classification for cross-domain adaptation. We create a new dataset, Nollywood movie reviews for five languages widely spoken in Nigeria (English, Hausa, Igbo, Nigerian Pidgin, and Yoruba). We provide an extensive empirical evaluation using classical machine learning methods and pre-trained language models.

By leveraging transfer learning, we compare the performance of cross-domain adaptation from Twitter domain, and cross-lingual adaptation from English language. Our evaluation shows that transfer from English in the same target domain leads to more than 5% improvement in accuracy compared to transfer from Twitter in the same language.

To further mitigate the domain difference, we leverage machine translation from English to other Nigerian languages, which leads to a further improvement of 7% over cross-lingual evaluation. While machine translation to low-resource languages are often of low quality, our analysis shows that sentiment related words are often preserved.

Session 4 - 11:00-12:30

Resources and Evaluation

11:00-12:30 (Metropolitan East)

Evaluating Open-Domain Dialogues in Latent Space with Next Sentence Prediction and Mutual Information
Kun Zhao, Bohao Yang, Chenghua Lin, Wenge Rong, Aline Villavicencio and Xiaohui Cui   11:00-11:15 (Metropolitan East)
The long-standing one-to-many issue of the open-domain dialogues poses significant challenges for automatic evaluation methods, i.e., there may be multiple suitable responses which differ in semantics for a given conversational context. To tackle this challenge, we propose a novel learning-based automatic evaluation metric (CMN), which can robustly evaluate open-domain dialogues by augmenting Conditional Variational Autoencoders (CVAEs) with a Next Sentence Prediction (NSP) objective and employing Mutual Information (MI) to model the

semantic similarity of text in the latent space. Experimental results on two open-domain dialogue datasets demonstrate the superiority of our method compared with a wide range of baselines, especially in handling responses which are distant to the "golden" reference responses in semantics.

SafeConv: Explaining and Correcting Conversational Unsafe Behavior

Mian Zhang, Lifeng Jin, Linfeng Song, Haitao Mi, Wenliang Chen and Dong Yu   11:15-11:30 (Metropolitan East)

One of the main challenges open-domain end-to-end dialogue systems, or chatbots, face is the prevalence of unsafe behavior, such as toxic languages and harmful suggestions. However, existing dialogue datasets do not provide enough annotation to explain and correct such unsafe behavior. In this work, we construct a new dataset called SafeConv for the research of conversational safety: (1) Besides the utterance-level safety labels, SafeConv also provides unsafe spans in an utterance, information able to indicate which words contribute to the detected unsafe behavior; (2) SafeConv provides safe alternative responses to continue the conversation when unsafe behavior detected, guiding the conversation to a gentle trajectory.

By virtue of the comprehensive annotation of SafeConv, we benchmark three powerful models for the mitigation of conversational unsafe behavior, including a checker to detect unsafe utterances, a tagger to extract unsafe spans, and a rewriter to convert an unsafe response to a safe version. Moreover, we explore the huge benefits brought by combining the models for explaining the emergence of unsafe behavior and detoxifying chatbots. Experiments show that the detected unsafe behavior could be well explained with unsafe spans and popular chatbots could be detoxified by a huge extent. The dataset is available at https://github.com/mianzhang/SafeConv.

Evaluating Open-Domain Question Answering in the Era of Large Language Models

Ehsan Kamalloo, Nouha Dziri, Charles Clarke and Davood Rafiei   11:30-11:45 (Metropolitan East)

Lexical matching remains the de facto evaluation method for open-domain question answering (QA). Unfortunately, lexical matching fails completely when a plausible candidate answer does not appear in the list of gold answers, which is increasingly the case as we shift from extractive to generative models. The recent success of large language models (LLMs) for QA aggravates lexical matching failures since candidate answers become longer, thereby making matching with the gold answers even more challenging. Without accurate evaluation, the true progress in open-domain QA remains unknown. In this paper, we conduct a thorough analysis of various open-domain QA models, including LLMs, by manually evaluating their answers on a subset of NQ-open, a popular benchmark. Our assessments reveal that while the true performance of all models is significantly underestimated, the performance of the InstructGPT (zero-shot) LLM increases by nearly +60%, making it on par with existing top models, and the InstructGPT (few-shot) model actually achieves a new state-of-the-art on NQ-open. We also find that more than 50% of lexical matching failures are attributed to semantically equivalent answers. We further demonstrate that regex matching ranks QA models consistent with human judgments, although still suffering from unnecessary strictness. Finally, we demonstrate that automated evaluation models are a reasonable surrogate for lexical matching in some circumstances, but not for long-form answers generated by LLMs. The automated models struggle in detecting hallucinations in LLM answers and are thus unable to evaluate LLMs. At this time, there appears to be no substitute for human evaluation.

DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover and Duen Horng Chau   11:45-12:00 (Metropolitan East)

With recent advancements in diffusion models, users can generate high-quality images by writing text prompts in natural language. However, generating images with desired details requires proper prompts, and it is often unclear how a model reacts to different prompts or what the best prompts are. To help researchers tackle these critical challenges, we introduce DiffusionDB, the first large-scale text-to-image prompt dataset totaling 6.5TB, containing 14 million images generated by Stable Diffusion, 1.8 million unique prompts, and hyperparameters specified by real users. We analyze the syntactic and semantic characteristics of prompts. We pinpoint specific hyperparameter values and prompt styles that can lead to model errors and present evidence of potentially harmful model usage, such as the generation of misinformation. The unprecedented scale and diversity of this human-actuated dataset provide exciting research opportunities in understanding the interplay between prompts and generative models, detecting deepfakes, and designing human-AI interaction tools to help users more easily use these models. DiffusionDB is publicly available at: https://poloclub.github.io/diffusiondb.

Tell2Design: A Dataset for Language-Guided Floor Plan Generation

Sicong Leng, Yang Zhou, Mohammed Haroon Dupty, Wee Sun Lee, Sam C Joyce and Wei Lu   12:00-12:15 (Metropolitan East)

We consider the task of generating designs directly from natural language descriptions, and consider floor plan generation as the initial research area. Language conditional generative models have recently been very successful in generating high-quality artistic images. However, designs must satisfy different constraints that are not present in generating artistic images, particularly spatial and relational constraints. We make multiple contributions to initiate research on this task. First, we introduce a novel dataset, Tell2Design (T2D), which contains more than 80k floor plan designs associated with natural language instructions. Second, we propose a Sequence-to-Sequence model that can serve as a strong baseline for future research. Third, we benchmark this task with several text-conditional image generation models. We conclude by conducting human evaluations on the generated samples and providing an analysis of human performance. We hope our contributions will propel the research on language-guided design generation forward.

CREPE: Open-Domain Question Answering with False Presuppositions

Xinyan Velocity Yu, Sewon Min, Luke Zettlemoyer and Hannaneh Hajishirzi   12:15-12:30 (Metropolitan East)

When asking about unfamiliar topics, information seeking users often pose questions with false presuppositions. Most existing question answering (QA) datasets, in contrast, assume all questions have well defined answers. We introduce CREPE, a QA dataset containing a natural distribution of presupposition failures from online information-seeking forums. We find that 25% of questions contain false presuppositions, and provide annotations for these presuppositions and their corrections. Through extensive baseline experiments, we show that adaptations of existing open-domain QA models can find presuppositions moderately well, but struggle when predicting whether a presupposition is factually correct. This is in large part due to difficulty in retrieving relevant evidence passages from a large text corpus. CREPE provides a benchmark to study question answering in the wild, and our analyses provide avenues for future work in better modeling and further studying the task.

Large Language Models

11:00-12:30 (Metropolitan Centre)

RetroMAE-2: Duplex Masked Auto-Encoder For Pre-Training Retrieval-Oriented Language Models

Zheng Liu, Shitao Xiao, Yingxia Shao and Zhao Cao   11:00-11:15 (Metropolitan Centre)

To better support information retrieval tasks such as web search and open-domain question answering, growing effort is made to develop retrieval-oriented language models, e.g., RetroMAE and many others. Most of the existing works focus on improving the semantic represen-

tation capability for the contextualized embedding of the [CLS] token. However, recent study shows that the ordinary tokens besides [CLS] may provide extra information, which help to produce a better representation effect. As such, it's necessary to extend the current methods where all contextualized embeddings can be jointly pre-trained for the retrieval tasks.

In this work, we propose a novel pre-training method called Duplex Masked Auto-Encoder, a.k.a. DupMAE. It is designed to improve the quality of semantic representation where all contextualized embeddings of the pre-trained model can be leveraged. It takes advantage of two complementary auto-encoding tasks: one reconstructs the input sentence on top of the [CLS] embedding; the other one predicts the bag-of-words feature of the input sentence based on the ordinary tokens' embeddings. The two tasks are jointly conducted to train a unified encoder, where the whole contextualized embeddings are aggregated in a compact way to produce the final semantic representation. DupMAE is simple but empirically competitive: it substantially improves the pre-trained model's representation capability and transferability, where superior retrieval performances can be achieved on popular benchmarks, like MS MARCO and BEIR. We make our code publicly available at https://github.com/staoxiao/RetroMAE.

Augmentation-Adapted Retriever Improves Generalization of Language Models as Generic Plug-In

Zichun Yu, Chenyan Xiong, Shi Yu and Zhiyuan Liu   11:15-11:30 (Metropolitan Centre)

Retrieval augmentation can aid language models (LMs) in knowledge-intensive tasks by supplying them with external information. Prior works on retrieval augmentation usually jointly fine-tune the retriever and the LM, making them closely coupled. In this paper, we explore the scheme of generic retrieval plug-in: the retriever is to assist target LMs that may not be known beforehand or are unable to be fine-tuned together. To retrieve useful documents for unseen target LMs, we propose augmentation-adapted retriever (AAR), which learns LM's preferences obtained from a known source LM. Experiments on the MMLU and PopQA datasets demonstrate that our AAR trained with a small source LM is able to significantly improve the zero-shot generalization of larger target LMs ranging from 250M Flan-T5 to 175B InstructGPT. Further analysis indicates that the preferences of different LMs overlap, enabling AAR trained with a single source LM to serve as a generic plug-in for various target LMs. Our code is open-sourced at https://github.com/OpenMatch/Augmentation-Adapted-Retriever.

Pre-trained Language Models Can be Fully Zero-Shot Learners

Xuandong Zhao, Siqi Ouyang, Zhiguo Yu, Ming Wu and Lei Li   11:30-11:45 (Metropolitan Centre)

How can we extend a pre-trained model to many language understanding tasks, without labeled or additional unlabeled data? Pre-trained language models (PLMs) have been effective for a wide range of NLP tasks. However, existing approaches either require fine-tuning on downstream labeled datasets or manually constructing proper prompts. In this paper, we propose nonparametric prompting PLM (NPPrompt) for fully zero-shot language understanding. Unlike previous methods, NPPrompt uses only pre-trained language models and does not require any labeled data or additional raw corpus for further fine-tuning, nor does it rely on humans to construct a comprehensive set of prompt label words. We evaluate NPPrompt against previous major few-shot and zero-shot learning methods on diverse NLP tasks: including text classification, text entailment, similar text retrieval, paraphrasing, and multiple-choice question answering. Experimental results demonstrate that our NPPrompt outperforms the previous best fully zero-shot method by big margins, with absolute gains of 12.8% in accuracy on text classification and 15.6% on the GLUE benchmark. Our source code is available at https://anonymous.4open. science/r/NPPrompt.

Multilingual LLMs are Better Cross-lingual In-context Learners with Alignment

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur and Tanmoy Chakraborty   11:45-12:00 (Metropolitan Centre)

In-context learning (ICL) unfolds as large language models become capable of inferring test labels conditioned on a few labeled samples without any gradient update. ICL-enabled large language models provide a promising step forward toward bypassing recurrent annotation costs in a low-resource setting. Yet, only a handful of past studies have explored ICL in a cross-lingual setting, in which the need for transferring label-knowledge from a high-resource language to a low-resource one is immensely crucial. To bridge the gap, we provide the first in-depth analysis of ICL for cross-lingual text classification. We find that the prevalent mode of selecting random input-label pairs to construct the prompt-context is severely limited in the case of cross-lingual ICL, primarily due to the lack of alignment in the input as well as the output spaces. To mitigate this, we propose a novel prompt construction strategy — Cross-lingual In-context Source Target Alignment (X-InSTA). With an injected coherence in the semantics of the input examples and a task-based alignment across the source and target languages, X-InSTA is able to outperform random prompt selection by a large margin across three different tasks using 44 different cross-lingual pairs.

Surface-Based Retrieval Reduces Perplexity of Retrieval-Augmented Language Models

Ehsan Doostmohammadi, Tobias Norlund, Marco Kuhlmann and Richard Johansson   12:00-12:15 (Metropolitan Centre)

Augmenting language models with a retrieval mechanism has been shown to significantly improve their performance while keeping the number of parameters low. Retrieval-augmented models commonly rely on a semantic retrieval mechanism based on the similarity between dense representations of the query chunk and potential neighbors. In this paper, we study the state-of-the-art Retro model and observe that its performance gain is better explained by surface-level similarities, such as token overlap. Inspired by this, we replace the semantic retrieval in Retro with a surface-level method based on BM25, obtaining a significant reduction in perplexity. As full BM25 retrieval can be computationally costly for large datasets, we also apply it in a re-ranking scenario, gaining part of the perplexity reduction with minimal computational overhead.

Understanding In-Context Learning via Supportive Pretraining Data

Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz and Tianlu Wang   12:15-12:30 (Metropolitan Centre)

In-context learning (ICL) improves language models' performance on a variety of NLP tasks by simply demonstrating a handful of examples at inference time. It is not well understood why ICL ability emerges, as the model has never been specifically trained on such demonstrations. Unlike prior work that explores implicit mechanisms behind ICL, we study ICL via investigating the pretraining data. Specifically, we first adapt an iterative, gradient-based approach to find a small subset of pretraining data that supports ICL. We observe that a continued pretraining on this small subset significantly improves the model's ICL ability, by up to 18%. We then compare the supportive subset constrastively with random subsets of pretraining data and discover: (1) The supportive pretraining data to ICL do not have a higher domain relevance to downstream tasks. (2) The supportive pretraining data have a higher mass of rarely occurring, long-tail tokens. (3) The supportive pretraining data are challenging examples where the information gain from long-range context is below average, indicating learning to incorporate difficult long-range context encourages ICL. Our work takes a first step towards understanding ICL via analyzing instance-level pretraining data. Our insights have a potential to enhance the ICL ability of language models by actively guiding the construction of pretraining data in the future.

Summarization

11:00-12:30 (Metropolitan West)

What are the Desired Characteristics of Calibration Sets? Identifying Correlates on Long Form Scientific Summarization

Griffin Adams, Bichlien H Nguyen, Jake Allen Smith, Yingce Xia, Shufang Xie, Anna Ostropolets, Budhaditya Deb, Yuan-Jyue Chen, Tristan

Naumann and Noémie Elhadad   11:00-11:15 (Metropolitan West)
Summarization models often generate text that is poorly calibrated to quality metrics because they are trained to maximize the likelihood of a single reference (MLE). To address this, recent work has added a calibration step, which exposes a model to its own ranked outputs to improve relevance or, in a separate line of work, contrasts positive and negative sets to improve faithfulness. While effective, much of this work has focused on *how* to generate and optimize these sets. Less is known about *why* one setup is more effective than another. In this work, we uncover the underlying characteristics of effective sets. For each training instance, we form a large, diverse pool of candidates and systematically vary the subsets used for calibration fine-tuning. Each selection strategy targets distinct aspects of the sets, such as lexical diversity or the size of the gap between positive and negatives. On three diverse scientific long-form summarization datasets (spanning biomedical, clinical, and chemical domains), we find, among others, that faithfulness calibration is optimal when the negative sets are extractive and more likely to be generated, whereas for relevance calibration, the metric margin between candidates should be maximized and surprise–the disagreement between model and metric defined candidate rankings–minimized.

Balancing Lexical and Semantic Quality in Abstractive Summarization
Jeewoo Sul and Yong Suk Choi   11:15-11:30 (Metropolitan West)
An important problem of the sequence-to-sequence neural models widely used in abstractive summarization is exposure bias. To alleviate this problem, re-ranking systems have been applied in recent years. Despite some performance improvements, this approach remains under-explored. Previous works have mostly specified the rank through the ROUGE score and aligned candidate summaries, but there can be quite a large gap between the lexical overlap metric and semantic similarity. In this paper, we propose a novel training method in which a re-ranker balances the lexical and semantic quality. We further newly define false positives in ranking and present a strategy to reduce their influence. Experiments on the CNN/DailyMail and XSum datasets show that our method can estimate the meaning of summaries without seriously degrading the lexical aspect. More specifically, it achieves an 89.67 BERTScore on the CNN/DailyMail dataset, reaching new state-of-the-art performance. Our code is publicly available at https://github.com/jeewoo1025/BalSum.

Socratic Pretraining: Question-Driven Pretraining for Controllable Summarization
Artidoro Pagnoni, Alex Fabbri, Wojciech Kryscinski and Chien-Sheng Jason Wu   11:30-11:45 (Metropolitan West)
In long document controllable summarization, where labeled data is scarce, pretrained models struggle to adapt to the task and effectively respond to user queries. In this paper, we introduce Socratic pretraining, a question-driven, unsupervised pretraining objective specifically designed to improve controllability in summarization tasks. By training a model to generate and answer relevant questions in a given context, Socratic pretraining enables the model to more effectively adhere to user-provided queries and identify relevant content to be summarized. We demonstrate the effectiveness of this approach through extensive experimentation on two summarization domains, short stories and dialogue, and multiple control strategies: keywords, questions, and factoid QA pairs. Our pretraining method relies only on unlabeled documents and a question generation system and outperforms pre-finetuning approaches that use additional supervised data. Furthermore, our results show that Socratic pretraining cuts task-specific labeled data requirements in half, is more faithful to user-provided queries, and achieves state-of-the-art performance on QMSum and SQuALITY.

Attributable and Scalable Opinion Summarization
Tom Hosking, Hao Tang and Mirella Lapata   11:45-12:00 (Metropolitan West)
We propose a method for unsupervised opinion summarization that encodes sentences from customer reviews into a hierarchical discrete latent space, then identifies common opinions based on the frequency of their encodings. We are able to generate both abstractive summaries by decoding these frequent encodings, and extractive summaries by selecting the sentences assigned to the same frequent encodings. Our method is attributable, because the model identifies sentences used to generate the summary as part of the summarization process. It scales easily to many hundreds of input reviews, because aggregation is performed in the latent space rather than over long sequences of tokens. We also demonstrate that our approach enables a degree of control, generating aspect-specific summaries by restricting the model to parts of the encoding space that correspond to desired aspects (e.g., location or food). Automatic and human evaluation on two datasets from different domains demonstrates that our method generates summaries that are more informative than prior work and better grounded in the input reviews.

Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation
Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Jason Wu, Caiming Xiong and Dragomir Radev   12:00-12:15 (Metropolitan West)
Human evaluation is the foundation upon which the evaluation of both summarization systems and automatic metrics rests. However, existing human evaluation studies for summarization either exhibit a low inter-annotator agreement or have insufficient scale, and an in-depth analysis of human evaluation is lacking. Therefore, we address the shortcomings of existing summarization evaluation along the following axes: (1) We propose a modified summarization salience protocol, Atomic Content Units (ACUs), which is based on fine-grained semantic units and allows for a high inter-annotator agreement. (2) We curate the Robust Summarization Evaluation (RoSE) benchmark, a large human evaluation dataset consisting of 22,000 summary-level annotations over 28 top-performing systems on three datasets. (3) We conduct a comparative study of four human evaluation protocols, underscoring potential confounding factors in evaluation setups. (4) We evaluate 50 automatic metrics and their variants using the collected human annotations across evaluation protocols and demonstrate how our benchmark leads to more statistically stable and significant results. The metrics we benchmarked include recent methods based on large language models (LLMs), GPTScore and G-Eval. Furthermore, our findings have important implications for evaluating LLMs, as we show that LLMs adjusted by human feedback (e.g., GPT-3.5) may overfit unconstrained human evaluation, which is affected by the annotators' prior, input-agnostic preferences, calling for more robust, targeted evaluation methods.

[TACL] MACSum: Controllable Summarization with Mixed Attributes
Yusen Zhang, Yang Liu, Ziyi Yang, Yuwei Fang, Chenguang Zhu, Yulong Chen, Michael Zeng, Rui Zhang and Dragomir Radev   12:15-12:30 (Metropolitan West)
Controllable summarization allows users to generate customized summaries with specified attributes. However, due to the lack of designated annotations of controlled summaries, existing works have to craft pseudo datasets by adapting generic summarization benchmarks. Furthermore, most research focuses on controlling single attributes individually (e.g., a short summary or a highly abstractive summary) rather than controlling a mix of attributes together (e.g., a short and highly abstractive summary). In this paper, we propose MACSum, the first human-annotated summarization dataset for controlling mixed attributes. It contains source texts from two domains, news articles and dialogues, with human-annotated summaries controlled by five designed attributes (Length, Extractiveness, Specificity, Topic, and Speaker). We propose two simple and effective parameter-efficient approaches for the new task of mixed controllable summarization based on hard prompt tuning and soft prefix tuning. Results and analysis demonstrate that hard prompt models yield the best performance on most metrics and human evaluations. However, mixed-attribute control is still challenging for summarization tasks. We will release our data and code upon paper acceptance.

Posters

11:00-12:30 (Frontenac Ballroom and Queen's Quay)

[TACL] FaithDial: A Faithful Benchmark for Information-Seeking Dialogue
*Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti and Siva Reddy*11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The goal of information-seeking dialogue is to respond to seeker queries with natural language utterances that are grounded on knowledge sources.However, dialogue systems often produce unsupported utterances, a phenomenon known as hallucination. To mitigate this behavior, we adopt a data-centric solution and create FaithDial, a new benchmark for hallucination-free dialogues, by editing hallucinated responses in the Wizard of Wikipedia (WoW) benchmark. We observe that FaithDial is more faithful than WoW while also maintaining engaging conversations.We show that FaithDial can serve as training signal for: i) a hallucination critic, which discriminates whether an utterance is faithful or not, and boosts the performance by 12.8 F1 score on the BEGIN benchmark compared to existing datasets for dialogue coherence; ii) high-quality dialogue generation. We benchmark a series of state-of-the-art models and propose an auxiliary contrastive objective that achieves the highest level of faithfulness and abstractiveness based on several automated metrics. Further, we find that the benefits of FaithDial generalize to zero-shot transfer on other datasets, such as CMU-Dog and TopicalChat. Finally, human evaluation reveals that responses generated by models trained on FaithDial are perceived as more interpretable, cooperative, and engaging.

[TACL] Temporal Effects on Pre-trained Models for Language Processing Tasks
*Oshin Agarwal and Ani Nenkova*11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Keeping the performance of language technologies optimal as time passes is of great practical interest. We study temporal effects on model performance on downstream language tasks, establishing a nuanced terminology for such discussion and identifying factors essential to conduct a robust study. We present experiments for several tasks in English where the label correctness is not dependent on time and demonstrate the importance of distinguishing between temporal model deterioration and temporal domain adaptation for systems using pre-trained representations. We find that depending on the task, temporal model deterioration is not necessarily a concern. Temporal domain adaptation however is beneficial in all cases, with better performance for a given time period possible when the system is trained on temporally more recent data. Therefore, we also examine the efficacy of two approaches for temporal domain adaptation without human annotations on new data. Self-labeling shows consistent improvement and notably, for named entity recognition, leads to better temporal adaptation than even human annotations.

[TACL] Dubbing in Practice: A Large Scale Study of Human Localization With Insights for Automatic Dubbing
*Brian Thompson, William Brannon and Yogesh Virkar*11:00-12:30 (Frontenac Ballroom and Queen's Quay)
We investigate how humans perform the task of dubbing video content from one language into another, leveraging a novel corpus of 319.57 hours of video from 54 professionally produced titles. This is the first such large-scale study we are aware of. The results challenge a number of assumptions commonly made in both qualitative literature on human dubbing and machine-learning literature on automatic dubbing, arguing for the importance of vocal naturalness and translation quality over commonly emphasized isometric (character length) and lip-sync constraints, and for a more qualified view of the importance of isochronic (timing) constraints. We also find substantial influence of the source-side audio on human dubs through channels other than the words of the translation, pointing to the need for research on ways to preserve speech characteristics, as well as semantic transfer such as emphasis/emotion, in automatic dubbing systems.

[TACL] Time-and-Space-Efficient Weighted Deduction
*Jason Eisner*11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Many NLP algorithms have been described in terms of deduction systems. Unweighted deduction allows a generic forward-chaining execution strategy. For weighted deduction, however, efficient execution should propagate the weight of each item only after it has converged. This means visiting the items in topologically sorted order (as in dynamic programming). Toposorting is fast on a materialized graph; unfortunately, materializing the graph would take extra space. Is there a generic weighted deduction strategy which, for every acyclic deduction system and every input, uses only a constant factor more time and space than generic unweighted deduction? After reviewing past strategies, we answer this question in the affirmative by combining ideas of Goodman (1999) and Kahn (1962). We also give an extension to cyclic deduction systems, based on Tarjan (1972).

[TACL] Expectations over unspoken alternatives predict pragmatic inferences
*Jennifer Hu, Roger Levy, Judith Degen and Sebastian Schuster*11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Scalar inferences (SI) are a signature example of how humans interpret language based on unspoken alternatives. While empirical studies have demonstrated that human SI rates are highly variable – both within instances of a single scale, and across different scales – there have been few proposals that quantitatively explain both cross- and within-scale variation. Here, we test a shared mechanism explaining SI rates within and across scales: context-driven expectations about the unspoken alternatives. Using neural language models to approximate human predictive distributions, we test to what extent SIs are affected by (1) expectedness of the strong scalemate as an alternative, and (2) uncertainty about the underlying scale. We find that both predictors capture variation in SI rates, although the expectedness of the alternative robustly predicts cross-scale variation only under a meaning-based view of alternatives. Our results suggest that pragmatic inferences arise from context-driven expectations over alternatives, and these expectations are captured by neural language models.

[TACL] Neuron-level Interpretation of Deep NLP Models: A Survey
*Nadir Durrani, Hassan Sajjad and Fahim Dalvi*11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The proliferation of deep neural networks in various domains has seen an increased need for interpretability of these models. Preliminary work done along this line and papers that surveyed such, are focused on high-level representation analysis. However, a recent branch of work has concentrated on interpretability at a more granular level of analyzing neurons within these models. In this paper, we survey the work done on neuron analysis including: i) methods to discover and understand neurons in a network, ii) evaluation methods, iii) major findings including cross architectural comparisons that neuron analysis has unraveled, iv) applica- tions of neuron probing such as: controlling the model, domain adaptation etc., and v) a discussion on open issues and future research directions.

[Demo] A Hyperparameter Optimization Toolkit for Neural Machine Translation Research
*Paul McNamee, Kevin Duh and Xuan Zhang*11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Hyperparameter optimization is an important but often overlooked process in the research of deep learning technologies. To obtain a good model, one must carefully tune hyperparameters that determine the architecture and training algorithm. Insufficient tuning may result in poor results, while inequitable tuning may lead to exaggerated differences between models. We present a hyperparameter optimization toolkit for neural machine translation (NMT) to help researchers focus their time on the creative rather than the mundane. The toolkit is implemented as a wrapper on top of the open-source Sockeye NMT software. Using the Asynchronous Successive Halving Algorithm (ASHA), we demonstrate that it is possible to discover near-optimal models under a computational budget with little effort.

Code: https://github.com/kevinduh/sockeye-recipes3

Video demo: https://cs.jhu.edu/ kevinduh/j/demo.mp4

[Demo] OpenTIPE: An Open-source Translation Framework for Interactive Post-Editing Research

Laura Mascarell, Thomas Steinmann and Fabian Landwehr   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Despite the latest improvements on machine translation, professional translators still must review and post-edit the automatic output to ensure high-quality translations. The research on automating this process lacks an interactive post-editing environment implemented for this purpose; therefore, current approaches do not consider the human interactions that occur in real post-editing scenarios. To address this issue, we present OpenTIPE, a flexible and extensible framework that aims at supporting research on interactive post-editing. Specifically, the interactive environment of OpenTIPE allows researchers to explore human-centered approaches for the post-editing task. We release the OpenTIPE source code and showcase its main functionalities with a demonstration video and an online live demo.

[Demo] Massively Multi-Lingual Event Understanding: Extraction, Visualization, and Search

Elizabeth Boschee, Steven Fincke, Joel Barry, Shantanu Agarwal and Chris Jenkins   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

In this paper, we present ISI-Clear, a state-of-the-art, cross-lingual, zero-shot event extraction system and accompanying user interface for event visualization & search. Using only English training data, ISI-Clear makes global events available on-demand, processing user-supplied text in 100 languages ranging from Afrikaans to Yiddish. We provide multiple event-centric views of extracted events, including both a graphical representation and a document-level summary. We also integrate existing cross-lingual search algorithms with event extraction capabilities to provide cross-lingual event-centric search, allowing English-speaking users to search over events automatically extracted from a corpus of non-English documents, using either English natural language queries (e.g., "cholera outbreaks in Iran") or structured queries (e.g. find all events of type Disease-Outbreak with agent "cholera" and location "Iran").

[Demo] Riveter: Measuring Power and Social Dynamics Between Entities

Maarten Sap, Lauren Klein, Melanie Walsh, Jimin Mun, Anjalie Field and Maria Antoniak   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Riveter provides a complete easy-to-use pipeline for analyzing verb connotations associated with entities in text corpora. We prepopulate the package with connotation frames of sentiment, power, and agency, which have demonstrated usefulness for capturing social phenomena, such as gender bias, in a broad range of corpora. For decades, lexical frameworks have been foundational tools in computational social science, digital humanities, and natural language processing, facilitating multifaceted analysis of text corpora. But working with verb-centric lexica specifically requires natural language processing skills, reducing their accessibility to other researchers. By organizing the language processing pipeline, providing complete lexicon scores and visualizations for all entities in a corpus, and providing functionality for users to target specific research questions, Riveter greatly improves the accessibility of verb lexica and can facilitate a broad range of future research.

[Demo] WizMap: Scalable Interactive Visualization for Exploring Large Machine Learning Embeddings

Duen Horng Chau, Fred Hohman and Zijie J. Wang   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Machine learning models often learn latent embedding representations that capture the domain semantics of their training data. These embedding representations are valuable for interpreting trained models, building new models, and analyzing new datasets. However, interpreting and using embeddings can be challenging due to their opaqueness, high dimensionality, and the large size of modern datasets. To tackle these challenges, we present WizMap, an interactive visualization tool to help researchers and practitioners easily explore large embeddings. With a novel multi-resolution embedding summarization method and a familiar map-like interaction design, WizMap enables users to navigate and interpret embedding spaces with ease. Leveraging modern web technologies such as WebGL and Web Workers, WizMap scales to millions of embedding points directly in users' web browsers and computational notebooks without the need for dedicated backend servers. WizMap is open-source and available at the following public demo link: https://poloclub.github.io/wizmap.

[Demo] BiSync: A Bilingual Editor for Synchronized Monolingual Texts

François Yvon, Jitao Xu and Josep Crego   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

In our globalized world, a growing number of situations arise where people are required to communicate in one or several foreign languages. In the case of written communication, users with a good command of a foreign language may find assistance from computer-aided translation (CAT) technologies. These technologies often allow users to access external resources, such as dictionaries, terminologies or bilingual concordancers, thereby interrupting and considerably hindering the writing process. In addition, CAT systems assume that the source sentence is fixed and also restrict the possible changes on the target side. In order to make the writing process smoother, we present BiSync, a bilingual writing assistant that allows users to freely compose text in two languages, while maintaining the two monolingual texts synchronized. We also include additional functionalities, such as the display of alternative prefix translations and paraphrases, which are intended to facilitate the authoring of texts. We detail the model architecture used for synchronization and evaluate the resulting tool, showing that high accuracy can be attained with limited computational resources. The interface and models are publicly available at https://github.com/jmcrego/BiSync and a demonstration video can be watched on YouTube.

Distilling Script Knowledge from Large Language Models for Constrained Language Planning

Siyu Yuan, Jiangjie Chen, Ziquan Fu, Xuyang Ge, Soham Pranav Shah, Charles Jankowski, Yanghua Xiao and Deqing Yang   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

In everyday life, humans often plan their actions by following step-by-step instructions in the form of goal-oriented scripts. Previous work has exploited language models (LMs) to plan for abstract goals of stereotypical activities (e.g., "make a cake"), but leaves more specific goals with multi-facet constraints understudied (e.g., "make a cake for diabetics"). In this paper, we define the task of constrained language planning for the first time. We propose an over-generate-then-filter approach to improve large language models (LLMs) on this task, and use it to distill a novel constrained language planning dataset, Coscript, which consists of 55,000 scripts. Empirical results demonstrate that our method significantly improves the constrained language planning ability of LLMs, especially on constraint faithfulness. Furthermore, Coscript is demonstrated to be quite effective in endowing smaller LMs with constrained language planning ability.

HistRED: A Historical Document-Level Relation Extraction Dataset

Soyoung Yang, Minseok Choi, Youngwoo Cho and Jaegul Choo   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Despite the extensive applications of relation extraction (RE) tasks in various domains, little has been explored in the historical context, which contains promising data across hundreds and thousands of years. To promote the historical RE research, we present HistRED constructed from Yeonhaengnok. Yeonhaengnok is a collection of records originally written in Hanja, the classical Chinese writing, which has later been translated into Korean. HistRED provides bilingual annotations such that RE can be performed on Korean and Hanja texts. In addition, HistRED supports various self-contained subtexts with different lengths, from a sentence level to a document level, supporting diverse context settings for researchers to evaluate the robustness of their RE models. To demonstrate the usefulness of our dataset, we propose a bilingual RE model that leverages both Korean and Hanja contexts to predict relations between entities. Our model outperforms monolingual baselines on HistRED, showing that employing multiple language contexts supplements the RE predictions. The dataset is publicly available at: https://huggingface.co/datasets/Soyoung/HistRED under CC BY-NC-ND 4.0 license.

Facilitating Fine-grained Detection of Chinese Toxic Language: Hierarchical Taxonomy, Resources, and Benchmarks
Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang and Hongfei LIN   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The widespread dissemination of toxic online posts is increasingly damaging to society. However, research on detecting toxic language in Chinese has lagged significantly due to limited datasets. Existing datasets suffer from a lack of fine-grained annotations, such as the toxic type and expressions with indirect toxicity. These fine-grained annotations are crucial factors for accurately detecting the toxicity of posts involved with lexical knowledge, which has been a challenge for researchers. To tackle this problem, we facilitate the fine-grained detection of Chinese toxic language by building a new dataset with benchmark results. First, we devised Monitor Toxic Frame, a hierarchical taxonomy to analyze the toxic type and expressions. Then, we built a fine-grained dataset ToxiCN, including both direct and indirect toxic samples. ToxiCN is based on an insulting vocabulary containing implicit profanity. We further propose a benchmark model, Toxic Knowledge Enhancement (TKE), by incorporating lexical features to detect toxic language. We demonstrate the usability of ToxiCN and the effectiveness of TKE based on a systematic quantitative and qualitative analysis.

Hints on the data for language modeling of synthetic languages with transformers
Rodolfo Joel Zevallos and Nuria Bel   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Language Models (LM) are becoming more and more useful for providing representations upon which to train Natural Language Processing applications. However, there is now clear evidence that attention-based transformers require a critical amount of language data to produce good enough LMs. The question we have addressed in this paper is to what extent the critical amount of data varies for languages of different morphological typology, in particular those that have a rich inflectional morphology, and whether the tokenization method to preprocess the data can make a difference. These details can be important for low-resourced languages that need to plan the production of datasets. We evaluated intrinsically and extrinsically the differences of five different languages with different pretraining dataset sizes and three different tokenization methods for each. The results confirm that the size of the vocabulary due to morphological characteristics is directly correlated with both the LM perplexity and the performance of two typical downstream tasks such as NER identification and POS labeling. The experiments also provide new evidence that a canonical tokenizer can reduce perplexity by more than a half for a polysynthetic language like Quechua as well as raising F1 from 0.8 to more than 0.9 in both downstream tasks with a LM trained with only 6M tokens.

How to Distill your BERT: An Empirical Study on the Impact of Weight Initialisation and Distillation Objectives
Xinpeng Wang, Leonie Weissweiler, Hinrich Schütze and Barbara Plank   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Recently, various intermediate layer distillation (ILD) objectives have been shown to improve compression of BERT models via Knowledge Distillation (KD). However, a comprehensive evaluation of the objectives in both task-specific and task-agnostic settings is lacking. To the best of our knowledge, this is the first work comprehensively evaluating distillation objectives in both settings. We show that attention transfer gives the best performance overall. We also study the impact of layer choice when initializing the student from the teacher layers, finding a significant impact on the performance in task-specific distillation. For vanilla KD and hidden states transfer, initialisation with lower layers of the teacher gives a considerable improvement over higher layers, especially on the task of QNLI (up to an absolute percentage change of 17.8 in accuracy). Attention transfer behaves consistently under different initialisation settings. We release our code as an efficient transformer-based model distillation framework for further studies.

BUMP: A Benchmark of Unfaithful Minimal Pairs for Meta-Evaluation of Faithfulness Metrics
Liang Ma, Shuyang Cao, Robert L Logan IV, Di Lu, Shihao Ran, Ke Zhang, Joel Tetreault and Alejandro Jaimes   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The proliferation of automatic faithfulness metrics for summarization has produced a need for benchmarks to evaluate them. While existing benchmarks measure the correlation with human judgements of faithfulness on model-generated summaries, they are insufficient for diagnosing whether metrics are: 1) consistent, i.e., indicate lower faithfulness as errors are introduced into a summary, 2) effective on human-written texts, and 3) sensitive to different error types (as summaries can contain multiple errors). To address these needs, we present a benchmark of unfaithful minimal pairs (BUMP), a dataset of 889 human-written, minimally different summary pairs, where a single error is introduced to a summary from the CNN/DailyMail dataset to produce an unfaithful summary. We find BUMP complements existing benchmarks in a number of ways: 1) the summaries in BUMP are harder to discriminate and less probable than SOTA summarization models, 2) unlike non-pair-based datasets, BUMP can be used to measure the consistency of metrics, and reveals that the most discriminative metrics tend not to be the most consistent, and 3) unlike datasets containing generated summaries with multiple errors, BUMP enables the measurement of metrics' performance on individual error types.

FACTIFY-5WQA: 5W Aspect-based Fact Verification through Question Answering
Anku Rani, S.M Towhidul Islam Tonmoy, Dwip D Dalal, Shreya Gautam, Megha Chakraborty, Aman Chadha, Amit Sheth and Amitava Das   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Automatic fact verification has received significant attention recently. Contemporary automatic fact-checking systems focus on estimating truthfulness using numerical scores which are not human-interpretable. A human fact-checker generally follows several logical steps to verify a verisimilitude claim and conclude whether it's truthful or a mere masquerade. Popular fact-checking websites follow a common structure for fact categorization such as half true, half false, false, pants on fire, etc. Therefore, it is necessary to have an aspect-based (delineating which part(s) are true and which are false) explainable system that can assist human fact-checkers in asking relevant questions related to a fact, which can then be validated separately to reach a final verdict. In this paper, we propose a 5W framework (who, what, when, where, and why) for question-answer-based fact explainability. To that end, we present a semi-automatically generated dataset called FACTIFY-5WQA, which consists of 391, 041 facts along with relevant 5W QAs – underscoring our major contribution to this paper. A semantic role labeling system has been utilized to locate 5Ws, which generates QA pairs for claims using a masked language model. Finally, we report a baseline QA system to automatically locate those answers from evidence documents, which can serve as a baseline for future research in the field. Lastly, we propose a robust fact verification system that takes paraphrased claims and automatically validates them. The dataset and the baseline model are available at https: //github.com/ankuranii/acl-5W-QA

A Better Way to Do Masked Language Model Scoring
Carina Kauf and Anna Ivanova   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Estimating the log-likelihood of a given sentence under an autoregressive language model is straightforward: one can simply apply the chain rule and sum the log-likelihood values for each successive token. However, for masked language models (MLMs), there is no direct way to estimate the log-likelihood of a sentence. To address this issue, Salazar et al. (2020) propose to estimate sentence pseudo-log-likelihood (PLL) scores, computed by successively masking each sentence token, retrieving its score using the rest of the sentence as context, and summing the resulting values. Here, we demonstrate that the original PLL method yields inflated scores for out-of-vocabulary words and propose an adapted metric, in which we mask not only the target token, but also all within-word tokens to the right of the target. We show that our adapted metric (PLL-word-l2r) outperforms both the original PLL metric and a PLL metric in which all within-word tokens are masked. In particular, it better satisfies theoretical desiderata and better correlates with scores from autoregressive models. Finally, we show that the choice of metric affects even tightly controlled, minimal pair evaluation benchmarks (such as BLiMP), underscoring the importance of selecting an appropriate scoring metric for evaluating MLM properties.

Automatic Annotation of Direct Speech in Written French Narratives
Noé Durandard, Viet Anh TRAN, Gaspard Michel and Elena V. Epure   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The automatic annotation of direct speech (AADS) in written text has been often used in computational narrative understanding. Methods based on either rules or deep neural networks have been explored, in particular for English or German languages. Yet, for French, our target language, not many works exist. Our goal is to create a unified framework to design and evaluate AADS models in French. For this, we consolidated the largest-to-date French narrative dataset annotated with DS per word; we adapted various baselines for sequence labelling or from AADS in other languages; and we designed and conducted an extensive evaluation focused on generalisation. Results show that the task still requires substantial efforts and emphasise characteristics of each baseline. Although this framework could be improved, it is a step further to encourage more research on the topic.

On the Evaluation of Neural Selective Prediction Methods for Natural Language Processing
Zhengyao Gu and Mark Hopkins   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
We provide a survey and empirical comparison of the state-of-the-art in neural selective classification for NLP tasks. We also provide a methodological blueprint, including a novel metric called refinement that provides a calibrated evaluation of confidence functions for selective prediction. Finally, we supply documented, open-source code to support the future development of selective prediction techniques.

A Textual Dataset for Situated Proactive Response Selection
Naoki Otani, Jun Araki, HyeongSik Kim and Eduard H Hovy   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Recent data-driven conversational models are able to return fluent, consistent, and informative responses to many kinds of requests and utterances in task-oriented scenarios. However, these responses are typically limited to just the immediate local topic instead of being wider-ranging and proactively taking the conversation further, for example making suggestions to help customers achieve their goals. This inadequacy reflects a lack of understanding of the interlocutor's situation and implicit goal. To address the problem, we introduce a task of proactive response selection based on situational information. We present a manually-curated dataset of 1.7k English conversation examples that include situational background information plus for each conversation a set of responses, only some of which are acceptable in the situation. A responsive and informed conversation system should select the appropriate responses and avoid inappropriate ones; doing so demonstrates the ability to adequately understand the initiating request and situation. Our benchmark experiments show that this is not an easy task even for strong neural models, offering opportunities for future research.

Towards Open-World Product Attribute Mining: A Lightly-Supervised Approach
Liyan Xu, Chenwei Zhang, Xian Li, Jingbo Shang and Jinho D. Choi   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
We present a new task setting for attribute mining on e-commerce products, serving as a practical solution to extract open-world attributes without extensive human intervention. Our supervision comes from a high-quality seed attribute set bootstrapped from existing resources, and we aim to expand the attribute vocabulary of existing seed types, and also to discover any new attribute types automatically. A new dataset is created to support our setting, and our approach Amacer is proposed specifically to tackle the limited supervision. Especially, given that no direct supervision is available for those unseen new attributes, our novel formulation exploits self-supervised heuristic and unsupervised latent attributes, which attains implicit semantic signals as additional supervision by leveraging product context. Experiments suggest that our approach surpasses various baselines by 12 F1, expanding attributes of existing types significantly by up to 12 times, and discovering values from 39% new types.

Typo-Robust Representation Learning for Dense Retrieval
Panuthep Tasawong, Wuttikorn Ponwitayarat, Peerat Limkonchotiwat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich and Sarana Nutanong   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Dense retrieval is a basic building block of information retrieval applications. One of the main challenges of dense retrieval in real-world settings is the handling of queries containing misspelled words. A popular approach for handling misspelled queries is minimizing the representations discrepancy between misspelled queries and their pristine ones. Unlike the existing approaches, which only focus on the alignment between misspelled and pristine queries, our method also improves the contrast between each misspelled query and its surrounding queries. To assess the effectiveness of our proposed method, we compare it against the existing competitors using two benchmark datasets and two base encoders. Our method outperforms the competitors in all cases with misspelled queries. Our code and models are available at https://github.com/panuthept/DST-DenseRetrieval.

MUSTIE: Multimodal Structural Transformer for Web Information Extraction
Qifan Wang, Jingang Wang, Xiaojun Quan, Fuli Feng, Zenglin Xu, Shaoliang Nie, Sinong Wang, Madian Khabsa, Hamed Firooz and Dongfang Liu   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The task of web information extraction is to extract target fields of an object from web pages, such as extracting the name, genre and actor from a movie page. Recent sequential modeling approaches have achieved state-of-the-art results on web information extraction. However, most of these methods only focus on extracting information from textual sources while ignoring the rich information from other modalities such as image and web layout. In this work, we propose a novel MUltimodal Structural Transformer (MUST) that incorporates multiple modalities for web information extraction. Concretely, we develop a structural encoder that jointly encodes the multimodal information based on the HTML structure of the web layout, where high-level DOM nodes, and low-level text and image tokens are introduced to represent the entire page. Structural attention patterns are designed to learn effective cross-modal embeddings for all DOM nodes and low-level tokens. An extensive set of experiments are conducted on WebSRC and Common Crawl benchmarks. Experimental results demonstrate the superior performance of MUST over several state-of-the-art baselines.

Faking Fake News for Real Fake News Detection: Propaganda-Loaded Training Data Generation
Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi and Heng Ji   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Despite recent advances in detecting fake news generated by neural models, their results are not readily applicable to effective detection of human-written disinformation. What limits the successful transfer between them is the sizable gap between machine-generated fake news and human-authored ones, including the notable differences in terms of style and underlying intent. With this in mind, we propose a novel framework for generating training examples that are informed by the known styles and strategies of human-authored propaganda. Specifically, we perform self-critical sequence training guided by natural language inference to ensure the validity of the generated articles, while also incorporating propaganda techniques, such as appeal to authority and loaded language. In particular, we create a new training dataset, PropaNews, with 2,256 examples, which we release for future use. Our experimental results show that fake news detectors trained on PropaNews are better at detecting human-written disinformation by 3.62–7.69% F1 score on two public datasets.

When Does Aggregating Multiple Skills with Multi-Task Learning Work? A Case Study in Financial NLP
Jingwei Ni, Zhijing Jin, QIAN WANG, Mrinmaya Sachan and Markus Leippold   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Multi-task learning (MTL) aims at achieving a better model by leveraging data and knowledge from multiple tasks. However, MTL does not always work – sometimes negative transfer occurs between tasks, especially when aggregating loosely related skills, leaving it an open question when MTL works. Previous studies show that MTL performance can be improved by algorithmic tricks. However, what tasks and

skills should be included is less well explored. In this work, we conduct a case study in Financial NLP where multiple datasets exist for skills relevant to the domain, such as numeric reasoning and sentiment analysis. Due to the task difficulty and data scarcity in the Financial NLP domain, we explore when aggregating such diverse skills from multiple datasets with MTL can work. Our findings suggest that the key to MTL success lies in skill diversity, relatedness between tasks, and choice of aggregation size and shared capacity. Specifically, MTL works well when tasks are diverse but related, and when the size of the task aggregation and the shared capacity of the model are balanced to avoid overwhelming certain tasks.

Rule By Example: Harnessing Logical Rules for Explainable Hate Speech Detection

Christopher Clarke, Matthew Hall, Gaurav Mittal, Ye Yu, Sandra Sajeev, Jason Mars and Mei Chen 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Classic approaches to content moderation typically apply a rule-based heuristic approach to flag content. While rules are easily customizable and intuitive for humans to interpret, they are inherently fragile and lack the flexibility or robustness needed to moderate the vast amount of undesirable content found online today. Recent advances in deep learning have demonstrated the promise of using highly effective deep neural models to overcome these challenges. However, despite the improved performance, these data-driven models lack transparency and explainability, often leading to mistrust from everyday users and a lack of adoption by many platforms. In this paper, we present Rule By Example (RBE): a novel exemplar-based contrastive learning approach for learning from logical rules for the task of textual content moderation. RBE is capable of providing rule-grounded predictions, allowing for more explainable and customizable predictions compared to typical deep learning-based approaches. We demonstrate that our approach is capable of learning rich rule embedding representations using only a few data examples. Experimental results on 3 popular hate speech classification datasets show that RBE is able to outperform state-of-the-art deep learning classifiers as well as the use of rules in both supervised and unsupervised settings while providing explainable model predictions via rule-grounding.

Fact-Checking Complex Claims with Program-Guided Reasoning

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan and Preslav Nakov 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Fact-checking real-world claims often requires collecting multiple pieces of evidence and applying complex multi-step reasoning. In this paper, we present Program-Guided Fact-Checking (ProgramFC), a novel fact-checking model that decomposes complex claims into simpler sub-tasks that can be solved using a shared library of specialized functions. We first leverage the in-context learning ability of large language models to generate reasoning programs to guide the verification process. Afterward, we execute the program by delegating each sub-task to the corresponding sub-task handler. This process makes our model both explanatory and data-efficient, providing clear explanations of its reasoning process and requiring minimal training data. We evaluate ProgramFC on two challenging fact-checking datasets and show that it outperforms seven fact-checking baselines across different settings of evidence availability, with explicit output programs that benefit human debugging. Our codes and data are publicly available at https://github.com/mbzuai-nlp/ProgramFC.

Unsupervised Subtitle Segmentation with Masked Language Models

David Ponce, Thierry Etchegoyhen and Victor Ruiz 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

We describe a novel unsupervised approach to subtitle segmentation, based on pretrained masked language models, where line endings and subtitle breaks are predicted according to the likelihood of punctuation to occur at candidate segmentation points. Our approach obtained competitive results in terms of segmentation accuracy across metrics, while also fully preserving the original text and complying with length constraints. Although supervised models trained on in-domain data and with access to source audio information can provide better segmentation accuracy, our approach is highly portable across languages and domains and may constitute a robust off-the-shelf solution for subtitle segmentation.

A Cognitive Stimulation Dialogue System with Multi-source Knowledge Fusion for Elders with Cognitive Impairment

Jiyue Jiang, Sheng Wang, Qintong Li, Lingpeng Kong and Chuan Wu 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

When communicating with elders with cognitive impairment, cognitive stimulation (CS) help to maintain the cognitive health of elders. Data sparsity is the main challenge in building CS-based dialogue systems, particularly in the Chinese language. To fill this gap, we construct a Chinese CS conversation (CSConv) dataset, which contains about 2.6K groups of dialogues with therapy principles and emotional support strategy labels. Making chit chat while providing emotional support is overlooked by the majority of existing cognitive dialogue systems. In this paper, we propose a multi-source knowledge fusion method for CS dialogue (CSD), to generate open-ended responses guided by the therapy principle and emotional support strategy. We first use a progressive mask method based on external knowledge to learn encoders as effective classifiers, which is the prerequisite to predict the therapy principle and emotional support strategy of the target response. Then a decoder interacts with the perceived therapy principle and emotional support strategy to generate responses. Extensive experiments conducted on the CSConv dataset demonstrate the effectiveness of the proposed method, while there is still a large space for improvement compared to human performance.

ACTC: Active Threshold Calibration for Cold-Start Knowledge Graph Completion

Anastasiia Sedova and Benjamin Roth 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Self-supervised knowledge-graph completion (KGC) relies on estimating a scoring model over (entity, relation, entity)-tuples, for example, by embedding an initial knowledge graph. Prediction quality can be improved by calibrating the scoring model, typically by adjusting the prediction thresholds using manually annotated examples. In this paper, we attempt for the first time cold-start calibration for KGC, where no annotated examples exist initially for calibration, and only a limited number of tuples can be selected for annotation. Our new method ACTC finds good per-relation thresholds efficiently based on a limited set of annotated tuples. Additionally to a few annotated tuples, ACTC also leverages unlabeled tuples by estimating their correctness with Logistic Regression or Gaussian Process classifiers. We also experiment with different methods for selecting candidate tuples for annotation: density-based and random selection. Experiments with five scoring models and an oracle annotator show an improvement of 7% points when using ACTC in the challenging setting with an annotation budget of only 10 tuples, and an average improvement of 4% points over different budgets.

Controllable Mixed-Initiative Dialogue Generation through Prompting

Maximillian Chen, Xiao Yu, Weiyan Shi, Urvi Awasthi and Zhou Yu 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Mixed-initiative dialogue tasks involve repeated exchanges of information and conversational control. Conversational agents gain control by generating responses that follow particular dialogue intents or strategies, prescribed by a policy planner. The standard approach has been fine-tuning pre-trained language models to perform generation conditioned on these intents. However, these supervised generation models are limited by the cost and quality of data annotation. We instead prompt large language models as a drop-in replacement to fine-tuning on conditional generation. We formalize prompt construction for controllable mixed-initiative dialogue. Our findings show improvements over fine-tuning and ground truth responses according to human evaluation and automatic metrics for two tasks: PersuasionForGood and Emotional Support Conversations.

ClarifyDelphi: Reinforced Clarification Questions with Defeasibility Rewards for Social and Moral Situations

Valentina Pyatkin, Jena D. Hwang, Vivek Srikumar, Ximing Lu, Liwei Jiang, Yejin Choi and Chandra Bhagavatula   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Context is everything, even in commonsense moral reasoning. Changing contexts can flip the moral judgment of an action; Lying to a friend is wrong in general, but may be morally acceptable if it is intended to protect their life.
We present ClarifyDelphi, an interactive system that learns to ask clarification questions (e.g., why did you lie to your friend?) in order to elicit additional salient contexts of a social or moral situation. We posit that questions whose potential answers lead to *diverging* moral judgments are the most informative. Thus, we propose a reinforcement learning framework with a defeasibility reward that aims to maximize the divergence between moral judgments of hypothetical answers to a question. Human evaluation demonstrates that our system generates more relevant, informative and defeasible questions compared to competitive baselines. Our work is ultimately inspired by studies in cognitive science that have investigated the flexibility in moral cognition (i.e., the diverse contexts in which moral rules can be bent), and we hope that research in this direction can assist both cognitive and computational investigations of moral judgments.

Privacy-Preserving Domain Adaptation of Semantic Parsers
Fatemehsadat Mireshghallah, Yu Su, Tatsunori Hashimoto, Jason Eisner and Richard Shin   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Task-oriented dialogue systems often assist users with personal or confidential matters. For this reason, the developers of such a system are generally prohibited from observing actual usage. So how can they know where the system is failing and needs more training data or new functionality? In this work, we study ways in which realistic user utterances can be generated synthetically, to help increase the linguistic and functional coverage of the system, without compromising the privacy of actual users. To this end, we propose a two-stage Differentially Private (DP) generation method which first generates latent semantic parses, and then generates utterances based on the parses. Our proposed approach improves MAUVE by 2.5X and parse tree function-type overlap by 1.3X relative to current approaches for private synthetic data generation, improving both on fluency and semantic coverage. We further validate our approach on a realistic domain adaptation task of adding new functionality from private user data to a semantic parser, and show overall gains of 8.5% points on its accuracy with the new feature.

ChatGPT for Zero-shot Dialogue State Tracking: A Solution or an Opportunity?
Michael Heck, Nurul Lubis, Benjamin Matthias Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk and Milica Gasic   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Recent research on dialog state tracking (DST) focuses on methods that allow few- and zero-shot transfer to new domains or schemas. However, performance gains heavily depend on aggressive data augmentation and fine-tuning of ever larger language model based architectures. In contrast, general purpose language models, trained on large amounts of diverse data, hold the promise of solving any kind of task without task-specific training. We present preliminary experimental results on the ChatGPT research preview, showing that ChatGPT achieves state-of-the-art performance in zero-shot DST. Despite our findings, we argue that properties inherent to general purpose models limit their ability to replace specialized systems. We further theorize that the in-context learning capabilities of such models will likely become powerful tools to support the development of dedicated dialog state trackers and enable dynamic methods.

PeaCoK: Persona Commonsense Knowledge for Consistent and Engaging Narratives
Silin Gao, Beatriz Borges, Soyoung Oh, Deniz Bayazit, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji and Antoine Bosselut   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Sustaining coherent and engaging narratives requires dialogue or storytelling agents to understand how the personas of speakers or listeners ground the narrative. Specifically, these agents must infer personas of their listeners to produce statements that cater to their interests. They must also learn to maintain consistent speaker personas for themselves throughout the narrative, so that their counterparts feel involved in a realistic conversation or story.
However, personas are diverse and complex: they entail large quantities of rich interconnected world knowledge that is challenging to robustly represent in general narrative systems (e.g., a singer is good at singing, and may have attended conservatoire). In this work, we construct a new large-scale persona commonsense knowledge graph, PeaCoK, containing 100K human-validated persona facts. Our knowledge graph schematizes five dimensions of persona knowledge identified in previous studies of human interactive behaviours, and distils facts in this schema from both existing commonsense knowledge graphs and large-scale pretrained language models. Our analysis indicates that PeaCoK contains rich and precise world persona inferences that help downstream systems generate more consistent and engaging narratives.

BREAK: Breaking the Dialogue State Tracking Barrier with Beam Search and Re-ranking
Seungpil Won, Heeyoung Kwak, Joongbo Shin, Janghoon Han and Kyomin Jung   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Despite the recent advances in dialogue state tracking (DST), the joint goal accuracy (JGA) of the existing methods on MultiWOZ 2.1 still remains merely 60%. In our preliminary error analysis, we find that beam search produces a pool of candidates that is likely to include the correct dialogue state. Motivated by this observation, we introduce a novel framework, called BREAK (Beam search and RE-rAnKing), that achieves outstanding performance on DST. BREAK performs DST in two stages: (i) generating k-best dialogue state candidates with beam search and (ii) re-ranking the candidates to select the correct dialogue state. This simple yet powerful framework shows state-of-the-art performance on all versions of MultiWOZ and M2M datasets. Most notably, we push the joint goal accuracy to 80-90% on MultiWOZ 2.1-2.4, which is an improvement of 23.6%, 26.3%, 21.7%, and 10.8% over the previous best-performing models, respectively. The data and code will be available at https://github.com/tony-won/DST-BREAK

Learning to Generate Equitable Text in Dialogue from Biased Training Data
Anthony B Sicilia and Malihe Alikhani   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The ingrained principles of fairness in a dialogue system's decision-making process and generated responses are crucial for user engagement, satisfaction, and task achievement. Absence of equitable and inclusive principles can hinder the formation of common ground, which in turn negatively impacts the overall performance of the system. For example, misusing pronouns in a user interaction may cause ambiguity about the intended subject. Yet, there is no comprehensive study of equitable text generation in dialogue. Aptly, in this work, we use theories of computational learning to study this problem. We provide formal definitions of equity in text generation, and further, prove formal connections between learning human-likeness and learning equity: algorithms for improving equity ultimately reduce to algorithms for improving human-likeness (on augmented data). With this insight, we also formulate reasonable conditions under which text generation algorithms can learn to generate equitable text without any modifications to the biased training data on which they learn. To exemplify our theory in practice, we look at a group of algorithms for the GuessWhat?! visual dialogue game and, using this example, test our theory empirically. Our theory accurately predicts relative-performance of multiple algorithms in generating equitable text as measured by both human and automated evaluation.

Diverse Demonstrations Improve In-context Compositional Generalization
Itay Levy, Ben Bogin and Jonathan Berant   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
In-context learning has shown great success in i.i.d semantic parsing splits, where the training and test sets are drawn from the same distribution. In this setup, models are typically prompted with demonstrations that are similar to the input utterance. However, in the setup

of compositional generalization, where models are tested on outputs with structures that are absent from the training set, selecting similar demonstrations is insufficient, as often no example will be similar enough to the input. In this work, we propose a method to select diverse demonstrations that aims to collectively cover all of the structures required in the output program, in order to encourage the model to generalize to new structures from these demonstrations. We empirically show that combining diverse demonstrations with in-context learning substantially improves performance across three compositional generalization semantic parsing datasets in the pure in-context learning setup and when combined with finetuning.

Context-Aware Transformer Pre-Training for Answer Sentence Selection
Luca Di Liello, Siddhant Garg and Alessandro Moschitti   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Answer Sentence Selection (AS2) is a core component for building an accurate Question Answering pipeline. AS2 models rank a set of candidate sentences based on how likely they answer a given question. The state of the art in AS2 exploits pre-trained transformers by transferring them on large annotated datasets, while using local contextual information around the candidate sentence. In this paper, we propose three pre-training objectives designed to mimic the downstream fine-tuning task of contextual AS2. This allows for specializing LMs when fine-tuning for contextual AS2. Our experiments on three public and two large-scale industrial datasets show that our pre-training approaches (applied to RoBERTa and ELECTRA) can improve baseline contextual AS2 accuracy by up to 8% on some datasets.

Cross-Domain Data Augmentation with Domain-Adaptive Language Modeling for Aspect-Based Sentiment Analysis
Jianfei Yu, Qiankun Zhao and Rui Xia   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Cross-domain Aspect-Based Sentiment Analysis (ABSA) aims to leverage the useful knowledge from a source domain to identify aspect-sentiment pairs in sentences from a target domain. To tackle the task, several recent works explore a new unsupervised domain adaptation framework, i.e., Cross-Domain Data Augmentation (CDDA), aiming to directly generate much labeled target-domain data based on the labeled source-domain data. However, these CDDA methods still suffer from several issues: 1) preserving many source-specific attributes such as syntactic structures; 2) lack of fluency and coherence; 3) limiting the diversity of generated data. To address these issues, we propose a new cross-domain Data Augmentation approach based on Domain-Adaptive Language Modeling named DA^2LM, which contains three stages: 1) assigning pseudo labels to unlabeled target-domain data; 2) unifying the process of token generation and labeling with a Domain-Adaptive Language Model (DALM) to learn the shared context and annotation across domains; 3) using the trained DALM to generate labeled target-domain data. Experiments show that DA^2LM consistently outperforms previous feature adaptation and CDDA methods on both ABSA and Aspect Extraction tasks. The source code is publicly released at https://github.com/NUSTM/DALM.

Do You Hear The People Sing? Key Point Analysis via Iterative Clustering and Abstractive Summarisation
Hao Li, Viktor Schlegel, Riza Batista-Navarro and Goran Nenadic   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Argument summarisation is a promising but currently under-explored field. Recent work has aimed to provide textual summaries in the form of concise and salient short texts, i.e., key points (KPs), in a task known as Key Point Analysis (KPA). One of the main challenges in KPA is finding high-quality key point candidates from dozens of arguments even in a small corpus. Furthermore, evaluating key points is crucial in ensuring that the automatically generated summaries are useful. Although automatic methods for evaluating summarisation have considerably advanced over the years, they mainly focus on sentence-level comparison, making it difficult to measure the quality of a summary (a set of KPs) as a whole. Aggravating this problem is the fact that human evaluation is costly and unreproducible. To address the above issues, we propose a two-step abstractive summarisation framework based on neural topic modelling with an iterative clustering procedure, to generate key points which are aligned with how humans identify key points. Our experiments show that our framework advances the state of the art in KPA, with performance improvement of up to 14 (absolute) percentage points, in terms of both ROUGE and our own proposed evaluation metrics. Furthermore, we evaluate the generated summaries using a novel set-based evaluation toolkit. Our quantitative analysis demonstrates the effectiveness of our proposed evaluation metrics in assessing the quality of generated KPs. Human evaluation further demonstrates the advantages of our approach and validates that our proposed evaluation metric is more consistent with human judgment than ROUGE scores.

Are Fairy Tales Fair? Analyzing Gender Bias in Temporal Narrative Event Chains of Children's Fairy Tales
Paulina Toro Isaza, Guangxuan Xu, Toye Oloko, Yufang Hou, Nanyun Peng and Dakuo Wang 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Social biases and stereotypes are embedded in our culture in part through their presence in our stories, as evidenced by the rich history of humanities and social science literature analyzing such biases in children stories. Because these analyses are often conducted manually and at a small scale, such investigations can benefit from the use of more recent natural language processing (NLP) methods that examine social bias in models and data corpora. Our work joins this interdisciplinary effort and makes a unique contribution by taking into account the event narrative structures when analyzing the social bias of stories. We propose a computational pipeline that automatically extracts a story's temporal narrative verb-based event chain for each of its characters as well as character attributes such as gender. We also present a verb-based event annotation scheme that can facilitate bias analysis by including categories such as those that align with traditional stereotypes. Through a case study analyzing gender bias in fairy tales, we demonstrate that our framework can reveal bias in not only the unigram verb-based events in which female and male characters participate but also in the temporal narrative order of such event participation.

Helping a Friend or Supporting a Cause? Disentangling Active and Passive Cosponsorship in the U.S. Congress
Giuseppe Russo, Christoph Gote, Laurence Brandenberger, Sophia Johanna Schlosser and Frank Schweitzer   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
In the U.S. Congress, legislators can use active and passive cosponsorship to support bills. We show that these two types of cosponsorship are driven by two different motivations: the backing of political colleagues and the backing of the bill's content. To this end, we develop an Encoder+RGCN based model that learns legislator representations from bill texts and speech transcripts. These representations predict active and passive cosponsorship with an F1-score of 0.88. Applying our representations to predict voting decisions, we show that they are interpretable and generalize to unseen tasks.

Class based Influence Functions for Error Detection
Thang Nguyen-Duc, Hoang Thanh-Tung, Quan Hung Tran, Dang Huu-Tien, Hieu Ngoc Nguyen, Anh T. V. Dau and Nghi D. Q. Bui   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Influence functions (IFs) are a powerful tool for detecting anomalous examples in large scale datasets. However, they are unstable when applied to deep networks. In this paper, we provide an explanation for the instability of IFs and develop a solution to this problem. We show that IFs are unreliable when the two data points belong to two different classes. Our solution leverages class information to improve the stability of IFs. Extensive experiments show that our modification significantly improves the performance and stability of IFs while incurring no additional computational cost.

Syntax and Geometry of Information
Raphaël Bailly, Laurent Leblond and Kata Gábor   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
This paper presents an information-theoretical model of syntactic generalization. We study syntactic generalization from the perspective

of the capacity to disentangle semantic and structural information, emulating the human capacity to assign a grammaticality judgment to semantically nonsensical sentences. In order to isolate the structure, we propose to represent the probability distribution behind a corpus as the product of the probability of a semantic context and the probability of a structure, the latter being independent of the former. We further elaborate the notion of abstraction as a relaxation of the property of independence. It is based on the measure of structural and contextual information for a given representation. We test abstraction as an optimization objective on the task of inducing syntactic categories from natural language data and show that it significantly outperforms alternative methods. Furthermore, we find that when syntax-unaware optimization objectives succeed in the task, their success is mainly due to an implicit disentanglement process rather than to the model structure. On the other hand, syntactic categories can be deduced in a principled way from the independence between structure and context.

FiD-ICL: A Fusion-in-Decoder Approach for Efficient In-Context Learning
Qinyuan Ye, Iz Beltagy, Matthew Peters, Xiang Ren and Hannaneh Hajishirzi   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Large pre-trained models are capable of few-shot in-context learning (ICL), i.e., performing a new task by prepending a few demonstrations before the test input. However, the concatenated demonstrations are often excessively long and induce additional computation. Inspired by fusion-in-decoder (FiD) models which efficiently aggregate more passages and thus outperforms concatenation-based models in open-domain QA, we hypothesize that similar techniques can be applied to improve the efficiency and end-task performance of ICL. To verify this, we present a comprehensive study on applying three fusion methods—concatenation-based (early fusion), FiD (intermediate), and ensemble-based (late)—to ICL. We adopt a meta-learning setup where a model is first trained to perform ICL on a mixture of tasks using one selected fusion method, then evaluated on held-out tasks for ICL. Results on 11 held-out tasks show that FiD-ICL matches or outperforms the other two fusion methods. Additionally, we show that FiD-ICL (1) is 10x faster at inference time compared to concat-based and ensemble-based ICL, as we can easily pre-compute the representations of in-context examples and reuse them; (2) enables scaling up to meta-training 3B-sized models, which would fail for concat-based ICL.

Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models
Junmo Kang, Wei Xu and Alan Ritter   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Fine-tuning large models is highly effective, however, inference can be expensive and produces carbon emissions. Knowledge distillation has been shown to be a practical solution to reduce inference costs, but the distillation process itself requires significant computational resources. Rather than buying or renting GPUs to fine-tune, then distill a large model, an NLP practitioner might instead choose to allocate the available budget to hire annotators and manually label additional fine-tuning data. In this paper, we investigate how to most efficiently use a fixed budget to build a compact model. Through extensive experiments on six diverse tasks, we show that distilling from T5-XXL (11B) to T5-Small (60M) is almost always a cost-efficient strategy compared to annotating more data to directly train a compact model (T5-Small). We further investigate how the optimal budget allocated towards computation varies across scenarios. We will make our code, datasets, annotation cost estimates, and baseline models available as a benchmark to support further work on cost-efficient training of compact models.

Learning Symbolic Rules over Abstract Meaning Representations for Textual Reinforcement Learning
Subhajit Chaudhury, Sarathkrishna Swaminathan, Daiki Kimura, Prithviraj Sen, Keerthiram Murugesan, Rosario Uceda-Sosa, Michiaki Tatsubori, Achille Fokoue, Pavan Kapanipathi, Asim Munawar and Alexander Gray   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Text-based reinforcement learning agents have predominantly been neural network-based models with embeddings-based representation, learning uninterpretable policies that often do not generalize well to unseen games. On the other hand, neuro-symbolic methods, specifically those that leverage an intermediate formal representation, are gaining significant attention in language understanding tasks. This is because of their advantages ranging from inherent interpretability, the lesser requirement of training data, and being generalizable in scenarios with unseen data. Therefore, in this paper, we propose a modular, NEuro-Symbolic Textual Agent (NESTA) that combines a generic semantic parser with a rule induction system to learn abstract interpretable rules as policies. Our experiments on established text-based game benchmarks show that the proposed NESTA method outperforms deep reinforcement learning-based techniques by achieving better generalization to unseen test games and learning from fewer training interactions.

LM-CPPF: Paraphrasing-Guided Data Augmentation for Contrastive Prompt-Based Few-Shot Fine-Tuning
Amirhossein Abaskohi, Sascha Rothe and Yadollah Yaghoobzadeh   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
In recent years, there has been significant progress in developing pre-trained language models for NLP. However, these models often struggle when fine-tuned on small datasets. To address this issue, researchers have proposed various adaptation approaches. Prompt-based tuning is arguably the most common way, especially for larger models. Previous research shows that adding contrastive learning to prompt-based fine-tuning is effective as it helps the model generate embeddings that are more distinguishable between classes, and it can also be more sample-efficient as the model learns from positive and negative examples simultaneously. One of the most important components of contrastive learning is data augmentation, but unlike computer vision, effective data augmentation for NLP is still challenging. This paper proposes LM-CPPF, Contrastive Paraphrasing-guided Prompt-based Fine-tuning of Language Models, which leverages prompt-based few-shot paraphrasing using generative language models, especially large language models such as GPT-3 and OPT-175B, for data augmentation. Our experiments on multiple text classification benchmarks show that this augmentation method outperforms other methods, such as easy data augmentation, back translation, and multiple templates.

A Measure-Theoretic Characterization of Tight Language Models
Li Du, Lucas Torroba Hennigen, Tiago Pimentel, Clara Meister, Jason Eisner and Ryan Cotterell   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Language modeling, a central task in natural language processing, involves estimating a probability distribution over strings. In most cases, the estimated distribution sums to 1 over all finite strings. However, in some pathological cases, probability mass can "leak" onto the set of infinite sequences. In order to characterize the notion of leakage more precisely, this paper offers a measure-theoretic treatment of language modeling. We prove that many popular language model families are in fact tight, meaning that they will not leak in this sense. We also generalize characterizations of tightness proposed in previous works.

Parallel Context Windows for Large Language Models
Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Dov Karpas, Amnon Shashua, Kevin Leyton-Brown and Yoav Shoham   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
When applied to processing long text, Large Language Models (LLMs) are limited by their context window. Existing efforts to address this limitation involve training specialized architectures, and cannot be easily applied to off- the-shelf LLMs. We present Parallel Context Windows (PCW), a method that alleviates the context window restriction for any off-the-shelf LLM without further training. The key to the approach is to carve a long context into chunks ("windows"), restrict the attention mechanism to apply only within each window, and re-use the positional embeddings across the windows. Our main results test the PCW approach on in-context learning with models that range in size from 750 million and 178 billion parameters, and show substantial improvements for tasks with diverse input and output spaces. We show additional benefits in other settings where long context windows may be beneficial: multi-hop questions and retrieval-augmented question answering with multiple retrieved documents. Our results highlight Parallel Context Windows as a promising method for applying off-the-shelf LLMs in a range of settings that require long text sequences. We make our code publicly available at https://github.com/

ai21labs/parallel-context-windows.

Two-Stage Fine-Tuning for Improved Bias and Variance for Large Pretrained Language Models

Lijing Wang, Yingya Li, Timothy A Miller, Steven Bethard and Guergana Savova   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The bias-variance tradeoff is the idea that learning methods need to balance model complexity with data size to minimize both under-fitting and over-fitting. Recent empirical work and theoretical analysis with over-parameterized neural networks challenges the classic bias-variance trade-off notion suggesting that no such trade-off holds: as the width of the network grows, bias monotonically decreases while variance initially increases followed by a decrease. In this work, we first provide a variance decomposition-based justification criteria to examine whether large pretrained neural models in a fine-tuning setting are generalizable enough to have low bias and variance. We then perform theoretical and empirical analysis using ensemble methods explicitly designed to decrease variance due to optimization. This results in essentially a two-stage fine-tuning algorithm that first ratchets down bias and variance iteratively, and then uses a selected fixed-bias model to further reduce variance due to optimization by ensembling. We also analyze the nature of variance change with the ensemble size in low- and high-resource classes. Empirical results show that this two-stage method obtains strong results on SuperGLUE tasks and clinical information extraction tasks. Code and settings are available: https://github.com/christa60/bias-var-fine-tuning-plms.git

Self-Instruct: Aligning Language Models with Self-Generated Instructions

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi and Hannaneh Hajishirzi   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Large "instruction-tuned" language models (i.e., finetuned to respond to instructions) have demonstrated a remarkable ability to generalize zero-shot to new tasks. Nevertheless, they depend heavily on human-written instruction data that is often limited in quantity, diversity, and creativity, therefore hindering the generality of the tuned model. We introduce Self-Instruct, a framework for improving the instruction-following capabilities of pretrained language models by bootstrapping off their own generations. Our pipeline generates instructions, input, and output samples from a language model, then filters invalid or similar ones before using them to finetune the original model. Applying our method to the vanilla GPT3, we demonstrate a 33% absolute improvement over the original model on Super-NaturalInstructions, on par with the performance of InstructGPT-001, which was trained with private user data and human annotations. For further evaluation, we curate a set of expert-written instructions for novel tasks, and show through human evaluation that tuning GPT3 with Self-Instruct outperforms using existing public instruction datasets by a large margin, leaving only a 5% absolute gap behind InstructGPT-001. Self-Instruct provides an almost annotation-free method for aligning pre-trained language models with instructions, and we release our large synthetic dataset to facilitate future studies on instruction tuning.

Knowledge Unlearning for Mitigating Privacy Risks in Language Models

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran and Minjoon Seo   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Pretrained Language Models (LMs) memorize a vast amount of knowledge during initial pretraining, including information that may violate the privacy of personal lives and identities. Previous work addressing privacy issues for LMs has mostly focused on data preprocessing and differential privacy methods, both requiring re-training the underlying LM. We propose knowledge unlearning as an alternative method to reduce privacy risks for LMs post hoc. We show that simply performing gradient ascent on target token sequences is effective at forgetting them with little to no degradation of general language modeling performances for larger-sized LMs. We also find that sequential unlearning is better than trying to unlearn all the data at once and that unlearning is highly dependent on which kind of data (domain) is forgotten. By showing comparisons with previous methods known to mitigate privacy risks for LMs, we show that our approach can give a stronger empirical privacy guarantee in scenarios where the data vulnerable to extraction attacks are known a priori while being much more efficient and robust.

Data Curation Alone Can Stabilize In-context Learning

Ting-Yun Chang and Robin Jia   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
In-context learning (ICL) enables large language models (LLMs) to perform new tasks by prompting them with a sequence of training examples. However, it is known that ICL is very sensitive to the choice of training examples: randomly sampling examples from a training set leads to high variance in performance. In this paper, we show that carefully curating a subset of training data greatly stabilizes ICL performance without any other changes to the ICL algorithm (e.g., prompt retrieval or calibration). We introduce two methods to choose training subsets—both score training examples individually, then select the highest-scoring ones. CondAcc scores a training example by its average dev-set ICL accuracy when combined with random training examples, while Datamodels learns linear regressors that estimate how the presence of each training example influences LLM outputs. Across five tasks and two LLMs, sampling from stable subsets selected by CondAcc and Datamodels improves average accuracy over sampling from the entire training set by 7.7% and 6.3%, respectively. Surprisingly, the stable subset examples are not especially diverse in content or low in perplexity, in contrast with other work suggesting that diversity and perplexity are important when prompting LLMs.

A Systematic Study of Knowledge Distillation for Natural Language Generation with Pseudo-Target Training

Nitay Calderon, Subhabrata Mukherjee, Roi Reichart and Amir Kantor   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Modern Natural Language Generation (NLG) models come with massive computational and storage requirements. In this work, we study the potential of compressing them, which is crucial for real-world applications serving millions of users. We focus on Knowledge Distillation (KD) techniques, in which a small student model learns to imitate a large teacher model, allowing to transfer knowledge from the teacher to the student. In contrast to much of the previous work, our goal is to optimize the model for a specific NLG task and a specific dataset. Typically in real-world applications, in addition to labeled data there is abundant unlabeled task-specific data, which is crucial for attaining high compression rates via KD. In this work, we conduct a systematic study of task-specific KD techniques for various NLG tasks under realistic assumptions. We discuss the special characteristics of NLG distillation and particularly the exposure bias problem. Following, we derive a family of Pseudo-Target (PT) augmentation methods, substantially extending prior work on sequence-level KD. We propose the Joint-Teaching method, which applies word-level KD to multiple PTs generated by both the teacher and the student. Finally, we validate our findings in an extreme setup with no labeled examples using GPT-4 as the teacher. Our study provides practical model design observations and demonstrates the effectiveness of PT training for task-specific KD in NLG.

A Natural Bias for Language Generation Models

Clara Meister, Wojciech Jan Stokowiec, Tiago Pimentel, Lei Yu, Laura Rimell and Adhiguna Kuncoro 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
After just a few hundred training updates, a standard probabilistic model for language generation has likely not yet learnt many semantic or syntactic rules of natural language, making it difficult to estimate the probability distribution over next tokens. Yet around this point, these models have identified a simple, loss-minimising behaviour: to output the unigram distribution of the target training corpus. The use of such a heuristic raises the question: Can we initialise our models with this behaviour and save precious compute resources and model capacity? Here we show that we can effectively endow standard neural language generation models with a separate module that reflects unigram frequency statistics as prior knowledge, simply by initialising the bias term in a model's final linear layer with the log-unigram distribution. We use neu-

ral machine translation as a test bed for this simple technique and observe that it: (i) improves learning efficiency; (ii) achieves better overall performance; and perhaps most importantly (iii) appears to disentangle strong frequency effects by encouraging the model to specialise in non-frequency-related aspects of language.

Efficient Transformers with Dynamic Token Pooling
Piotr Nawrot, Jan Chorowski, Adrian Łańcucki and Edoardo Maria Ponti   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Transformers achieve unrivalled performance in modelling language, but remain inefficient in terms of memory and time complexity. A possible remedy is to reduce the sequence length in the intermediate layers by pooling fixed-length segments of tokens. Nevertheless, natural units of meaning, such as words or phrases, display varying sizes. To address this mismatch, we equip language models with a dynamic-pooling mechanism, which predicts segment boundaries in an autoregressive fashion. We compare several methods to infer boundaries, including end-to-end learning through stochastic re-parameterisation, supervised learning (based on segmentations from subword tokenizers or spikes in conditional entropy), as well as linguistically motivated boundaries. We perform character-level evaluation on texts from multiple datasets and morphologically diverse languages. The results demonstrate that dynamic pooling, which jointly segments and models language, is both faster and more accurate than vanilla Transformers and fixed-length pooling within the same computational budget.

Unsupervised Melody-to-Lyrics Generation
Yufei Tian, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Gunnar A Sigurdsson, Chenyang Tao, Wenbo Zhao, Tagyoung Chung, Jing Huang and Nanyun Peng   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Automatic melody-to-lyric generation is a task in which song lyrics are generated to go with a given melody. It is of significant practical interest and more challenging than unconstrained lyric generation as the music imposes additional constraints onto the lyrics. The training data is limited as most songs are copyrighted, resulting in models that underfit the complicated cross-modal relationship between melody and lyrics. In this work, we propose a method for generating high-quality lyrics without training on any aligned melody-lyric data. Specifically, we design a hierarchical lyric generation framework that first generates a song outline and second the complete lyrics. The framework enables disentanglement of training (based purely on text) from inference (melody-guided text generation) to circumvent the shortage of parallel data.

We leverage the segmentation and rhythm alignment between melody and lyrics to compile the given melody into decoding constraints as guidance during inference. The two-step hierarchical design also enables content control via the lyric outline, a much-desired feature for democratizing collaborative song creation. Experimental results show that our model can generate high-quality lyrics that are more on-topic, singable, intelligible, and coherent than strong baselines, for example SongMASS, a SOTA model trained on a parallel dataset, with a 24% relative overall quality improvement based on human ratings. Our code is available at https://github.com/amazon-science/unsupervised-melody-to-lyrics-generation.

Seen to Unseen: Exploring Compositional Generalization of Multi-Attribute Controllable Dialogue Generation
Weihao Zeng, Lulu Zhao, Keqing He, Ruotong Geng, Jingang Wang, Wei Wu and Weiran Xu   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Existing controllable dialogue generation work focuses on the single-attribute control and lacks generalization capability to out-of-distribution multiple attribute combinations. In this paper, we explore the compositional generalization for multi-attribute controllable dialogue generation where a model can learn from seen attribute values and generalize to unseen combinations. We propose a prompt-based disentangled controllable dialogue generation model, DCG. It learns attribute concept composition by generating attribute-oriented prompt vectors and uses a disentanglement loss to disentangle different attributes for better generalization. Besides, we design a unified reference-free evaluation framework for multiple attributes with different levels of granularities. Experiment results on two benchmarks prove the effectiveness of our method and the evaluation metric.

UniSumm and SummZoo: Unified Model and Diverse Benchmark for Few-Shot Summarization
Yulong Chen, Yang Liu, Ruochen Xu, Ziyi Yang, Chenguang Zhu, Michael Zeng and Yue Zhang   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The high annotation costs and diverse demands of various summarization tasks motivate the development of few-shot summarization. However, despite the emergence of many summarization tasks and datasets, the current training paradigm for few-shot summarization systems ignores potentially shareable knowledge in heterogeneous datasets. To this end, we propose UNISUMM, a unified few-shot summarization model pre-trained with multiple summarization tasks and can be prefix-tuned to excel at any few-shot summarization task. Meanwhile, to better evaluate few-shot summarizers, under the principles of diversity and robustness, we assemble and release a new benchmark SUMMZOO. It consists of 8 summarization tasks with multiple sets of few-shot samples for each task, covering diverse domains. Experimental results and analysis show that UNISUMM outperforms strong baselines by a large margin across all sub-tasks in SUMMZOO under both automatic and human evaluations and achieves comparable results in human evaluation compared with a GPT-3.5 model.

ExplainMeetSum: A Dataset for Explainable Meeting Summarization Aligned with Human Intent
Hyun Kim, Minsoo Cho and Seung-Hoon Na   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
To enhance the explainability of meeting summarization, we construct a new dataset called "ExplainMeetSum," an augmented version of QMSum, by newly annotating evidence sentences that faithfully "explain" a summary. Using ExplainMeetSum, we propose a novel multiple extractor guided summarization, namely Multi-DYLE, which extensively generalizes DYLE to enable using a supervised extractor based on human-aligned extractive oracles. We further present an explainability-aware task, named "Explainable Evidence Extraction" (E3), which aims to automatically detect all evidence sentences that support a given summary. Experimental results on the QMSum dataset show that the proposed Multi-DYLE outperforms DYLE with gains of up to 3.13 in the ROUGE-1 score. We further present the initial results on the E3 task, under the settings using supervised and joint evaluation metrics.

Unsupervised Extractive Summarization of Emotion Triggers
Tiberiu Sosea, Hongli Zhan, Junyi Jessy Li and Cornelia Caragea   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Understanding what leads to emotions during large-scale crises is important as it can provide groundings for expressed emotions and subsequently improve the understanding of ongoing disasters. Recent approaches trained supervised models to both detect emotions and explain emotion triggers (events and appraisals) via abstractive summarization. However, obtaining timely and qualitative abstractive summaries is expensive and extremely time-consuming, requiring highly-trained expert annotators. In time-sensitive, high-stake contexts, this can block necessary responses. We instead pursue unsupervised systems that extract triggers from text. First, we introduce CovidET-EXT, augmenting (Zhan et al., 2022)'s abstractive dataset (in the context of the COVID-19 crisis) with extractive triggers. Second, we develop new unsupervised learning models that can jointly detect emotions and summarize their triggers. Our best approach, entitled Emotion-Aware Pagerank, incorporates emotion information from external sources combined with a language understanding module, and outperforms strong baselines. We release our data and code at https://github.com/tsosea2/CovidET-EXT.

SESCORE2: Learning Text Generation Evaluation via Synthesizing Realistic Mistakes
Wenda Xu, Xian Qian, Mingxuan Wang, Lei Li and William Yang Wang   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Is it possible to train a general metric for evaluating text generation quality without human-annotated ratings? Existing learned metrics either

perform unsatisfactory across text generation tasks or require human ratings for training on specific tasks. In this paper, we propose SEScore2, a self-supervised approach for training a model-based metric for text generation evaluation. The key concept is to synthesize realistic model mistakes by perturbing sentences retrieved from a corpus. We evaluate SEScore2 and previous methods on four text generation tasks across three languages. SEScore2 outperforms all prior unsupervised metrics on four text generation evaluation benchmarks, with an average Kendall improvement of 0.158. Surprisingly, SEScore2 even outperforms the supervised BLEURT and COMET on multiple text generation tasks.

kNN-TL: k-Nearest-Neighbor Transfer Learning for Low-Resource Neural Machine Translation
Shudong Liu, Xuebo Liu, Derek F. Wong, Zhaocong Li, Wenxiang Jiao, Lidia S. Chao and Min Zhang   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Transfer learning has been shown to be an effective technique for enhancing the performance of low-resource neural machine translation (NMT). This is typically achieved through either fine-tuning a child model with a pre-trained parent model, or by utilizing the out- put of the parent model during the training of the child model. However, these methods do not make use of the parent knowledge during the child in- ference, which may limit the translation performance. In this paper, we propose a k-Nearest-Neighbor Transfer Learning (kNN-TL) approach for low-resource NMT, which leverages the parent knowledge throughout the entire developing process of the child model. Our approach in- cludes a parent-child representation alignment method, which ensures consistency in the output representations between the two models, and a child-aware datastore construction method that improves inference efficiency by selectively distilling the parent datastore based on relevance to the child model. Experimental results on four low-resource translation tasks show that kNN-TL outperforms strong baselines. Extensive analyses further demonstrate the effectiveness of our approach. Code and scripts are freely available at https://github.com/NLP2CT/kNN-TL.

Did the Models Understand Documents? Benchmarking Models for Language Understanding in Document-Level Relation Extrac- tion
Haotian Chen, Bingsheng Chen and Xiangdong Zhou   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Document-level relation extraction (DocRE) attracts more research interest recently. While models achieve consistent performance gains in DocRE, their underlying decision rules are still understudied: Do they make the right predictions according to rationales? In this paper, we take the first step toward answering this question and then introduce a new perspective on comprehensively evaluating a model. Specifically, we first conduct annotations to provide the rationales considered by humans in DocRE. Then, we conduct investigations and discover the fact that: In contrast to humans, the representative state-of-the-art (SOTA) models in DocRE exhibit different reasoning processes. Through our proposed RE-specific attacks, we next demonstrate that the significant discrepancy in decision rules between models and humans severely damages the robustness of models. After that, we introduce mean average precision (MAP) to evaluate the understanding and reasoning capa- bilities of models. According to the extensive experimental results, we finally appeal to future work to consider evaluating the understanding ability of models because the improved ability renders models more trustworthy and robust to be deployed in real-world scenarios. We make our annotations and code publicly available.

Prompts Can Play Lottery Tickets Well: Achieving Lifelong Information Extraction via Lottery Prompt Tuning
Zujie Liang, feng wei, Yin Jie, YUXI QIAN, Zhenghong Hao and Bing Han   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Thanks to the recent success of Pre-trained Language Models (PLMs), it has become a promising research direction to develop a universal model (UIE) that can solve all typical information extraction tasks within one generative framework. Nonetheless, in real-world scenarios of UIE applications, new data of different IE tasks and domains usually come in a stream over time. A desirable UIE system should be capable of continually learning new tasks without forgetting old ones, thereby allowing knowledge and functionalities expansion without re-training the whole system. In this paper, we study the UIE system under a more challenging yet practical scenario, i.e., "lifelong learning" settings, to evaluate its abilities in three aspects, including knowledge sharing and expansion, catastrophic forgetting prevention, and rapid generalization on few-shot and unseen tasks. To achieve these three goals, we present a novel parameter- and deployment-efficient prompt tuning method namely Lottery Prompt Tuning (LPT). LPT freezes the PLM's parameters and sequentially learns compact pruned prompt vectors for each task leveraging a binary prompt mask, while keeping the prompt parameters selected by the previous tasks insusceptible. Furthermore, we use a simple yet effective method to perform mask selection and show the powerful transferability of Lottery Prompts to novel tasks. Exten- sive experiments demonstrate that LPT consistently sets state-of-the-art performance on multiple lifelong learning settings of UIE, including task-incremental setting on seen tasks, few-shot adaptation, and zero-shot generalization on novel tasks.

SPEECH: Structured Prediction with Energy-Based Event-Centric Hyperspheres
Shumin Deng, Shengyu Mao, Ningyu Zhang and Bryan Hooi   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Event-centric structured prediction involves predicting structured outputs of events. In most NLP cases, event structures are complex with manifold dependency, and it is challenging to effectively represent these complicated structured events. To address these issues, we propose Structured Prediction with Energy-based Event-Centric Hyperspheres (SPEECH). SPEECH models complex dependency among event struc- tured components with energy-based modeling, and represents event classes with simple but effective hyperspheres. Experiments on two unified-annotated event datasets indicate that SPEECH is predominant in event detection and event-relation extraction tasks.

Resolving Ambiguities in Text-to-Image Generative Models
Ninareh Mehrabi, Palash Goyal, Apurv Verma, Jwala Dhamala, Varun Kumar, Qian Hu, Kai-Wei Chang, Richard Zemel, Aram Galstyan and Rahul Gupta   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Natural language often contains ambiguities that can lead to misinterpretation and miscommunication. While humans can handle ambigu- ities effectively by asking clarifying questions and/or relying on contextual cues and common-sense knowledge, resolving ambiguities can be notoriously hard for machines. In this work, we study ambiguities that arise in text-to-image generative models. We curate the Text-to- image Ambiguity Benchmark (TAB) dataset to study different types of ambiguities in text-to-image generative models. We then propose the Text-to-ImagE Disambiguation (TIED) framework to disambiguate the prompts given to the text-to-image generative models by soliciting clarifications from the end user. Through automatic and human evaluations, we show the effectiveness of our framework in generating more faithful images aligned with end user intention in the presence of ambiguities.

lilGym: Natural Language Visual Reasoning with Reinforcement Learning
Anne Wu, Kiante Brantley, Noriyuki Kojima and Yoav Artzi   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
We present lilGym, a new benchmark for language-conditioned reinforcement learning in visual environments. lilGym is based on 2,661 highly-compositional human-written natural language statements grounded in an interactive visual environment. We introduce a new ap- proach for exact reward computation in every possible world state by annotating all statements with executable Python programs. Each statement is paired with multiple start states and reward functions to form thousands of distinct Markov Decision Processes of varying dif- ficulty. We experiment with lilGym with different models and learning regimes. Our results and analysis show that while existing methods are able to achieve non-trivial performance, lilGym forms a challenging open problem. lilGym is available at https://lil.nlp.cornell.edu/lilgym/.

A Facial Expression-Aware Multimodal Multi-task Learning Framework for Emotion Recognition in Multi-party Conversations
Wenjie Zheng, Jianfei Yu, Rui Xia and Shijin Wang   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Multimodal Emotion Recognition in Multiparty Conversations (MERMC) has recently attracted considerable attention. Due to the complex-

ity of visual scenes in multi-party conversations, most previous MERMC studies mainly focus on text and audio modalities while ignoring visual information. Recently, several works proposed to extract face sequences as visual features and have shown the importance of visual information in MERMC. However, given an utterance, the face sequence extracted by previous methods may contain multiple people's faces, which will inevitably introduce noise to the emotion prediction of the real speaker. To tackle this issue, we propose a two-stage framework named Facial expressionaware Multimodal Multi-Task learning (FacialMMT). Specifically, a pipeline method is first designed to extract the face sequence of the real speaker of each utterance, which consists of multimodal face recognition, unsupervised face clustering, and face matching. With the extracted face sequences, we propose a multimodal facial expression-aware emotion recognition model, which leverages the frame-level facial emotion distributions to help improve utterance-level emotion recognition based on multi-task learning. Experiments demonstrate the effectiveness of the proposed FacialMMT framework on the benchmark MELD dataset. The source code is publicly released at https://github.com/NUSTM/FacialMMT.

A Fast Algorithm for Computing Prefix Probabilities
Franz Nowak and Ryan Cotterell   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Multiple algorithms are known for efficiently calculating the prefix probability of a string under a probabilistic context-free grammar (PCFG). Good algorithms for the problem have a runtime cubic in the length of the input string. However, some proposed algorithms are suboptimal with respect to the size of the grammar. This paper proposes a new speed-up of Jelinek and Lafferty's (1991) algorithm, which runs in $O(n^3|N|^3 + |N|^4)$, where n is the input length and $|N|$ is the number of non-terminals in the grammar. In contrast, our speed-up runs in $O(n^2|N|^3 + n^3|N|^2)$.

No clues good clues: out of context Lexical Relation Classification
Lucia Pitarch, Jordi Bernad, Lacramioara Dranca, Carlos Bobed Lisbona and Jorge Gracia 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The accurate prediction of lexical relations between words is a challenging task in Natural Language Processing (NLP). The most recent advances in this direction come with the use of pre-trained language models (PTLMs). A PTLM typically needs "well-formed" verbalized text to interact with it, either to fine-tune it or to exploit it. However, there are indications that commonly used PTLMs already encode enough linguistic knowledge to allow the use of minimal (or none) textual context for some linguistically motivated tasks, thus notably reducing human effort, the need for data pre-processing, and favoring techniques that are language neutral since do not rely on syntactic structures.

In this work, we explore this idea for the tasks of lexical relation classification (LRC) and graded Lexical Entailment (LE). After fine-tuning PTLMs for LRC with different verbalizations, our evaluation results show that very simple prompts are competitive for LRC and significantly outperform graded LE SoTA. In order to gain a better insight into this phenomenon, we perform a number of quantitative statistical analyses on the results, as well as a qualitative visual exploration based on embedding projections.

Human Inspired Progressive Alignment and Comparative Learning for Grounded Word Acquisition
Yuwei Bao, Barrett Martin Lattimer and Joyce Chai   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Human language acquisition is an efficient, supervised, and continual process. In this work, we took inspiration from how human babies acquire their first language, and developed a computational process for word acquisition through comparative learning. Motivated by cognitive findings, we generated a small dataset that enables the computation models to compare the similarities and differences of various attributes, learn to filter out and extract the common information for each shared linguistic label. We frame the acquisition of words as not only the information filtration process, but also as representation-symbol mapping. This procedure does not involve a fixed vocabulary size, nor a discriminative objective, and allows the models to continually learn more concepts efficiently. Our results in controlled experiments have shown the potential of this approach for efficient continual learning of grounded words.

What does the Failure to Reason with "Respectively" in Zero/Few-Shot Settings Tell Us about Language Models?
Ruixiang Cui, Seolhwa Lee, Daniel Hershcovich and Anders Søgaard   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Humans can effortlessly understand the coordinate structure of sentences such as "Niels Bohr and Kurt Cobain were born in Copenhagen and Seattle, *respectively*". In the context of natural language inference (NLI), we examine how language models (LMs) reason with respective readings (Gawron and Kehler, 2004) from two perspectives: syntactic-semantic and commonsense-world knowledge. We propose a controlled synthetic dataset WikiResNLI and a naturally occurring dataset NatResNLI to encompass various explicit and implicit realizations of "respectively". We show that fine-tuned NLI models struggle with understanding such readings without explicit supervision. While few-shot learning is easy in the presence of explicit cues, longer training is required when the reading is evoked implicitly, leaving models to rely on common sense inferences. Furthermore, our fine-grained analysis indicates models fail to generalize across different constructions. To conclude, we demonstrate that LMs still lag behind humans in generalizing to the long tail of linguistic constructions.

Just Like a Human Would, Direct Access to Sarcasm Augmented with Potential Result and Reaction
Changrong Min, Ximing Li, Liang Yang, Zhilin Wang, Bo Xu and Hongfei LIN   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Sarcasm, as a form of irony conveying mockery and contempt, has been widespread in social media such as Twitter and Weibo, where the sarcastic text is commonly characterized as an incongruity between the surface positive and negative situation. Naturally, it has an urgent demand to automatically identify sarcasm from social media, so as to illustrate people's real views toward specific targets. In this paper, we develop a novel sarcasm detection method, namely Sarcasm Detector with Augmentation of Potential Result and Reaction (SD-APRR). Inspired by the direct access view, we treat each sarcastic text as an incomplete version without latent content associated with implied negative situations, including the result and human reaction caused by its observable content. To fill the latent content, we estimate the potential result and human reaction for each given training sample by [xEffect] and [xReact] relations inferred by the pre-trained commonsense reasoning tool COMET, and integrate the sample with them as an augmented one. We can then employ those augmented samples to train the sarcasm detector, whose encoder is a graph neural network with a denoising module. We conduct extensive empirical experiments to evaluate the effectiveness of SD-APRR. The results demonstrate that SD-APRR can outperform strong baselines on benchmark datasets.

How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech
Aditya Yedetore, Tal Linzen, Robert Frank and R. Thomas McCoy   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
When acquiring syntax, children consistently choose hierarchical rules over competing non-hierarchical possibilities. Is this preference due to a learning bias for hierarchical structure, or due to more general biases that interact with hierarchical cues in children's linguistic input? We explore these possibilities by training LSTMs and Transformers - two types of neural networks without a hierarchical bias - on data similar in quantity and content to children's linguistic input: text from the CHILDES corpus. We then evaluate what these models have learned about English yes/no questions, a phenomenon for which hierarchical structure is crucial. We find that, though they perform well at capturing the surface statistics of child-directed speech (as measured by perplexity), both model types generalize in a way more consistent with an incorrect linear rule than the correct hierarchical rule. These results suggest that human-like generalization from text alone requires stronger biases than the general sequence-processing biases of standard neural network architectures.

Modeling Structural Similarities between Documents for Coherence Assessment with Graph Convolutional Networks

Wei Liu, Xiyan Fu and Michael Strube   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Coherence is an important aspect of text quality, and various approaches have been applied to coherence modeling. However, existing methods solely focus on a single document's coherence patterns, ignoring the underlying correlation between documents. We investigate a GCN-based coherence model that is capable of capturing structural similarities between documents. Our model first creates a graph structure for each document, from where we mine different subgraph patterns. We then construct a heterogeneous graph for the training corpus, connecting documents based on their shared subgraphs. Finally, a GCN is applied to the heterogeneous graph to model the connectivity relationships. We evaluate our method on two tasks, assessing discourse coherence and automated essay scoring. Results show that our GCN-based model outperforms all baselines, achieving a new state-of-the-art on both tasks.

Massively Multilingual Lexical Specialization of Multilingual Transformers
Tommaso Green, Simone Paolo Ponzetto and Goran Glavaš   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
While pretrained language models (PLMs) primarily serve as general-purpose text encoders that can be fine-tuned for a wide variety of downstream tasks, recent work has shown that they can also be rewired to produce high-quality word representations (i.e., static word embeddings) and yield good performance in type-level lexical tasks. While existing work primarily focused on the lexical specialization of monolingual PLMs with immense quantities of monolingual constraints, in this work we expose massively multilingual transformers (MMTs, e.g., mBERT or XLM-R) to multilingual lexical knowledge at scale, leveraging BabelNet as the readily available rich source of multilingual and cross-lingual type-level lexical knowledge. Concretely, we use BabelNet's multilingual synsets to create synonym pairs (or synonym-gloss pairs) across 50 languages and then subject the MMTs (mBERT and XLM-R) to a lexical specialization procedure guided by a contrastive objective. We show that such massively multilingual lexical specialization brings substantial gains in two standard cross-lingual lexical tasks, bilingual lexicon induction and cross-lingual word similarity, as well as in cross-lingual sentence retrieval. Crucially, we observe gains for languages unseen in specialization, indicating that multilingual lexical specialization enables generalization to languages with no lexical constraints. In a series of subsequent controlled experiments, we show that the number of specialization constraints plays a much greater role than the set of languages from which they originate.

Free Lunch: Robust Cross-Lingual Transfer via Model Checkpoint Averaging
Fabian David Schmidt, Ivan Vulić and Goran Glavaš   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Massively multilingual language models have displayed strong performance in zero-shot (ZS-XLT) and few-shot (FS-XLT) cross-lingual transfer setups, where models fine-tuned on task data in a source language are transferred without any or with only a few annotated instances to the target language(s). However, current work typically overestimates model performance as fine-tuned models are frequently evaluated at model checkpoints that generalize best to validation instances in the target languages. This effectively violates the main assumptions of 'true' ZS-XLT and FS-XLT. Such XLT setups require robust methods that do not depend on labeled target language data for validation and model selection. In this work, aiming to improve the robustness of 'true' ZS-XLT and FS-XLT, we propose a simple and effective method that averages different checkpoints (i.e., model snapshots) during task fine-tuning. We conduct exhaustive ZS-XLT and FS-XLT experiments across higher-level semantic tasks (NLI, extractive QA) and lower-level token classification tasks (NER, POS). The results indicate that averaging model checkpoints yields systematic and consistent performance gains across diverse target languages in all tasks. Importantly, it simultaneously substantially desensitizes XLT to varying hyperparameter choices in the absence of target language validation. We also show that checkpoint averaging benefits performance when further combined with run averaging (i.e., averaging the parameters of models fine-tuned over independent runs).

What the DAAM: Interpreting Stable Diffusion Using Cross Attention
Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin and Ferhan Ture   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Diffusion models are a milestone in text-to-image generation, but they remain poorly understood, lacking interpretability analyses. In this paper, we perform a text-image attribution analysis on Stable Diffusion, a recently open-sourced model. To produce attribution maps, we upscale and aggregate cross-attention maps in the denoising module, naming our method DAAM. We validate it by testing its segmentation ability on nouns, as well as its generalized attribution quality on all parts of speech, rated by humans. On two generated datasets, we attain a competitive 58.8-64.8 mIoU on noun segmentation and fair to good mean opinion scores (3.4-4.2) on generalized attribution. Then, we apply DAAM to study the role of syntax in the pixel space across head–dependent heat map interaction patterns for ten common dependency relations. We show that, for some relations, the head map consistently subsumes the dependent, while the opposite is true for others. Finally, we study several semantic phenomena, focusing on feature entanglement; we find that the presence of cohyponyms worsens generation quality by 9%, and descriptive adjectives attend too broadly. We are the first to interpret large diffusion models from a visuolinguistic perspective, which enables future research. Our code is at https://github.com/castorini/daam.

DIP: Dead code Insertion based Black-box Attack for Programming Language Model
CheolWon Na, YunSeok Choi and Jee-Hyong Lee   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Automatic processing of source code, such as code clone detection and software vulnerability detection, is very helpful to software engineers. Large pre-trained Programming Language (PL) models (such as CodeBERT, GraphCodeBERT, CodeT5, etc.), show very powerful performance on these tasks. However, these PL models are vulnerable to adversarial examples that are generated with slight perturbation. Unlike natural language, an adversarial example of code must be semantic-preserving and compilable. Due to the requirements, it is hard to directly apply the existing attack methods for natural language models. In this paper, we propose DIP (Dead code Insertion based Black-box Attack for Programming Language Model), a high-performance and effective black-box attack method to generate adversarial examples using dead code insertion. We evaluate our proposed method on 9 victim downstream-task large code models. Our method outperforms the state-of-the-art black-box attack in both attack efficiency and attack quality, while generated adversarial examples are compiled preserving semantic functionality.

On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning
Omar Shaikh, Hongxin Zhang, William Held, Michael S Bernstein and Diyi Yang   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Generating a Chain of Thought (CoT) has been shown to consistently improve large language model (LLM) performance on a wide range of NLP tasks. However, prior work has mainly focused on logical reasoning tasks (e.g. arithmetic, commonsense QA); it remains unclear whether improvements hold for more diverse types of reasoning, especially in socially situated contexts. Concretely, we perform a controlled evaluation of zero-shot CoT across two socially sensitive domains: harmful questions and stereotype benchmarks. We find that zero-shot CoT reasoning in sensitive domains significantly increases a model's likelihood to produce harmful or undesirable output, with trends holding across different prompt formats and model variants. Furthermore, we show that harmful CoTs increase with model size, but decrease with improved instruction following. Our work suggests that zero-shot CoT should be used with caution on socially important tasks, especially when marginalized groups or sensitive topics are involved.

Language Models Get a Gender Makeover: Mitigating Gender Bias with Few-Shot Data Interventions
Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang and Louis-Philippe Morency   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Societal biases present in pre-trained large language models are a critical issue as these models have been shown to propagate biases in countless downstream applications, rendering them unfair towards specific groups of people. Since large-scale retraining of these models from scratch is both time and compute-expensive, a variety of approaches have been previously proposed that de-bias a pre-trained model. While the majority of current state-of-the-art debiasing methods focus on changes to the training regime, in this paper, we propose data intervention strategies as a powerful yet simple technique to reduce gender bias in pre-trained models. Specifically, we empirically show that by fine-tuning a pre-trained model on only 10 debiased (intervened) training examples, the tendency to favor any gender is significantly reduced. Since our proposed method only needs a few training examples, we argue that our few-shot de-biasing approach is highly feasible and practical. Through extensive experimentation, we show that our de-biasing technique performs better than competitive state-of-the-art baselines with minimal loss in language modeling ability.

MISGENDERED: Limits of Large Language Models in Understanding Pronouns
Tamanna Hossain, Sunipa Dev and Sameer Singh　　　　　　　　　11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Content Warning: This paper contains examples of misgendering and erasure that could be offensive and potentially triggering.

Gender bias in language technologies has been widely studied, but research has mostly been restricted to a binary paradigm of gender. It is essential also to consider non-binary gender identities, as excluding them can cause further harm to an already marginalized group. In this paper, we comprehensively evaluate popular language models for their ability to correctly use English gender-neutral pronouns (e.g., singular they, them) and neo-pronouns (e.g., ze, xe, thon) that are used by individuals whose gender identity is not represented by binary pronouns. We introduce Misgendered, a framework for evaluating large language models' ability to correctly use preferred pronouns, consisting of (i) instances declaring an individual's pronoun, followed by a sentence with a missing pronoun, and (ii) an experimental setup for evaluating masked and auto-regressive language models using a unified method. When prompted out-of-the-box, language models perform poorly at correctly predicting neo-pronouns (averaging 7.6% accuracy) and gender-neutral pronouns (averaging 31.0% accuracy). This inability to generalize results from a lack of representation of non-binary pronouns in training data and memorized associations. Few-shot adaptation with explicit examples in the prompt improves the performance but plateaus at only 45.4% for neo-pronouns. We release the full dataset, code, and demo at https://tamannahossainkay.github.io/misgendered/.

Improving Gender Fairness of Pre-Trained Language Models without Catastrophic Forgetting
Zahra Fatemi, Chen Xing, Wenhao Liu and Caimming Xiong　　　　　11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Existing studies addressing gender bias of pre-trained language models, usually build a small gender-neutral data set and conduct a second phase pre-training on the model with such data. However, given the limited size and concentrated focus of the gender-neutral data, catastrophic forgetting would occur during second-phase pre-training. Forgetting information in the original training data may damage the model's downstream performance by a large margin. In this work, we empirically show that catastrophic forgetting occurs in such methods by evaluating them with general NLP tasks in GLUE. Then, we propose a new method, GEnder Equality Prompt (GEEP), to improve gender fairness of pre-trained models with less forgetting. GEEP freezes the pre-trained model and learns gender-related prompts with gender-neutral data.Empirical results show that GEEP not only achieves SOTA performances on gender fairness tasks, but also forgets less and performs better on GLUE by a large margin.

Considerations for meaningful sign language machine translation based on glosses
Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios and Sarah Ebling　　11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Automatic sign language processing is gaining popularity in Natural Language Processing (NLP) research (Yin et al., 2021). In machine translation (MT) in particular, sign language translation based on glosses is a prominent approach. In this paper, we review recent works on neural gloss translation. We find that limitations of glosses in general and limitations of specific datasets are not discussed in a transparent manner and that there is no common standard for evaluation.

To address these issues, we put forward concrete recommendations for future research on gloss translation. Our suggestions advocate awareness of the inherent limitations of gloss-based approaches, realistic datasets, stronger baselines and convincing evaluation.

Dialect-robust Evaluation of Generated Text
Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein and Sebastian Gehrmann
11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Text generation metrics that are not robust to dialect variation make it impossible to tell how well systems perform for many groups of users, and can even penalize systems for producing text in lower-resource dialects. In this paper, we introduce a suite of methods to assess whether metrics are dialect robust. These methods show that state-of-the-art metrics are not dialect robust: they often prioritize dialect similarity over semantics, preferring outputs that are semantically incorrect over outputs that match the semantics of the reference but contain dialect differences. As a step towards dialect-robust metrics for text generation, we propose NANO, which introduces regional and language information to the metric's pretraining. NANO significantly improves dialect robustness while preserving the correlation between automated metrics and human ratings. It also enables a more ambitious approach to evaluation, dialect awareness, in which system outputs are scored by both semantic match to the reference and appropriateness in any specified dialect.

What about "em"? How Commercial Machine Translation Fails to Handle (Neo-)Pronouns
Anne Lauscher, Debora Nozza, Ehm Miltersen, Archie Crowley and Dirk Hovy　　11:00-12:30 (Frontenac Ballroom and Queen's Quay)
As 3rd-person pronoun usage shifts to include novel forms, e.g., neopronouns, we need more research on identity-inclusive NLP. Exclusion is particularly harmful in one of the most popular NLP applications, machine translation (MT). Wrong pronoun translations can discriminate against marginalized groups, e.g., non-binary individuals (Dev et al., 2021). In this "reality check", we study how three commercial MT systems translate 3rd-person pronouns. Concretely, we compare the translations of gendered vs. gender-neutral pronouns from English to five other languages (Danish, Farsi, French, German, Italian), and vice versa, from Danish to English. Our error analysis shows that the presence of a gender-neutral pronoun often leads to grammatical and semantic translation errors. Similarly, gender neutrality is often not preserved. By surveying the opinions of affected native speakers from diverse languages, we provide recommendations to address the issue in future MT research.

Dealing with Semantic Underspecification in Multimodal NLP
Sandro Pezzelle　　　　　　　　　　　　　　　　　　　　　11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Intelligent systems that aim at mastering language as humans do must deal with its semantic underspecification, namely, the possibility for a linguistic signal to convey only part of the information needed for communication to succeed. Consider the usages of the pronoun they, which can leave the gender and number of its referent(s) underspecified. Semantic underspecification is not a bug but a crucial language feature that boosts its storage and processing efficiency. Indeed, human speakers can quickly and effortlessly integrate semantically-underspecified linguistic signals with a wide range of non-linguistic information, e.g., the multimodal context, social or cultural conventions, and shared knowledge. Standard NLP models have, in principle, no or limited access to such extra information, while multimodal systems grounding language into other modalities, such as vision, are naturally equipped to account for this phenomenon. However, we show that they struggle with it, which could negatively affect their performance and lead to harmful consequences when used for applications. In this position paper, we argue that our community should be aware of semantic underspecification if it aims to develop language technology that can successfully

interact with human users. We discuss some applications where mastering it is crucial and outline a few directions toward achieving this goal.

Forgotten Knowledge: Examining the Citational Amnesia in NLP
Janvijay Singh, Mukund Rungta, Diyi Yang and Saif M. Mohammad   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Citing papers is the primary method through which modern scientific writing discusses and builds on past work. Collectively, citing a diverse set of papers (in time and area of study) is an indicator of how widely the community is reading. Yet, there is little work looking at broad temporal patterns of citation. This work systematically and empirically examines: How far back in time do we tend to go to cite papers? How has that changed over time, and what factors correlate with this citational attention/amnesia? We chose NLP as our domain of interest and analyzed approximately 71.5K papers to show and quantify several key trends in citation. Notably, around 62% of cited papers are from the immediate five years prior to publication, whereas only about 17% are more than ten years old. Furthermore, we show that the median age and age diversity of cited papers were steadily increasing from 1990 to 2014, but since then, the trend has reversed, and current NLP papers have an all-time low temporal citation diversity. Finally, we show that unlike the 1990s, the highly cited papers in the last decade were also papers with the least citation diversity, likely contributing to the intense (and arguably harmful) recency focus. Code, data, and a demo are available on the project homepage.

Theory-Grounded Computational Text Analysis
Arya D. McCarthy and Giovanna Maria Dora Dore   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
In this position paper, we argue that computational text analysis lacks and requires organizing principles. A broad space separates its two constituent disciplines—natural language processing and social science—which has to date been sidestepped rather than filled by applying increasingly complex computational models to problems in social science research. We contrast descriptive and integrative findings, and our review of approximately 60 papers on computational text analysis reveals that those from *ACL venues are typically descriptive. The lack of theory began at the area's inception and has over the decades, grown more important and challenging. A return to theoretically grounded research questions will propel the area from both theoretical and methodological points of view.

What Do NLP Researchers Believe? Results of the NLP Community Metasurvey
Julian Michael, Ari Holtzman, Alicia Parrish, Aaron Mueller, Alex Wang, Angelica Chen, Divyam Madaan, Nikita Nangia, Richard Yuanzhe Pang, Jason Phang and Samuel R. Bowman   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
We present the results of the NLP Community Metasurvey. Run from May to June 2022, it elicited opinions on controversial issues, including industry influence in the field, concerns about AGI, and ethics. Our results put concrete numbers to several controversies: For example, respondents are split in half on the importance of artificial general intelligence, whether language models understand language, and the necessity of linguistic structure and inductive bias for solving NLP problems. In addition, the survey posed meta-questions, asking respondents to predict the distribution of survey responses. This allows us to uncover false sociological beliefs where the community's predictions don't match reality. Among other results, we find that the community greatly overestimates its own belief in the usefulness of benchmarks and the potential for scaling to solve real-world problems, while underestimating its belief in the importance of linguistic structure, inductive bias, and interdisciplinary science.

Student Research Workshop
11:00-12:30 (Pier 2&3)

[SRW] Assessing Chain-of-Thought Reasoning against Lexical Negation: A Case Study on Syllogism
Mengyu Ye, Tatsuki Kuribayashi, Jun Suzuki, Hiroaki Funayama and Goro Kobayashi   11:00-11:15 (Pier 2&3)
A investage on LLMs

[SRW] Is a Knowledge-based Response Engaging?: An Analysis on Knowledge-Grounded Dialogue with Information Source Annotation
Takashi Kodama, Hirokazu Kiyomaru, Yin Jou Huang, Taro Okahisa and Sadao Kurohashi   11:15-11:30 (Pier 2&3)
This paper investigates how humans incorporate speaker-derived information by annotating the utterances in a knowledge-grounded dialogue corpus.

[SRW] LECO: Improving Early Exiting via Learned Exits and Comparison-based Exiting Mechanism
Jingfan Zhang, Ming Tan, Pengyu Dai and Wei Zhu   11:30-11:45 (Pier 2&3)
Speeding up the inference of pretrained models by designing better intermediate early exits and a comparison-based early exiting mechanism

[SRW] How-to Guides for Specific Audiences: A Corpus and Initial Findings
Nicola Fanton, Agnieszka Falenska and Michael Roth   11:45-12:00 (Pier 2&3)
We collect how-to guides for different target audiences and investigate qualitative and quantitative differences.

Language Grounding to Vision, Robotics, and Beyond
11:00-12:30 (Pier 4&5)

Learning to Imagine: Visually-Augmented Natural Language Generation
Tianyi Tang, Yushuo Chen, Yifan Du, Junyi Li, Wayne Xin Zhao and Ji-Rong Wen   11:00-11:15 (Pier 4&5)
People often imagine relevant scenes to aid in the writing process. In this work, we aim to utilize visual information for composition in the same manner as humans. We propose a method, LIVE, that makes pre-trained language models (PLMs) Learn to Imagine for Visually-augmented natural language gEneration. First, we imagine the scene based on the text: we use a diffusion model to synthesize high-quality images conditioned on the input texts. Second, we use CLIP to determine whether the text can evoke the imagination in a posterior way. Finally, our imagination is dynamic, and we conduct synthesis for each sentence rather than generate only one image for an entire paragraph. Technically, we propose a novel plug-and-play fusion layer to obtain visually-augmented representations for each text. Our vision-text fusion layer is compatible with Transformer-based architecture. We have conducted extensive experiments on four generation tasks using BART and T5, and the automatic results and human evaluation demonstrate the effectiveness of our proposed method. We will release the code, model, and data at the link: https://github.com/RUCAIBox/LIVE.

NUWA-XL: Diffusion over Diffusion for eXtremely Long Video Generation
Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, Jianlong Fu, Ming Gong, Lijuan Wang, Zicheng Liu, Houqiang Li and Nan Duan 11:15-11:30 (Pier 4&5)
In this paper, we propose NUWA-XL, a novel Diffusion over Diffusion architecture for eXtremely Long video generation. Most current work generates long videos segment by segment sequentially, which normally leads to the gap between training on short videos and inferring long videos, and the sequential generation is inefficient. Instead, our approach adopts a "coarse-to-fine" process, in which the video can be generated in parallel at the same granularity. A global diffusion model is applied to generate the keyframes across the entire time range, and then local diffusion models recursively fill in the content between nearby frames. This simple yet effective strategy allows us to directly train on long videos (3376 frames) to reduce the training-inference gap and makes it possible to generate all segments in parallel. To evaluate our model, we build FlintstonesHD dataset, a new benchmark for long video generation. Experiments show that our model not only generates high-quality long videos with both global and local coherence, but also decreases the average inference time from 7.55min to 26s (by 94.26%) at the same hardware setting when generating 1024 frames. The homepage link is [NUWA-XL](https://msra-nuwa.azurewebsites.net)

Weakly-Supervised Spoken Video Grounding via Semantic Interaction Learning
Ye Wang, Wang Lin, Shengyu Zhang, Tao Jin, Linjun Li, Xize Cheng and Zhou Zhao 11:30-11:45 (Pier 4&5)
The task of spoken video grounding aims to localize moments in videos that are relevant to descriptive spoken queries. However, extracting semantic information from speech and modeling the cross-modal correlation pose two critical challenges. Previous studies solve them by representing spoken queries based on the matched video frames, which require tremendous effort for frame-level labeling. In this work, we investigate weakly-supervised spoken video grounding, i.e., learning to localize moments without expensive temporal annotations. To effectively represent the cross-modal semantics, we propose Semantic Interaction Learning (SIL), a novel framework consisting of the acoustic-semantic pre-training (ASP) and acoustic-visual contrastive learning (AVCL). In ASP, we pre-train an effective encoder for the grounding task with three comprehensive tasks, where the robustness task enhances stability by explicitly capturing the invariance between time- and frequency-domain features, the conciseness task avoids over-smooth attention by compressing long sequence into segments, and the semantic task improves spoken language understanding by modeling the precise semantics. In AVCL, we mine pseudo labels with discriminative sampling strategies and directly strengthen the interaction between speech and video by maximizing their mutual information. Extensive experiments demonstrate the effectiveness and superiority of our method.

Virtual Poster

11:00-12:30 (Pier 7&8)

[Industry] GKD: A General Knowledge Distillation Framework for Large-scale Pre-trained Language Model
Shicheng Tan, Weng Lam Tam, Yuanchun Wang, Wenwen Gong, Shu Zhao, Peng Zhang and Jie Tang 11:00-12:30 (Pier 7&8)
Currently, the reduction in the parameter scale of large-scale pre-trained language models (PLMs) through knowledge distillation has greatly facilitated their widespread deployment on various devices. However, the deployment of knowledge distillation systems faces great challenges in real-world industrial-strength applications, which require the use of complex distillation methods on even larger-scale PLMs (over 10B), limited by memory on GPUs and the switching of methods. To overcome these challenges, we propose GKD, a general knowledge distillation framework that supports distillation on larger-scale PLMs using various distillation methods. With GKD, developers can build larger distillation models on memory-limited GPUs and easily switch and combine different distillation methods within a single framework. Experimental results show that GKD can support the distillation of at least 100B-scale PLMs and 25 mainstream methods on 8 NVIDIA A100 (40GB) GPUs.

[Industry] Towards Building a Robust Toxicity Predictor
Dmitriy Bespalov, Sourav Bhabesh, Yi Xiang, Liutong Zhou and Yanjun Qi 11:00-12:30 (Pier 7&8)
Recent NLP literature pays little attention to the robustness of toxicity language predictors, while these systems are most likely to be used in adversarial contexts. This paper presents a novel adversarial attack, \texttt{ToxicTrap}, introducing small word-level perturbations to fool SOTA text classifiers to predict toxic text samples as benign. \texttt{ToxicTrap} exploits greedy based search strategies to enable fast and effective generation of toxic adversarial examples. Two novel goal function designs allow \texttt{ToxicTrap} to identify weaknesses in both multiclass and multilabel toxic language detectors. Our empirical results show that SOTA toxicity text classifiers are indeed vulnerable to the proposed attacks, attaining over 98\% attack success rates in multilabel cases. We also show how a vanilla adversarial training and its improved version can help increase robustness of a toxicity detector even against unseen attacks.

[Industry] "Knowledge is Power": Constructing Knowledge Graph of Abdominal Organs and Using Them for Automatic Radiology Report Generation
Kaveri Kale, Pushpak Bhattacharyya, Aditya Shetty, Milind Gune, Kush Shrivastava, Rustom Lawyer and Spriha Biswas 11:00-12:30 (Pier 7&8)
In conventional radiology practice, the radiologist dictates the diagnosis to the transcriptionist, who then prepares a preliminary formatted report referring to the notes, after which the radiologist reviews the report, corrects the errors, and signs off. This workflow is prone to delay and error. In this paper, we report our work on automatic radiology report generation from radiologists' dictation, which is in collaboration with a startup about to become Unicorn. A major contribution of our work is the set of knowledge graphs (KGs) of ten abdominal organs-Liver, Kidney, Gallbladder, Uterus, Urinary bladder, Ovary, Pancreas, Prostate, Biliary Tree, and Bowel. Our method for constructing these KGs relies on extracting entity1-relation-entity2 triplets from a large collection (about 10,000) of free-text radiology reports. The quality and coverage of the KGs are verified by two experienced radiologists (practicing for the last 30 years and 8 years, respectively). The dictation of the radiologist is automatically converted to what is called a pathological description which is the clinical description of the findings of the radiologist during ultrasonography (USG). Our knowledge-enhanced deep learning model improves the reported BLEU-3, ROUGE-L, METEOR, and CIDEr scores of the pathological description generation by 2%, 4%, 2% and 2% respectively. To the best of our knowledge, this is the first attempt at representing the abdominal organs in the form of knowledge graphs and utilising these graphs for the automatic generation of USG reports. A Minimum Viable Product (MVP) has been made available to the beta users, i.e., radiologists of reputed hospitals, for testing and evaluation. Our solution guarantees report generation within 30 seconds of running a scan.

[Industry] Label efficient semi-supervised conversational intent classification
Mandar Kulkarni, Kyung Kim, Nikesh Garera and Anusua Trivedi 11:00-12:30 (Pier 7&8)
To provide a convenient shopping experience and to answer user queries at scale, conversational platforms are essential for e-commerce. The user queries can be pre-purchase questions, such as product specifications and delivery time related, or post-purchase queries, such as exchange and return. A chatbot should be able to understand and answer a variety of such queries to help users with relevant information. One of the important modules in the chatbot is automated intent identification, i.e., understanding the user's intention from the query text.

Due to non-English speaking users interacting with the chatbot, we often get a significant percentage of code mix queries and queries with grammatical errors, which makes the problem more challenging. This paper proposes a simple yet competent Semi-Supervised Learning (SSL) approach for label-efficient intent classification. We use a small labeled corpus and relatively larger unlabeled query data to train a transformer model. For training the model with labeled data, we explore supervised MixUp data augmentation. To train with unlabeled data, we explore label consistency with dropout noise. We experiment with different pre-trained transformer architectures, such as BERT and sentence-BERT. Experimental results demonstrate that the proposed approach significantly improves over the supervised baseline, even with a limited labeled set. A variant of the model is currently deployed in production.

[Industry] Tab-CQA: A Tabular Conversational Question Answering Dataset on Financial Reports

Chuang Liu, Junzhuo Li and Deyi Xiong   11:00-12:30 (Pier 7&8)
Existing conversational question answering (CQA) datasets have been usually constructed from unstructured texts in English. In this paper, we propose Tab-CQA, a tabular CQA dataset created from Chinese financial reports that are extracted from listed companies in a wide range of different sectors in the past 30 years. From these reports, we select 2,463 tables, and manually generate 2,463 conversations with 35,494 QA pairs. Additionally, we select 4,578 tables, from which 4,578 conversations with 73,595 QA pairs are automatically created via a template-based method. With the manually- and automatically-generated conversations, Tab-CQA contains answerable and unanswerable questions. For the answerable questions, we further diversify them to cover a wide range of skills, e.g., table retrieval, fact checking, numerical reasoning, so as to accommodate real-world scenarios. We further propose two different tabular CQA models, a text-based model and an operation-based model, and evaluate them on Tab-CQA. Experiment results show that Tab-CQA is a very challenging dataset, where a huge performance gap exists between human and neural models. We will publicly release Tab-CQA as a benchmark testbed to promote further research on Chinese tabular CQA.

[Industry] Improving Knowledge Production Efficiency With Question Answering on Conversation

Changlin Yang, Siye Liu, Sen Hu, Wangshu Zhang, Teng Xu and Jing Zheng   11:00-12:30 (Pier 7&8)
Through an online customer service application, we have collected many conversations between customer service agents and customers. Building a knowledge production system can help reduce the labor cost of maintaining the FAQ database for the customer service chatbot, whose core module is question answering (QA) on these conversations. However, most existing researches focus on document-based QA tasks, and there is a lack of researches on conversation-based QA and related datasets, especially in Chinese language. The challenges of conversation-based QA include: 1) answers may be scattered among multiple dialogue turns; 2) understanding complex dialogue contexts is more complicated than documents. To address these challenges, we propose a multi-span extraction model on this task and introduce continual pre-training and multi-task learning schemes to further improve model performance. To validate our approach, we construct two Chinese datasets using dialogues as the knowledge source, namely cs-qaconv and kd-qaconv, respectively. Experimental results demonstrate that the proposed model outperforms the baseline on both datasets. The online application also verifies the effectiveness of our method. The dataset kd-qaconv will be released publicly for research purposes.

[Industry] Domain-specific transformer models for query translation

Mandar Kulkarni, Nikesh Garera and Anusua Trivedi   11:00-12:30 (Pier 7&8)
Due to the democratization of e-commerce, many product companies are listing their goods for online shopping. For periodic buying within a domain such as Grocery, consumers are generally inclined to buy certain brands of products. Due to a large non-English speaking population in India, we observe a significant percentage of code-mix Hinglish search queries e.g., sasta atta. An intuitive approach to dealing with code-mix queries is to train an encoder-decoder model to translate the query to English to perform the search. However, the problem becomes non-trivial when the brand names themselves have Hinglish names and possibly have a literal English translation. In such queries, only the context (non-brand name) Hinglish words needs to be translated. In this paper, we propose a simple yet effective modification to the transformer training to preserve/correct Grocery brand names in the output while selectively translating the context words. To achieve this, we use an additional dataset of popular Grocery brand names. Brand names are added as tokens to the model vocabulary, and the token embeddings are randomly initialized. Further, we introduce a Brand loss in training the translation model. Brand loss is a cross entropy loss computed using a denoising auto-encoder objective with brand name data. We warm-start the training from a public pre-trained checkpoint (such as BART/T5) and further adapt it for query translation using the domain data. The proposed model is generic and can be used with English as well as code-mix Hinglish queries alleviating the need for language detection. To reduce the latency of the model for the production deployment, we use knowledge distillation and quantization. Experimental evaluation indicates that the proposed approach improves translation results by preserving/correcting English/Hinglish brand names. After positive results with A/B testing, the model is currently deployed in production.

[Industry] Hunt for Buried Treasures: Extracting Unclaimed Embodiments from Patent Specifications

Chikara Hashimoto, Gautam Kumar, Shuichiro Hashimoto and Jun Suzuki   11:00-12:30 (Pier 7&8)
Patent applicants write patent specifications that describe embodiments of inventions. Some embodiments are claimed for a patent, while others may be unclaimed due to strategic considerations. Unclaimed embodiments may be extracted by applicants later and claimed in continuing applications to gain advantages over competitors. Despite being essential for corporate intellectual property (IP) strategies, unclaimed embodiment extraction is conducted manually, and little research has been conducted on its automation. This paper presents a novel task of unclaimed embodiment extraction (UEE) and a novel dataset for the task. Our experiments with Transformer-based models demonstrated that the task was challenging as it required conducting natural language inference on patent specifications, which consisted of technical, long, syntactically and semantically involved sentences. We release the dataset and code to foster this new area of research.

[Industry] Learn over Past, Evolve for Future: Forecasting Temporal Trends for Fake News Detection

Beizhe Hu, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Zhengjia Wang and Zhiwei Jin   11:00-12:30 (Pier 7&8)
Fake news detection has been a critical task for maintaining the health of the online news ecosystem. However, very few existing works consider the temporal shift issue caused by the rapidly-evolving nature of news data in practice, resulting in significant performance degradation when training on past data and testing on future data. In this paper, we observe that the appearances of news events on the same topic may display discernible patterns over time, and posit that such patterns can assist in selecting training instances that could make the model adapt better to future data. Specifically, we design an effective framework FTT (Forecasting Temporal Trends), which could forecast the temporal distribution patterns of news data and then guide the detector to fast adapt to future distribution. Experiments on the real-world temporally split dataset demonstrate the superiority of our proposed framework.

[Industry] Tab-Cleaner: Weakly Supervised Tabular Data Cleaning via Pre-training for E-commerce Catalog

Kewei Cheng, Xian Li, Zhengyang Wang, Chenwei Zhang, Binxuan Huang, Yifan Ethan Xu, Xin Luna Dong and Yizhou Sun   11:00-12:30 (Pier 7&8)
Product catalogs, conceptually in the form of text-rich tables, are self-reported by individual retailers and thus inevitably contain noisy facts. Verifying such textual attributes in product catalogs is essential to improve their reliability. However, popular methods for processing free-text content, such as pre-trained language models, are not particularly effective on structured tabular data since they are typically trained on free-form natural language texts. In this paper, we present Tab-Cleaner, a model designed to handle error detection over text-rich tabular data following a pre-training / fine-tuning paradigm. We train Tab-Cleaner on a real-world Amazon Product Catalog table w.r.t millions of products

and show improvements over state-of-the-art methods by 16\% on PR AUC over attribute applicability classification task and by 11\% on PR AUC over attribute value validation task.

[Industry] Boosting Transformers and Language Models for Clinical Prediction in Immunotherapy

Zekai Chen, Mariann Micsinai Balan and Kevin Brown　　　　　　　　　　　　　　　11:00-12:30 (Pier 7&8)

Clinical prediction is an essential task in the healthcare industry. However, the recent success of transformers, on which large language models are built, has not been extended to this domain. In this research, we explore the use of transformers and language models in prognostic prediction for immunotherapy using real-world patients' clinical data and molecular profiles. This paper investigates the potential of transformers to improve clinical prediction compared to conventional machine learning approaches and addresses the challenge of few-shot learning in predicting rare disease areas. The study benchmarks the efficacy of baselines and language models on prognostic prediction across multiple cancer types and investigates the impact of different pretrained language models under few-shot regimes. The results demonstrate significant improvements in accuracy and highlight the potential of NLP in clinical research to improve early detection and intervention for different diseases.

[Industry] Chemical Language Understanding Benchmark

Yunsoo Kim, Hyuk Ko, Jane Lee, Hyun Young Heo, Jinyoung Yang, Sungsoo Lee and Kyu-hwang Lee　　　　　11:00-12:30 (Pier 7&8)

In this paper, we introduce the benchmark datasets named CLUB (Chemical Language Understanding Benchmark) to facilitate NLP research in the chemical industry. We have 4 datasets consisted of text and token classification tasks. As far as we have recognized, it is one of the first examples of chemical language understanding benchmark datasets consisted of tasks for both patent and literature articles provided by industrial organization. All the datasets are internally made by chemists from scratch. Finally, we evaluate the datasets on the various language models based on BERT and RoBERTa, and demonstrate the model performs better when the domain of the pretrained models are closer to chemistry domain. We provide baselines for our benchmark as 0.8054 in average, and we hope this benchmark is used by many researchers in both industry and academia.

[Industry] Automated Digitization of Unstructured Medical Prescriptions

Megha Sharma, Tushar Vatsal, Srujana Merugu and Aruna Rajan　　　　　　　　　　　11:00-12:30 (Pier 7&8)

Automated digitization of prescription images is a critical prerequisite to scale digital healthcare services such as online pharmacies. This is challenging in emerging markets since prescriptions are not digitized at source and patients lack the medical expertise to interpret prescriptions to place orders. In this paper, we present prescription digitization system for online medicine ordering built with minimal supervision. Our system uses a modular pipeline comprising a mix of ML and rule-based components for (a) image to text extraction, (b) segmentation into blocks and medication items, (c) medication attribute extraction, (d) matching against medicine catalog, and (e) shopping cart building. Our approach efficiently utilizes multiple signals like layout, medical ontologies, and semantic embeddings via LayoutLMv2 model to yield substantial improvement relative to strong baselines on medication attribute extraction. Our pipeline achieves +5.9% gain in precision@3 and +5.6% in recall@3 over catalog-based fuzzy matching baseline for shopping cart building for printed prescriptions.

[Industry] Constrained Policy Optimization for Controlled Self-Learning in Conversational AI Systems

Mohammad Kachuee and Sungjin Lee　　　　　　　　　　　　　　　　　　　　11:00-12:30 (Pier 7&8)

Recently, self-learning methods based on user satisfaction metrics and contextual bandits have shown promising results to enable consistent improvements in conversational AI systems. However, directly targeting such metrics by off-policy bandit learning objectives often increases the risk of making abrupt policy changes that break the current user experience. In this study, we introduce a scalable framework for supporting fine-grained exploration targets for individual domains via user-defined constraints. For example, we may want to ensure fewer policy deviations in business-critical domains such as shopping, while allocating more exploration budget to domains such as music. We present a novel meta-gradient learning approach that is scalable and practical to address this problem. The proposed method adjusts constraint violation penalty terms adaptively through a meta objective that encourages balanced constraint satisfaction across domains. We conducted extensive experiments on a real-world conversational AI and using a set of realistic constraint benchmarks. The proposed approach has been deployed in production for a large-scale commercial assistant, enabling the best balance between the policy value and constraint satisfaction rate.

[Industry] CocaCLIP: Exploring Distillation of Fully-Connected Knowledge Interaction Graph for Lightweight Text-Image Retrieval

Jiapeng Wang, Chengyu Wang, Xiaodan Wang, Jun Huang and Lianwen Jin　　　　　　　11:00-12:30 (Pier 7&8)

Large-scale pre-trained text-image models with dual-encoder architectures (such as CLIP) are typically adopted for various vision-language applications, including text-image retrieval. However, these models are still less practical on edge devices or for real-time situations, due to the substantial indexing and inference time and the large consumption of computational resources. Although knowledge distillation techniques have been widely utilized for uni-modal model compression, how to expand them to the situation when the numbers of modalities and teachers/students are doubled has been rarely studied. In this paper, we conduct comprehensive experiments on this topic and propose the fully-Connected knowledge interaction graph (Coca) technique for cross-modal pre-training distillation. Based on our findings, the resulting CocaCLIP achieves SOTA performances on the widely-used Flickr30K and MSCOCO benchmarks under the lightweight setting. An industry application of our method on an e-commercial platform further demonstrates the significant effectiveness of CocaCLIP.

[Industry] HyperT5: Towards Compute-Efficient Korean Language Modeling

Dongju Park, Soonwon Ka, Kang Min Yoo, Gichang Lee and Jaewook Kang　　　　　　11:00-12:30 (Pier 7&8)

Pretraining and fine-tuning language models have become the standard practice in industrial natural language processing (NLP), but developing and deploying general-purpose language models without the abundant computation or data resources is a real-world issue faced by smaller organizations or communities whose main focus is languages with less accessible resources (e.g., non-English). This paper explores the sequence-to-sequence (seq2seq) language model architecture as a more practical and compute-efficient alternative to the decoder-oriented approach (e.g., GPT-3), accompanied by novel findings in compute-optimality analyses. We successfully trained billion-scale Korean-language seq2seq language models that strongly outperform other competitive models in Korean benchmarks. Moreover, we demonstrate that such language models can be more efficiently utilized by employing a heavy pre-finetuning strategy, by showcasing a case study on dialog-task adaptation. Our case study shows that adopting language models with more readily available domain-specific unlabeled data greatly improves fine-tuning data efficiency in low-resource settings.

[Industry] SPM: A Split-Parsing Method for Joint Multi-Intent Detection and Slot Filling

Sheng Jiang, Su Zhu, Ruisheng Cao, Qingliang Miao and Kai Yu　　　　　　　　　　11:00-12:30 (Pier 7&8)

In a task-oriented dialogue system, joint intent detection and slot filling for multi-intent utterances become meaningful since users tend to query more. The current state-of-the-art studies choose to process multi-intent utterances through a single joint model of sequence labelling and multi-label classification, which cannot generalize to utterances with more intents than training samples. Meanwhile, it lacks the ability to assign slots to each corresponding intent. To overcome these problems, we propose a Split-Parsing Method (SPM) for joint multiple intent detection and slot filling, which is a two-stage method. It first splits an input sentence into multiple sub-sentences which contain a single-intent, and then a joint single intent detection and slot filling model is applied to parse each sub-sentence recurrently. Finally, we integrate the parsed results. The sub-sentence split task is also treated as a sequence labelling problem with only one entity-label, which can effectively

generalize to a sentence with more intents unseen in the training set. Experimental results on three multi-intent datasets show that our method obtains substantial improvements over different baselines.

[Industry] FashionKLIP: Enhancing E-Commerce Image-Text Retrieval with Fashion Multi-Modal Conceptual Knowledge Graph
Xiaodan Wang, Chengyu Wang, Lei Li, Zhixu Li, Ben Chen, Linbo Jin, Jun Huang, Yanghua Xiao and Ming Gao   11:00-12:30 (Pier 7&8)
Image-text retrieval is a core task in the multi-modal domain, which arises a lot of attention from both research and industry communities. Recently, the booming of visual-language pre-trained (VLP) models has greatly enhanced the performance of cross-modal retrieval. However, the fine-grained interactions between objects from different modalities are far from well-established. This issue becomes more severe in the e-commerce domain, which lacks sufficient training data and fine-grained cross-modal knowledge. To alleviate the problem, this paper proposes a novel e-commerce knowledge-enhanced VLP model FashionKLIP. We first automatically establish a multi-modal conceptual knowledge graph from large-scale e-commerce image-text data, and then inject the prior knowledge into the VLP model to align across modalities at the conceptual level. The experiments conducted on a public benchmark dataset demonstrate that FashionKLIP effectively enhances the performance of e-commerce image-text retrieval upon state-of-the-art VLP models by a large margin. The application of the method in real industrial scenarios also proves the feasibility and efficiency of FashionKLIP.

[Industry] SaFER: A Robust and Efficient Framework for Fine-tuning BERT-based Classifier with Noisy Labels
Zhenting Qi, Xiaoyu Tan, Chao Qu, Yinghui Xu and Yuan Qi   11:00-12:30 (Pier 7&8)
Learning on noisy datasets is a challenging problem when pre-trained language models are applied to real-world text classification tasks. In numerous industrial applications, acquiring task-specific datasets with 100% accurate labels is difficult, thus many datasets are accompanied by label noise at different levels. Previous work has shown that existing noise-handling methods could not improve the peak performance of BERT on noisy datasets, and might even deteriorate it. In this paper, we propose SaFER, a robust and efficient fine-tuning framework for BERT-based text classifiers, combating label noises without access to any clean data for training or validation. Utilizing a label-agnostic early-stopping strategy and self-supervised learning, our proposed framework achieves superior performance in terms of both accuracy and speed on multiple text classification benchmarks. The trained model is finally fully deployed in several industrial biomedical literature mining tasks and demonstrates high effectiveness and efficiency.

[Industry] Event-Centric Query Expansion in Web Search
Yanan Zhang, Weijie Cui, Yangfan Zhang, Xiaoling Bai, Zhe Zhang, Jin Ma, Xiang Chen and Tianhua Zhou   11:00-12:30 (Pier 7&8)
In search engines, query expansion (QE) is a crucial technique to improve search experience. Previous studies often rely on long-term search log mining, which leads to slow updates and is sub-optimal for time-sensitive news searches. In this work, we present Event-Centric Query Expansion (EQE), the QE system used in a famous Chinese search engine. EQE utilizes a novel event retrieval framework that consists of four stages, i.e., event collection, event reformulation, semantic retrieval and online ranking, which can select the best expansion from a significant amount of potential events rapidly and accurately. Specifically, we first collect and filter news headlines from websites. Then we propose a generation model that incorporates contrastive learning and prompt-tuning techniques to reformulate these headlines to concise candidates. Additionally, we fine-tune a dual-tower semantic model to serve as an encoder for event retrieval and explore a two-stage contrastive training approach to enhance the accuracy of event retrieval. Finally, we rank the retrieved events and select the optimal one as QE, which is then used to improve the retrieval of even-related documents. Through offline analysis and online A/B testing, we observed that the EQE system has significantly improved many indicators compared to the baseline. The system has been deployed in a real production environment and serves hundreds of millions of users.

[Industry] Scalable and Safe Remediation of Defective Actions in Self-Learning Conversational Systems
Sarthak Ahuja, Mohammad Kachuee, Fatemeh Sheikholeslami, Weiqing Liu and Jaeyoung Do   11:00-12:30 (Pier 7&8)
Off-Policy reinforcement learning has been the driving force for the state-of-the-art conversational AIs leading to more natural human-agent interactions and improving the user satisfaction for goal-oriented agents. However, in large-scale commercial settings, it is often challenging to balance between policy improvements and experience continuity on the broad spectrum of applications handled by such system. In the literature, off-policy evaluation and guard-railing on aggregate statistics has been commonly used to address this problem. In this paper, we propose method for curating and leveraging high-precision samples sourced from historical regression incident reports to validate, safe-guard, and improve policies prior to the online deployment. We conducted extensive experiments using data from a real-world conversational system and actual regression incidents. The proposed method is currently deployed in our production system to protect customers against broken experiences and enable long-term policy improvements.

[Industry] Content Moderation for Evolving Policies using Binary Question Answering
Sankha Subhra Mullick, Mohan Bhambhani, Suhit Sinha, Akshat Mathur, Somya Gupta and Jidnya Shah   11:00-12:30 (Pier 7&8)
Content moderation on social media is governed by policies that are intricate and frequently updated with evolving world events. However, automated content moderation systems often restrict easy adaptation to policy changes and are expected to learn policy intricacies from limited amounts of labeled data, which make effective policy compliance challenging. We propose to model content moderation as a binary question answering problem where the questions validate the loosely coupled themes constituting a policy. A decision logic is applied on top to aggregate the theme-specific validations. This way the questions pass theme information to a transformer network as explicit policy prompts, that in turn enables explainability. This setting further allows for faster adaptation to policy updates by leveraging zero-shot capabilities of pre-trained transformers. We showcase improved recall for our proposed method at 95\% precision on two proprietary datasets of social media posts and comments respectively annotated under curated Hate Speech and Commercial Spam policies.

[Industry] Building Accurate Low Latency ASR for Streaming Voice Search in E-commerce
Abhinav Goyal and Nikesh Garera   11:00-12:30 (Pier 7&8)
Automatic Speech Recognition (ASR) is essential for any voice-based application. The streaming capability of ASR becomes necessary to provide immediate feedback to the user in applications like Voice Search. LSTM/RNN and CTC based ASR systems are very simple to train and deploy for low latency streaming applications but have lower accuracy when compared to the state-of-the-art models. In this work, we build accurate LSTM, attention and CTC based streaming ASR models for large-scale Hinglish (blend of Hindi and English) Voice Search. We evaluate various modifications in vanilla LSTM training improve the system's accuracy while preserving the streaming capabilities. We also discuss a simple integration of end-of-speech (EOS) detection with CTC models, which helps reduce the overall search latency. Our model achieves a word error rate (WER) of 3.69% without EOS and 4.78% with EOS, with ~1300 ms (~46.64%) reduction in latency.

[Industry] Consistent Text Categorization using Data Augmentation in e-Commerce
Noa Avigdor, Guy Horowitz, Ariel Raviv and Stav Yanovsky Daye   11:00-12:30 (Pier 7&8)
The categorization of massive e-Commerce data is a crucial, well-studied task, which is prevalent in industrial settings. In this work, we aim to improve an existing product categorization model that is already in use by a major web company, serving multiple applications. At its core, the product categorization model is a text classification model that takes a product title as an input and outputs the most suitable category out of thousands of available candidates. Upon a closer inspection, we found inconsistencies in the labeling of similar items. For example, minor modifications of the product title pertaining to colors or measurements majorly impacted the model's output. This phenomenon can

negatively affect downstream recommendation or search applications, leading to a sub-optimal user experience.

To address this issue, we propose a new framework for consistent text categorization. Our goal is to improve the model's consistency while maintaining its production-level performance. We use a semi-supervised approach for data augmentation and presents two different methods for utilizing unlabeled samples. One method relies directly on existing catalogs, while the other uses a generative model. We compare the pros and cons of each approach and present our experimental results.

[Industry] Transferable and Efficient: Unifying Dynamic Multi-Domain Product Categorization
Shansan Gong, Zelin Zhou, Shuo Wang, Fengjiao Chen, Xiujie Song, Xuezhi Cao, Yunsen Xian and Kenny Zhu   11:00-12:30 (Pier 7&8)
As e-commerce platforms develop different business lines, a special but challenging product categorization scenario emerges, where there are multiple domain-specific category taxonomies and each of them evolves dynamically over time. In order to unify the categorization process and ensure efficiency, we propose a two-stage taxonomy-agnostic framework that relies solely on calculating the semantic relatedness between product titles and category names in the vector space. To further enhance domain transferability and better exploit cross-domain data, we design two plug-in modules: a heuristic mapping scorer and a pretrained contrastive ranking module with the help of meta concepts, which represent keyword knowledge shared across domains. Comprehensive offline experiments show that our method outperforms strong baselines on three dynamic multi-domain product categorization (DMPC) tasks, and online experiments reconfirm its efficacy with a 5% increase on seasonal purchase revenue. Related datasets will be released.

[Industry] Rapid Diffusion: Building Domain-Specific Text-to-Image Synthesizers with Fast Inference Speed
*Bingyan Liu, Weifeng Lin, Zhongjie Duan, Chengyu Wang, Wu Ziheng, Zhang Zipeng, Kui Jia, Lianwen Jin, Cen Chen and Jun Huang*11:00-12:30 (Pier 7&8)
Text-to-Image Synthesis (TIS) aims to generate images based on textual inputs. Recently, several large pre-trained diffusion models have been released to create high-quality images with pre-trained text encoders and diffusion-based image synthesizers. However, popular diffusion-based models from the open-source community cannot support industrial domain-specific applications due to the lack of entity knowledge and low inference speed. In this paper, we propose Rapid Diffusion, a novel framework for training and deploying super-resolution, text-to-image latent diffusion models with rich entity knowledge injected and optimized networks. Furthermore, we employ BladeDISC, an end-to-end Artificial Intelligence (AI) compiler, and FlashAttention techniques to optimize computational graphs of the generated models for online deployment. Experiments verify the effectiveness of our approach in terms of image quality and inference speed. In addition, we present industrial use cases and integrate Rapid Diffusion to an AI platform to show its practical values.

[TACL] Less is More: Mitigate Spurious Correlations for Open-Domain Dialogue Response Generation Models by Causal Discovery
Tao Feng, Lizhen Qu and Gholamreza Haffari   11:00-12:30 (Pier 7&8)
In this paper, we conduct the first study on spurious correlations for open-domain response generation models based on a corpus CGDIALOG curated in our work. The current models indeed suffer from spurious correlations and have a tendency of generating irrelevant and generic responses. Inspired by causal discovery algorithms, we propose a novel model-agnostic method for training and inference of response generation model using a conditional independence classifier. The classifier is trained by a constrained self-training method, coined CONSTRAIN, to overcome data scarcity. The experimental results based on both human and automatic evaluation show that our method significantly outperforms the competitive baselines in terms of relevance, informativeness, and fluency.

[TACL] Bridging the Gap between Synthetic and Natural Questions via Sentence Decomposition for Semantic Parsing
Yilin Niu, Fei Huang and Minlie Huang   11:00-12:30 (Pier 7&8)
Semantic parsing maps natural language questions into logical forms, which can be executed against a knowledge base for answers. In real-world applications, the performance of a parser is often limited by the lack of training data. To facilitate zero-shot learning, data synthesis have been widely studied to automatically generate paired questions and logical forms. However, the data synthesis methods can hardly cover the diverse structures in natural languages, leading to a large gap in sentence structure between synthetic and natural questions. In this paper, we propose a decomposition-based method to unify the sentence structures of questions, which benefits the generalization to the natural questions. Experiments demonstrate that our method significantly improves the semantic parser trained on synthetic data (+7.9

[SRW] Detection and Comparison of Abusive and Hate Speech in English and Hinglish with Emojis Using Deep Learning and Non-Deep Learning Techniques
Sneha P, Sneha Bhaskara, Srishti Seth, Stuti Mohanty and Preet Kanwal   11:00-12:30 (Pier 7&8)
Automated Identification of Multilingual Abusive Content on Social Media

[SRW] Towards Efficient Dialogue Processing in the Emergency Response Domain
Tatiana Anikina   11:00-12:30 (Pier 7&8)
This paper is about dialogue act classification and slot tagging in the emergency response domain.

[SRW] Authorship Attribution of Late 19th Century Novels using GAN-BERT
Kanishka Silva, Burcu Can, Frédéric Blain, Raheem Sarwar, Laura Ugolini and Ruslan Mitkov   11:00-12:30 (Pier 7&8)
This paper is about performing authorship attribution of long 19th-century novels using the GAN-BERT model, comparing author counts, author combinations and sample text sizes.

[SRW] Combining Tradition with Modernness: Exploring Event Representations in Vision-and-Language Models for Visual Goal-Step Inference
Chong Shen and Carina Silberer   11:00-12:30 (Pier 7&8)
This paper studies various methods and their effects on multimodal procedural knowledge understanding of injecting the early shallow017 event representations to nowadays multimodal deep learning-based models.

[SRW] Moral Mimicry: Large Language Models Produce Moral Rationalizations Tailored to Political Identity
Gabriel Simmons   11:00-12:30 (Pier 7&8)
This paper studies whether LLMs moral preferences based on prompted political ideology replicate known results obtained in social science studies, using tools from Moral Foundations Theory

[SRW] Math Word Problem Solving by Generating Linguistic Variants of Problem Statements
Syed Rifat Raiyan, Md Nafis Faiyaz, Shah Md. Jawad Kabir, Mohsinul Kabir, Hasan Mahmud and Md Kamrul Hasan   11:00-12:30 (Pier 7&8)
In order to ameliorate the issue of spurious correlation, in this paper, we propose a framework for MWP solvers based on the generation of linguistic variants of the problem text and introduce a dataset containing paraphrased, adversarial, and inverse variants of the MWPs.

[Demo] MetaPro Online: A Computational Metaphor Processing Online System
Erik Cambria, Mengshi Ge, Kai He, Xiao Li and Rui Mao   11:00-12:30 (Pier 7&8)

Metaphoric expressions are a special linguistic phenomenon, frequently appearing in everyday language. Metaphors do not take their literal meanings in contexts, which may cause obstacles for language learners to understand them. Metaphoric expressions also reflect the cognition of humans via concept mappings, attracting great attention from cognitive science and psychology communities. Thus, we aim to develop a computational metaphor processing online system, termed MetaPro Online, that allows users without a coding background, e.g., language learners and linguists, to easily query metaphoricity labels, metaphor paraphrases, and concept mappings for non-domain-specific text. The outputs of MetaPro can be directly used by language learners and natural language processing downstream tasks because MetaPro is an end-to-end system.

[Demo] A System for Answering Simple Questions in Multiple Languages
Alexander Panchenko, Pavel Braslavski, Valentin Malykh, Mikhail Salnikov and Anton Razzhigaev   11:00-12:30 (Pier 7&8)
Our research focuses on the most prevalent type of queries— simple questions —exemplified by questions like "What is the capital of France?". These questions reference an entity such as "France", which is directly connected (one hop) to the answer entity "Paris" in the underlying knowledge graph (KG). We propose a multilingual Knowledge Graph Question Answering (KGQA) technique that orders potential responses based on the distance between the question's text embeddings and the answer's graph embeddings. A system incorporating this novel method is also described in our work.

Through comprehensive experimentation using various English and multilingual datasets and two KGs — Freebase and Wikidata — we illustrate the comparative advantage of the proposed method across diverse KG embeddings and languages. This edge is apparent even against robust baseline systems, including seq2seq QA models, search-based solutions and intricate rule-based pipelines. Interestingly, our research underscores that even advanced AI systems like ChatGPT encounter difficulties when tasked with answering simple questions. This finding emphasizes the relevance and effectiveness of our approach, which consistently outperforms such systems. We are making the source code and trained models from our study publicly accessible to promote further advancements in multilingual KGQA.

[Demo] Self-Supervised Sentence Polishing by Adding Engaging Modifiers
Minlie Huang, Bo Liu, Yu Ran, Xin Cui, Jian Guan and Zhexin Zhang   11:00-12:30 (Pier 7&8)
Teachers often guide students to improve their essays by adding engaging modifiers to polish the sentences. In this work, we present the first study on automatic sentence polishing by adding modifiers. Since there is no available dataset for the new task, we first automatically construct a large number of parallel data by removing modifiers in the engaging sentences collected from public resources. Then we fine-tune LongLM to reconstruct the original sentences from the corrupted ones. Considering that much overlap between inputs and outputs may bias the model to completely copy the inputs, we split each source sentence into sub-sentences and only require the model to generate the modified sub-sentences. Furthermore, we design a retrieval augmentation algorithm to prompt the model to add suitable modifiers. Automatic and manual evaluation on the auto-constructed test set and real human texts show that our model can generate more engaging sentences with suitable modifiers than strong baselines while keeping fluency. We deploy the model at http://coai.cs.tsinghua.edu.cn/static/polishSent/. A demo video is available at https://youtu.be/Y6gFHOgSv8Y.

[Demo] Japanese-to-English Simultaneous Dubbing Prototype
Eiichiro Sumita, Masao Utiyama and Xiaolin Wang   11:00-12:30 (Pier 7&8)
Live video streaming has become an important form of communication such as virtual conferences. However, for cross-language communication in live video streaming, reading subtitles degrades the viewing experience. To address this problem, our simultaneous dubbing prototype translates and replaces the original speech of a live video stream in a simultaneous manner. Tests on a collection of 90 public videos show that our system achieves a low average latency of 11.90 seconds for smooth playback. Our method is general and can be extended to other language pairs.

[Demo] Pipeline for modeling causal beliefs from natural language
Fred Morstatter, Ishaan Verma and John Priniski   11:00-12:30 (Pier 7&8)
We present a causal language analysis pipeline that leverages a Large Language Model to identify causal claims made in natural language documents, and aggregates claims across a corpus to produce a causal claim network. The pipeline then applies a clustering algorithm that groups causal claims based on their semantic topics. We demonstrate the pipeline by modeling causal belief systems surrounding the Covid-19 vaccine from tweets.

Exploiting Hierarchically Structured Categories in Fine-grained Chinese Named Entity Recognition
Jiuding Yang, Jinwen Luo, Weidong Guo, Di Niu and Yu Xu   11:00-12:30 (Pier 7&8)
Chinese Named Entity Recognition (CNER) is a widely used technology in various applications. While recent studies have focused on utilizing additional information of the Chinese language and characters to enhance CNER performance, this paper focuses on a specific aspect of CNER known as fine-grained CNER (FG-CNER). FG-CNER involves the use of hierarchical, fine-grained categories (e.g. Person-MovieStar) to label named entities. To promote research in this area, we introduce the FiNE dataset, a dataset for FG-CNER consisting of 30,000 sentences from various domains and containing 67,651 entities in 54 fine-grained flattened hierarchical categories. Additionally, we propose SoftFiNE, a novel approach for FG-CNER that utilizes a custom-designed relevance scoring function based on label structures to learn the potential relevance between different flattened hierarchical labels. Our experimental results demonstrate that the proposed SoftFiNE method outperforms the state-of-the-art baselines on the FiNE dataset. Furthermore, we conduct extensive experiments on three other datasets, including OntoNotes 4.0, Weibo, and Resume, where SoftFiNE achieved state-of-the-art performance on all three datasets.

Exploring the Capacity of Pretrained Language Models for Reasoning about Actions and Change
Weinan He, Canming Huang, Zhanhao Xiao and Yongmei Liu   11:00-12:30 (Pier 7&8)
Reasoning about actions and change (RAC) is essential to understand and interact with the ever-changing environment. Previous AI research has shown the importance of fundamental and indispensable knowledge of actions, i.e., preconditions and effects. However, traditional methods rely on logical formalization which hinders practical applications. With recent transformer-based language models (LMs), reasoning over text is desirable and seemingly feasible, leading to the question of whether LMs can effectively and efficiently learn to solve RAC problems. We propose four essential RAC tasks as a comprehensive textual benchmark and generate problems in a way that minimizes the influence of other linguistic requirements (e.g., grounding) to focus on RAC. The resulting benchmark, TRAC, encompassing problems of various complexities, facilitates a more granular evaluation of LMs, precisely targeting the structural generalization ability much needed for RAC. Experiments with three high-performing transformers indicate that additional efforts are needed to tackle challenges raised by TRAC.

Personality Understanding of Fictional Characters during Book Reading
Mo Yu, Jiangnan Li, Shunyu Yao, Wenjie Pang, Xiaochen Zhou, Zhou Xiao, Fandong Meng and Jie Zhou   11:00-12:30 (Pier 7&8)
Comprehending characters' personalities is a crucial aspect of story reading. As readers engage with a story, their understanding of a character evolves based on new events and information; and multiple fine-grained aspects of personalities can be perceived. This leads to a natural problem of situated and fine-grained personality understanding. The problem has not been studied in the NLP field, primarily due to the lack of appropriate datasets mimicking the process of book reading. We present the first labeled dataset PersoNet for this problem. Our novel annotation strategy involves annotating user notes from online reading apps as a proxy for the original books. Experiments and human

studies indicate that our dataset construction is both efficient and accurate; and our task heavily relies on long-term context to achieve accurate predictions for both machines and humans.

Correction of Errors in Preference Ratings from Automated Metrics for Text Generation

Jan Deriu, Pius von Däniken, Don Tuggener and Mark Cieliebak   11:00-12:30 (Pier 7&8)

A major challenge in the field of Text Generation is evaluation: Human evaluations are cost-intensive, and automated metrics often display considerable disagreements with human judgments. In this paper, we propose to apply automated metrics for Text Generation in a preference-based evaluation protocol. The protocol features a statistical model that incorporates various levels of uncertainty to account for the error-proneness of the metrics. We show that existing metrics are generally over-confident in assigning significant differences between systems. As a remedy, the model allows to combine human ratings with automated ratings. We show that it can reduce the required amounts of human ratings to arrive at robust and statistically significant results by more than 50%, while yielding the same evaluation outcome as the pure human evaluation in 95% of cases. We showcase the benefits of the evaluation protocol for three text generation tasks: dialogue systems, machine translation, and text summarization.

HeGeL: A Novel Dataset for Geo-Location from Hebrew Text

Tzuf Paz-Argaman, Tal Bauman, Itai Mondshine, Itzhak Omer, Sagi Dalyot and Reut Tsarfaty   11:00-12:30 (Pier 7&8)

The task of textual geolocation — retrieving the coordinates of a place based on a free-form language description — calls for not only grounding but also natural language understanding and geospatial reasoning. Even though there are quite a few datasets in English used for geolocation, they are currently based on open-source data (Wikipedia and Twitter), where the location of the described place is mostly implicit, such that the location retrieval resolution is limited. Furthermore, there are no datasets available for addressing the problem of textual geolocation in morphologically rich and resource-poor languages, such as Hebrew. In this paper, we present the Hebrew Geo-Location (HeGeL) corpus, designed to collect literal place descriptions and analyze lingual geospatial reasoning. We crowdsourced 5,649 literal Hebrew place descriptions of various place types in three cities in Israel. Qualitative and empirical analysis show that the data exhibits abundant use of geospatial reasoning and requires a novel environmental representation.

RQUGE: Reference-Free Metric for Evaluating Question Generation by Answering the Question

Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan, James Henderson and Marzieh Saeidi   11:00-12:30 (Pier 7&8)

Existing metrics for evaluating the quality of automatically generated questions such as BLEU, ROUGE, BERTScore, and BLEURT compare the reference and predicted questions, providing a high score when there is a considerable lexical overlap or semantic similarity between the candidate and the reference questions. This approach has two major shortcomings. First, we need expensive human-provided reference questions. Second, it penalises valid questions that may not have high lexical or semantic similarity to the reference questions. In this paper, we propose a new metric, RQUGE, based on the answerability of the candidate question given the context. The metric consists of a question-answering and a span scorer modules, using pre-trained models from existing literature, thus it can be used without any further training. We demonstrate that RQUGE has a higher correlation with human judgment without relying on the reference question. Additionally, RQUGE is shown to be more robust to several adversarial corruptions. Furthermore, we illustrate that we can significantly improve the performance of QA models on out-of-domain datasets by fine-tuning on synthetic data generated by a question generation model and reranked by RQUGE.

C-XNLI: Croatian Extension of XNLI Dataset

Leo Obadić, Andrej Jertec, Marko Rajnović and Branimir Dropuljić   11:00-12:30 (Pier 7&8)

Comprehensive multilingual evaluations have been encouraged by emerging cross-lingual benchmarks and constrained by existing parallel datasets. To partially mitigate this limitation, we extended the Cross-lingual Natural Language Inference (XNLI) corpus with Croatian. The development and test sets were translated by a professional translator, and we show that Croatian is consistent with other XNLI dubs. The train set is translated using Facebook's 1.2B parameter m2m_100 model. We thoroughly analyze the Croatian train set and compare its quality with the existing machine-translated German set. The comparison is based on 2000 manually scored sentences per language using a variant of the Direct Assessment (DA) score commonly used at the Conference on Machine Translation (WMT). Our findings reveal that a less-resourced language like Croatian is still lacking in translation quality of longer sentences compared to German. However, both sets have a substantial amount of poor quality translations, which should be considered in translation-based training or evaluation setups.

ANALOGICAL - A Novel Benchmark for Long Text Analogy Evaluation in Large Language Models

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal Gajera, Shreeyash Mukul Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth and Amitava Das   11:00-12:30 (Pier 7&8)

Over the past decade, analogies, in the form of word-level analogies, have played a significant role as an intrinsic measure of evaluating the quality of word embedding methods such as word2vec. Modern large language models (LLMs), however, are primarily evaluated on extrinsic measures based on benchmarks such as GLUE and SuperGLUE, and there are only a few investigations on whether LLMs can draw analogies between long texts. In this paper, we present ANALOGICAL, a new benchmark to intrinsically evaluate LLMs across a taxonomy of analogies of long text with six levels of complexity – (i) word, (ii) word vs. sentence, (iii) syntactic, (iv) negation, (v) entailment, and (vi) metaphor. Using thirteen datasets and three different distance measures, we evaluate the abilities of eight LLMs in identifying analogical pairs in the semantic vector space. Our evaluation finds that it is increasingly challenging for LLMs to identify analogies when going up the analogy taxonomy.

ORCA: A Challenging Benchmark for Arabic Language Understanding

AbdelRahim Elmadany, ElMoatez Billah Nagoudi and Muhammad Abdul-Mageed   11:00-12:30 (Pier 7&8)

Due to the crucial role pretrained language models play in modern NLP, several benchmarks have been proposed to evaluate their performance. In spite of these efforts, no public benchmark of diverse nature currently exists for evaluating Arabic NLU. This makes it challenging to measure progress for both Arabic and multilingual language models. This challenge is compounded by the fact that any benchmark targeting Arabic needs to take into account the fact that Arabic is not a single language but rather a collection of languages and language varieties. In this work, we introduce a publicly available benchmark for Arabic language understanding evaluation dubbed ORCA. It is carefully constructed to cover diverse Arabic varieties and a wide range of challenging Arabic understanding tasks exploiting 60 different datasets (across seven NLU task clusters). To measure current progress in Arabic NLU, we use ORCA to offer a comprehensive comparison between 18 multilingual and Arabic language models. We also provide a public leaderboard with a unified single-number evaluation metric (ORCA score) to facilitate future research.

FACTUAL: A Benchmark for Faithful and Consistent Textual Scene Graph Parsing

Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji and Quan Hung Tran 11:00-12:30 (Pier 7&8)

Textual scene graph parsing has become increasingly important in various vision-language applications, including image caption evaluation and image retrieval. However, existing scene graph parsers that convert image captions into scene graphs often suffer from two types of errors. First, the generated scene graphs fail to capture the true semantics of the captions or the corresponding images, resulting in a lack of faithfulness. Second, the generated scene graphs have high inconsistency, with the same semantics represented by different annotations.

To address these challenges, we propose a novel dataset, which involves re-annotating the captions in Visual Genome (VG) using a new intermediate representation called FACTUAL-MR. FACTUAL-MR can be directly converted into faithful and consistent scene graph annotations. Our experimental results clearly demonstrate that the parser trained on our dataset outperforms existing approaches in terms of faithfulness and consistency. This improvement leads to a significant performance boost in both image caption evaluation and zero-shot image retrieval tasks. Furthermore, we introduce a novel metric for measuring scene graph similarity, which, when combined with the improved scene graph parser, achieves state-of-the-art (SOTA) results on multiple benchmark datasets for the aforementioned tasks.

An Annotated Dataset for Explainable Interpersonal Risk Factors of Mental Disturbance in Social Media Posts
Muskan Garg, Amirmohammad Shahbandegan, Amrit Chadha and Vijay Mago 11:00-12:30 (Pier 7&8)
With a surge in identifying suicidal risk and its severity in social media posts, we argue that a more consequential and explainable research is required for optimal impact on clinical psychology practice and personalized mental healthcare. The success of computational intelligence techniques for inferring mental illness from social media resources, points to natural language processing as a lens for determining Interpersonal Risk Factors (IRF) in human writings. Motivated with limited availability of datasets for social NLP research community, we construct and release a new annotated dataset with human-labelled explanations and classification of IRF affecting mental disturbance on social media: (i) Thwarted Belongingness (TBe), and (ii) Perceived Burdensomeness (PBu). We establish baseline models on our dataset facilitating future research directions to develop real-time personalized AI models by detecting patterns of TBe and PBu in emotional spectrum of user's historical social media profile.

Task-aware Retrieval with Instructions
Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi and Wen-tau Yih 11:00-12:30 (Pier 7&8)
We study the problem of retrieval with instructions, where users provide explicit descriptions of their intent along with their queries to guide a retrieval system. Our solution is a general-purpose task-aware retrieval system, trained using multi-task instruction tuning and can follow human-written instructions to find relevant documents to a given query. We introduce the first large-scale collection of 37 retrieval datasets with instructions, BERRI, and present TART, a single multi-task retrieval system trained on BERRI with instructions that can adapt to a new task without any parameter updates. TART advances the state of the art on two zero-shot retrieval benchmarks, BEIR and LOTTE, outperforming models up to three times larger. We further introduce a new evaluation setup, $X^{\hat{2}}$-Retrieval, to better reflect real-world scenarios in which diverse domains and tasks are pooled. TART significantly outperforms competitive baselines in this setup, further highlighting the effectiveness of guiding retrieval with instructions.

Aggregating Multiple Heuristic Signals as Supervision for Unsupervised Automated Essay Scoring
Cong Wang, Zhiwei Jiang, Yafeng Yin, Zifeng Cheng, Shiping Ge and Qing Gu 11:00-12:30 (Pier 7&8)
Automated Essay Scoring (AES) aims to evaluate the quality score for input essays. In this work, we propose a novel unsupervised AES approach ULRA, which does not require groundtruth scores of essays for training. The core idea of our ULRA is to use multiple heuristic quality signals as the pseudo-groundtruth, and then train a neural AES model by learning from the aggregation of these quality signals. To aggregate these inconsistent quality signals into a unified supervision, we view the AES task as a ranking problem, and design a special Deep Pairwise Rank Aggregation (DPRA) loss for training. In the DPRA loss, we set a learnable confidence weight for each signal to address the conflicts among signals, and train the neural AES model in a pairwise way to disentangle the cascade effect among partial-order pairs. Experiments on eight prompts of ASPA dataset show that ULRA achieves the state-of-the-art performance compared with previous unsupervised methods in terms of both transductive and inductive settings. Further, our approach achieves comparable performance with many existing domain-adapted supervised models, showing the effectiveness of ULRA. The code is available at https://github.com/tenvence/ulra.

Songs Across Borders: Singable and Controllable Neural Lyric Translation
Longshen Ou, Xichu Ma, Min-Yen Kan and Ye Wang 11:00-12:30 (Pier 7&8)
The development of general-domain neural machine translation (NMT) methods has advanced significantly in recent years, but the lack of naturalness and musical constraints in the outputs makes them unable to produce singable lyric translations. This paper bridges the singability quality gap by formalizing lyric translation into a constrained translation problem, converting theoretical guidance and practical techniques from translatology literature to prompt-driven NMT approaches, exploring better adaptation methods, and instantiating them to an English-Chinese lyric translation system. Our model achieves 99.85%, 99.00%, and 95.52% on length accuracy, rhyme accuracy, and word boundary recall. In our subjective evaluation, our model shows a 75% relative enhancement on overall quality, compared against naive fine-tuning (Code available at https://github.com/Sonata165/ControllableLyricTranslation).

Selecting Better Samples from Pre-trained LLMs: A Case Study on Question Generation
Xingdi Yuan, Tong Wang, Yen-Hsiang Wang, Emery Fine, Rania Abdelghani, Hélène Sauzéon and Pierre-Yves Oudeyer 11:00-12:30 (Pier 7&8)
Large Language Models (LLMs) have in recent years demonstrated impressive prowess in natural language generation. A common practice to improve generation diversity is to sample multiple outputs from the model. However, partly due to the inaccessibility of LLMs, there lacks a simple and robust way of selecting the best output from these stochastic samples. As a case study framed in the context of question generation, we propose two prompt-based approaches, namely round-trip and prompt-based score, to selecting high-quality questions from a set of LLM-generated candidates. Our method works without the need to modify the underlying model, nor does it rely on human-annotated references — both of which are realistic constraints for real-world deployment of LLMs. With automatic as well as human evaluations, we empirically demonstrate that our approach can effectively select questions of higher qualities than greedy generation.

Multilingual Knowledge Graph Completion from Pretrained Language Models with Knowledge Constraints
Ran Song, Shizhu He, Shengxiang Gao, Li Cai, Kang Liu, Zhengtao YU and Jun Zhao 11:00-12:30 (Pier 7&8)
Multilingual Knowledge Graph Completion (mKGC) aim at solving queries in different languages by reasoning a tail entity thus improving multilingual knowledge graphs. Previous studies leverage multilingual pretrained language models (PLMs) and the generative paradigm to achieve mKGC. Although multilingual pretrained language models contain extensive knowledge of different languages, its pretraining tasks cannot be directly aligned with the mKGC tasks. Moreover, the majority of KGs and PLMs currently available exhibit a pronounced English-centric bias. This makes it difficult for mKGC to achieve good results, particularly in the context of low-resource languages. To overcome previous problems, this paper introduces global and local knowledge constraints for mKGC. The former is used to constrain the reasoning of answer entities , while the latter is used to enhance the representation of query contexts. The proposed method makes the pretrained model better adapt to the mKGC task. Experimental results on public datasets demonstrate that our method outperforms the previous SOTA on Hits@1 and Hits@10 by an average of 12.32% and 16.03%, which indicates that our proposed method has significant enhancement on mKGC.

Leveraging Prefix Transfer for Multi-Intent Text Revision
Ruining Chong, Cunliang Kong, Liu Wu, Zhenghao Liu, Ziye Jin, Liner Yang, Yange Fan, Hanghang Fan and Erhong Yang 11:00-12:30 (Pier 7&8)
Text revision is a necessary process to improve text quality. During this process, writers constantly edit texts out of different edit intentions.

Identifying edit intention for a raw text is always an ambiguous work, and most previous work on revision systems mainly focuses on editing texts according to one specific edit intention. In this work, we aim to build a multi-intent text revision system that could revise texts without explicit intent annotation. Our system is based on prefix-tuning, which first gets prefixes for every edit intent, and then trains a prefix transfer module, enabling the system to selectively leverage the knowledge from various prefixes according to the input text. We conduct experiments on the IteraTeR dataset, and the results show that our system outperforms baselines. The system can significantly improve the SARI score with more than 3% improvements, which thrives on the learned editing intention prefixes.

Detecting Adversarial Samples through Sharpness of Loss Landscape
Rui Zheng, Shihan Dou, Yuhao Zhou, Qin Liu, Tao Gui, Qi Zhang, Zhongyu Wei, Xuanjing Huang and Menghan Zhang   11:00-12:30 (Pier 7&8)
Deep neural networks (DNNs) have been proven to be sensitive towards perturbations on input samples, and previous works highlight that adversarial samples are even more vulnerable than normal ones. In this work, this phenomenon is illustrated frWe first show that adversarial samples locate in steep and narrow local minima of the loss landscape (high sharpness) while normal samples, which differs distinctly from adversarial ones, reside in the loss surface that is more flatter (low sharpness).om the perspective of sharpness via visualizing the input loss landscape of models. Based on this, we propose a simple and effective sharpness-based detector to distinct adversarial samples by maximizing the loss increment within the region where the inference sample is located. Considering that the notion of sharpness of a loss landscape is relative, we further propose an adaptive optimization strategy in an attempt to fairly compare the relative sharpness among different samples. Experimental results show that our approach can outperform previous detection methods by large margins (average +6.6 F1 score) for four advanced attack strategies considered in this paper across three text classification tasks.

Prototype-Based Interpretability for Legal Citation Prediction
Chu Fei Luo, Rohan Bhambhoria, Samuel Dahan and Xiaodan Zhu   11:00-12:30 (Pier 7&8)
Deep learning has made significant progress in the past decade, and demonstrates potential to solve problems with extensive social impact. In high-stakes decision making areas such as law, experts often require interpretability for automatic systems to be utilized in practical settings. In this work, we attempt to address these requirements applied to the important problem of legal citation prediction (LCP). We design the task with parallels to the thought-process of lawyers, i.e., with reference to both precedents and legislative provisions. After initial experimental results, we refine the target citation predictions with the feedback of legal experts. Additionally, we introduce a prototype architecture to add interpretability, achieving strong performance while adhering to decision parameters used by lawyers. Our study builds on and leverages the state-of-the-art language processing models for law, while addressing vital considerations for high-stakes tasks with practical societal impact.

KGA: A General Machine Unlearning Framework Based on Knowledge Gap Alignment
Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong and Hongzhi Yin   11:00-12:30 (Pier 7&8)
Recent legislation of the "right to be forgotten" has led to the interest in machine unlearning, where the learned models are endowed with the function to forget information about specific training instances as if they never existed in the training set. Previous work mainly focuses on computer vision scenarios and largely ignores the essentials of unlearning in NLP field, where text data contains more explicit and sensitive personal information than images. In this paper, we propose a general unlearning framework called KGA to induce forgetfulness. Different from previous work that tries to recover gradients or forces models to perform close to one specific distribution, KGA maintains distribution differences (i.e., knowledge gap). This relaxes the distribution assumption. Furthermore, we first apply the unlearning method to various NLP tasks (i.e., classification, translation, response generation) and propose several unlearning evaluation metrics with pertinence. Experiments on large-scale datasets show that KGA yields comprehensive improvements over baselines, where extensive analyses further validate the effectiveness of KGA and provide insight into unlearning for NLP tasks.

Tucker Decomposition with Frequency Attention for Temporal Knowledge Graph Completion
Likang Xiao, Richong Zhang, Zijie Chen and Junfan Chen   11:00-12:30 (Pier 7&8)
Temporal Knowledge Graph Completion aims to complete missing entities or relations under temporal constraints. Previous tensor decomposition-based models for TKGC only independently consider the combination of one single relation with one single timestamp, ignoring the global nature of the embedding. We propose a Frequency Attention (FA) model to capture the global temporal dependencies between one relation and the entire timestamp. Specifically, we use Discrete Cosine Transform (DCT) to capture the frequency of the timestamp embedding and further compute the frequency attention weight to scale embedding. Meanwhile, the previous temporal tucker decomposition method uses a simple norm regularization to constrain the core tensor, which limits the optimization performance. Thus, we propose Orthogonal Regularization (OR) variants for the core tensor, which can limit the non-superdiagonal elements of the 3-rd core tensor. Experiments on three standard TKGC datasets demonstrate that our method outperforms the state-of-the-art results on several metrics. The results suggest that the direct-current component is not the best feature for TKG representation learning. Additional analysis shows the effectiveness of our FA and OR models, even with smaller embedding dimensions.

Towards Diverse and Effective Question-Answer Pair Generation from Children Storybooks
Sugyeong Eo, Hyeonseok Moon, Jinsung Kim, Yuna Hur, Jeongwook Kim, SongEun Lee, Changwoo Chun, SUNGSOO PARK and Heuiseok Lim   11:00-12:30 (Pier 7&8)
Recent advances in QA pair generation (QAG) have raised interest in applying this technique to the educational field. However, the diversity of QA types remains a challenge despite its contributions to comprehensive learning and assessment of children. In this paper, we propose a QAG framework that enhances QA type diversity by producing different interrogative sentences and implicit/explicit answers. Our framework comprises a QFS-based answer generator, an iterative QA generator, and a relevancy-aware ranker. The two generators aim to expand the number of candidates while covering various types. The ranker trained on the in-context negative samples clarifies the top-N outputs based on the ranking score. Extensive evaluations and detailed analyses demonstrate that our approach outperforms previous state-of-the-art results by significant margins, achieving improved diversity and quality. Our task-oriented processes are consistent with real-world demand, which highlights our system's high applicability.

Python Code Generation by Asking Clarification Questions
Haau-Sing (Xiaocheng) Li, Mohsen Mesgar, André Martins and Iryna Gurevych   11:00-12:30 (Pier 7&8)
Code generation from text requires understanding the user's intent from a natural language description and generating an executable code snippet that satisfies this intent. While recent pretrained language models demonstrate remarkable performance for this task, these models fail when the given natural language description is under-specified. In this work, we introduce a novel and more realistic setup for this task. We hypothesize that the under-specification of a natural language description can be resolved by asking clarification questions. Therefore, we collect and introduce a new dataset named CodeClarQA containing pairs of natural language descriptions and code with created synthetic clarification questions and answers. The empirical results of our evaluation of pretrained language model performance on code generation show that clarifications result in more precisely generated code, as shown by the substantial improvement of model performance in all evaluation metrics. Alongside this, our task and dataset introduce new challenges to the community, including when and what clarification questions should be asked. Our code and dataset are available on GitHub.

Dating Greek Papyri with Text Regression

John Pavlopoulos, Maria Konstantinidou, Isabelle Marthot-Santaniello, Holger Essler and Asimina Paparigopoulou 11:00-12:30 (Pier 7&8)
Dating Greek papyri accurately is crucial not only to edit their texts but also to understand numerous other aspects of ancient writing, document and book production and circulation, as well as various other aspects of administration, everyday life and intellectual history of antiquity. Although a substantial number of Greek papyri documents bear a date or other conclusive data as to their chronological placement, an even larger number can only be dated tentatively or in approximation, due to the lack of decisive evidence. By creating a dataset of 389 transcriptions of documentary Greek papyri, we train 389 regression models and we predict a date for the papyri with an average MAE of 54 years and an MSE of 1.17, outperforming image classifiers and other baselines. Last, we release date estimations for 159 manuscripts, for which only the upper limit is known.

Sequential Path Signature Networks for Personalised Longitudinal Language Modeling

Talia Tseriotou, Adam Tsakalidis, Peter Foster, Terence J Lyons and Maria Liakata 11:00-12:30 (Pier 7&8)
Longitudinal user modeling can provide a strong signal for various downstream tasks. Despite the rapid progress in representation learning, dynamic aspects of modelling individuals' language have only been sparsely addressed. We present a novel extension of neural sequential models using the notion of path signatures from rough path theory, which constitute graduated summaries of continuous paths and have the ability to capture non-linearities in trajectories. By combining path signatures of users' history with contextual neural representations and recursive neural networks we can produce compact time-sensitive user representations. Given the magnitude of mental health conditions with symptoms manifesting in language, we show the applicability of our approach on the task of identifying changes in individuals' mood by analysing their online textual content. By directly integrating signature transforms of users' history in the model architecture we jointly address the two most important aspects of the task, namely sequentiality and temporality. Our approach achieves state-of-the-art performance on macro-average F1 score on the two available datasets for the task, outperforming or performing on-par with state-of-the-art models utilising only historical posts and even outperforming prior models which also have access to future posts of users.

Backdooring Neural Code Search

Weisong Sun, Yuchen Chen, Guanhong Tao, Chunrong Fang, Xiangyu Zhang, Quanjun Zhang and Bin Luo 11:00-12:30 (Pier 7&8)
Reusing off-the-shelf code snippets from online repositories is a common practice, which significantly enhances the productivity of software developers. To find desired code snippets, developers resort to code search engines through natural language queries. Neural code search models are hence behind many such engines. These models are based on deep learning and gain substantial attention due to their impressive performance. However, the security aspect of these models is rarely studied. Particularly, an adversary can inject a backdoor in neural code search models, which return buggy or even vulnerable code with security/privacy issues. This may impact the downstream software (e.g., stock trading systems and autonomous driving) and cause financial loss and/or life-threatening incidents. In this paper, we demonstrate such attacks are feasible and can be quite stealthy. By simply modifying one variable/function name, the attacker can make buggy/vulnerable code rank in the top 11%. Our attack BADCODE features a special trigger generation and injection procedure, making the attack more effective and stealthy. The evaluation is conducted on two neural code search models and the results show our attack outperforms baselines by 60%. Our user study demonstrates that our attack is more stealthy than the baseline by two times based on the F1 score.

Bidirectional Transformer Reranker for Grammatical Error Correction

Ying Zhang, Hidetaka Kamigaito and Manabu Okumura 11:00-12:30 (Pier 7&8)
Pre-trained seq2seq models have achieved state-of-the-art results in the grammatical error correction task. However, these models still suffer from a prediction bias due to their unidirectional decoding. Thus, we propose a bidirectional Transformer reranker (BTR), that re-estimates the probability of each candidate sentence generated by the pre-trained seq2seq model. The BTR preserves the seq2seq-style Transformer architecture but utilizes a BERT-style self-attention mechanism in the decoder to compute the probability of each target token by using masked language modeling to capture bidirectional representations from the target context. For guiding the reranking, the BTR adopts negative sampling in the objective function to minimize the unlikelihood. During inference, the BTR gives final results after comparing the reranked top-1 results with the original ones by an acceptance threshold. Experimental results show that, in reranking candidates from a pre-trained seq2seq model, T5-base, the BTR on top of T5-base could yield 65.47 and 71.27 F0.5 scores on the CoNLL-14 and BEA test sets, respectively, and yield 59.52 GLEU score on the JFLEG corpus, with improvements of 0.36, 0.76 and 0.48 points compared with the original T5-base. Furthermore, when reranking candidates from T5-large, the BTR on top of T5-base improved the original T5-large by 0.26 points on the BEA test set.

TransGEC: Improving Grammatical Error Correction with Translationese

Tao Fang, Xuebo Liu, Derek F. Wong, Runzhe Zhan, Liang Ding, Lidia S. Chao, Dacheng Tao and Min Zhang 11:00-12:30 (Pier 7&8)
Data augmentation is an effective way to improve model performance of grammatical error correction (GEC). This paper identifies a critical side-effect of GEC data augmentation, which is due to the style discrepancy between the data used in GEC tasks (i.e., texts produced by non-native speakers) and data augmentation (i.e., native texts). To alleviate this issue, we propose to use an alternative data source, translationese (i.e., human-translated texts), as input for GEC data augmentation, which 1) is easier to obtain and usually has better quality than non-native texts, and 2) has a more similar style to non-native texts. Experimental results on the CoNLL14 and BEA19 English, NLPCC18 Chinese, Falko-MERLIN German, and RULEC-GEC Russian GEC benchmarks show that our approach consistently improves correction accuracy over strong baselines. Further analyses reveal that our approach is helpful for overcoming mainstream correction difficulties such as the corrections of frequent words, missing words, and substitution errors. Data, code, models and scripts are freely available at https://github.com/NLP2CT/TransGEC.

Cross Encoding as Augmentation: Towards Effective Educational Text Classification

Hyun Seung Lee, Seungtaek Choi, Yunsung Lee, Hyeongdon Moon, Shinhyeok Oh, Myeongho Jeong, Hyojun Go and Christian Wallraven 11:00-12:30 (Pier 7&8)
Text classification in education, usually called auto-tagging, is the automated process of assigning relevant tags to educational content, such as questions and textbooks. However, auto-tagging suffers from a data scarcity problem, which stems from two major challenges: 1) it possesses a large tag space and 2) it is multi-label. Though a retrieval approach is reportedly good at low-resource scenarios, there have been fewer efforts to directly address the data scarcity problem. To mitigate these issues, here we propose a novel retrieval approach CEAA that provides effective learning in educational text classification. Our main contributions are as follows: 1) we leverage transfer learning from question-answering datasets, and 2) we propose a simple but effective data augmentation method introducing cross-encoder style texts to a bi-encoder architecture for more efficient inference. An extensive set of experiments shows that our proposed method is effective in multi-label scenarios and low-resource tags compared to state-of-the-art models.

Causality-Guided Multi-Memory Interaction Network for Multivariate Stock Price Movement Prediction

Di Luo, Weiheng Liao, Shuqi Li, Xin Cheng and Rui Yan 11:00-12:30 (Pier 7&8)
Over the past few years, we've witnessed an enormous interest in stock price movement prediction using AI techniques. In recent literature, auxiliary data has been used to improve prediction accuracy, such as textual news. When predicting a particular stock, we assume that information from other stocks should also be utilized as auxiliary data to enhance performance. In this paper, we propose the Causality-guided Multi-memory Interaction Network (CMIN), a novel end-to-end deep neural network for stock movement prediction which, for the first

time, models the multi-modality between financial text data and causality-enhanced stock correlations to achieve higher prediction accuracy. CMIN transforms the basic attention mechanism into Causal Attention by calculating transfer entropy between multivariate stocks in order to avoid attention on spurious correlations. Furthermore, we introduce a fusion mechanism to model the multi-directional interactions through which CMIN learns not only the self-influence but also the interactive influence in information flows representing the interrelationship between text and stock correlations. The effectiveness of the proposed approach is demonstrated by experiments on three real-world datasets collected from the U.S. and Chinese markets, where CMIN outperforms existing models to establish a new state-of-the-art prediction accuracy.

Learning Non-linguistic Skills without Sacrificing Linguistic Proficiency
Mandar Sharma, Nikhil Muralidhar and Naren Ramakrishnan 11:00-12:30 (Pier 7&8)
The field of Math-NLP has witnessed significant growth in recent years, motivated by the desire to expand LLM performance to the leaning of non-linguistic notions (numerals, and subsequently, arithmetic reasoning). However, non-linguistic skill injection typically comes at a cost for LLMs: it leads to catastrophic forgetting of core linguistic skills, a consequence that often remains unaddressed in the literature. As Math-NLP has been able to create LLMs that can closely approximate the mathematical skills of a grade schooler or the arithmetic reasoning skills of a calculator, the practicality of these models fail if they concomitantly shed their linguistic capabilities. In this work, we take a closer look into the phenomena of catastrophic forgetting as it pertains to LLMs and subsequently offer a novel framework for non-linguistic skill injection for LLMs based on information-theoretic interventions and skill-specific losses that enable the learning of strict arithmetic reasoning. Our model outperforms the state-of-the-art both on injected non-linguistic skills and on linguistic knowledge retention, and does so with a fraction of the non-linguistic training data (1/4) and zero additional synthetic linguistic training data.

Financial Numeric Extreme Labelling: A dataset and benchmarking
Soumya Sharma, Subhendu Khatuya, Manjunath Hegde, Afreen Shaikh, Koustuv Dasgupta, Pawan Goyal and Niloy Ganguly 11:00-12:30 (Pier 7&8)
The U.S. Securities and Exchange Commission (SEC) mandates all public companies to file periodic financial statements that should contain numerals annotated with a particular label from a taxonomy. In this paper, we formulate the task of automating the assignment of a label to a particular numeral span in a sentence from an extremely large label set. Towards this task, we release a dataset, Financial Numeric Extreme Labelling (FNXL), annotated with 2,794 labels. We benchmark the performance of the FNXL dataset by formulating the task as (a) a sequence labelling problem and (b) a pipeline with span extraction followed by Extreme Classification. Although the two approaches perform comparably, the pipeline solution provides a slight edge for the least frequent labels.

Commonsense Knowledge Graph Completion Via Contrastive Pretraining and Node Clustering
Siwei Wu, Xiangqing Shen and Rui Xia 11:00-12:30 (Pier 7&8)
The nodes in the commonsense knowledge graph (CSKG) are normally represented by free-form short text (e.g., word or phrase). Different nodes may represent the same concept. This leads to the problems of edge sparsity and node redundancy, which challenges CSKG representation and completion. On the one hand, edge sparsity limits the performance of graph representation learning; On the other hand, node redundancy makes different nodes corresponding to the same concept have inconsistent relations with other nodes. To address the two problems, we propose a new CSKG completion framework based on Contrastive Pretraining and Node Clustering (CPNC). Contrastive Pretraining constructs positive and negative head-tail node pairs on CSKG and utilizes contrastive learning to obtain better semantic node representation. Node Clustering aggregates nodes with the same concept into a latent concept, assisting the task of CSKG completion. We evaluate our CPNC approach on two CSKG completion benchmarks (CN-100K and ATOMIC), where CPNC outperforms the state-of-the-art methods. Extensive experiments demonstrate that both Contrastive Pretraining and Node Clustering can significantly improve the performance of CSKG completion. The source code of CPNC is publicly available on https://github.com/NUSTM/CPNC.

AtTGen: Attribute Tree Generation for Real-World Attribute Joint Extraction
Yanzeng Li, Bingcong Xue, Ruoyu Zhang and Lei Zou 11:00-12:30 (Pier 7&8)
Attribute extraction aims to identify attribute names and the corresponding values from descriptive texts, which is the foundation for extensive downstream applications such as knowledge graph construction, search engines, and e-Commerce. In previous studies, attribute extraction is generally treated as a classification problem for predicting attribute types or a sequence tagging problem for labeling attribute values, where two paradigms, i.e., closed-world and open-world assumption, are involved. However, both of these paradigms have limitations in terms of real-world applications. And prior studies attempting to integrate these paradigms through ensemble, pipeline, and co-training models, still face challenges like cascading errors, high computational overhead, and difficulty in training. To address these existing problems, this paper presents Attribute Tree, a unified formulation for real-world attribute extraction application, where closed-world, open-world, and semi-open attribute extraction tasks are modeled uniformly. Then a text-to-tree generation model, AtTGen, is proposed to learn annotations from different scenarios efficiently and consistently. Experiments demonstrate that our proposed paradigm well covers various scenarios for real-world applications, and the model achieves state-of-the-art, outperforming existing methods by a large margin on three datasets. Our code, pre-trained model, and datasets are available at https://github.com/lsvih/AtTGen.

Rethinking Masked Language Modeling for Chinese Spelling Correction
Hongqiu Wu, Shaohua Zhang, Yuchen Zhang and Hai Zhao 11:00-12:30 (Pier 7&8)
In this paper, we study Chinese Spelling Correction (CSC) as a joint decision made by two separate models: a language model and an error model. Through empirical analysis, we find that fine-tuning BERT tends to over-fit the error model while under-fit the language model, resulting in poor generalization to out-of-distribution error patterns. Given that BERT is the backbone of most CSC models, this phenomenon has a significant negative impact. To address this issue, we are releasing a multi-domain benchmark LEMON, with higher quality and diversity than existing benchmarks, to allow a comprehensive assessment of the open domain generalization of CSC models. Then, we demonstrate that a very simple strategy – randomly masking 20% non-error tokens from the input sequence during fine-tuning – is sufficient for learning a much better language model without sacrificing the error model. This technique can be applied to any model architecture and achieves new state-of-the-art results on SIGHAN, ECSpell, and LEMON.

Pre-training Multi-party Dialogue Models with Latent Discourse Inference
Yiyang Li, Xinting Huang, Wei Bi and Hai Zhao 11:00-12:30 (Pier 7&8)
Multi-party dialogues are more difficult for models to understand than one-to-one two-party dialogues, since they involve multiple interlocutors, resulting in interweaving reply-to relations and information flows. To step over these obstacles, an effective way is to pre-train a model that understands the discourse structure of multi-party dialogues, namely, to whom each utterance is replying. However, due to the lack of explicitly annotated discourse labels in multi-party dialogue corpora, previous works fail to scale up the pre-training process by putting aside the unlabeled multi-party conversational data for nothing. To fully utilize the unlabeled data, we propose to treat the discourse structures as latent variables, then jointly infer them and pre-train the discourse-aware model by unsupervised latent variable inference methods. Experiments on multiple downstream tasks show that our pre-trained model outperforms strong baselines by large margins and achieves state-of-the-art (SOTA) results, justifying the effectiveness of our method. The official implementation of this paper is available at https://github.com/EricLee8/MPD_EMVI.

AutoConv: Automatically Generating Information-seeking Conversations with Large Language Models
Siheng Li, Cheng Yang, Yichun Yin, Xinyu Zhu, Zesen Cheng, Lifeng Shang, Xin Jiang, Qun Liu and Yujiu Yang   11:00-12:30 (Pier 7&8)
Information-seeking conversation, which aims to help users gather information through conversation, has achieved great progress in recent years. However, the research is still stymied by the scarcity of training data. To alleviate this problem, we propose AutoConv for synthetic conversation generation, which takes advantage of the few-shot learning ability and generation capacity of large language models (LLM). Specifically, we formulate the conversation generation problem as a language modeling task, then finetune an LLM with a few human conversations to capture the characteristics of the information-seeking process and use it for generating synthetic conversations with high quality. Experimental results on two frequently-used datasets verify that AutoConv has substantial improvements over strong baselines and alleviates the dependence on human annotation. In addition, we also provide several analysis studies to promote future research.

A Probabilistic Framework for Discovering New Intents
Yunhua Zhou, Guofeng Quan and Xipeng Qiu   11:00-12:30 (Pier 7&8)
Discovering new intents is of great significance for establishing the Task-Oriented Dialogue System. Most existing methods either cannot transfer prior knowledge contained in known intents or fall into the dilemma of forgetting prior knowledge in the follow-up. Furthermore, these methods do not deeply explore the intrinsic structure of unlabeled data, and as a result, cannot seek out the characteristics that define an intent in general. In this paper, starting from the intuition that discovering intents could be beneficial for identifying known intents, we propose a probabilistic framework for discovering intents where intent assignments are treated as latent variables. We adopt the Expectation Maximization framework for optimization. Specifically, in the E-step, we conduct intent discovery and explore the intrinsic structure of unlabeled data by the posterior of intent assignments. In the M-step, we alleviate the forgetting of prior knowledge transferred from known intents by optimizing the discrimination of labeled data. Extensive experiments conducted on three challenging real-world datasets demonstrate the generality and effectiveness of the proposed framework and implementation.

Facilitating Multi-turn Emotional Support Conversation with Positive Emotion Elicitation: A Reinforcement Learning Approach
Jinfeng Zhou, Zhuang Chen, Bo Wang and Minlie Huang   11:00-12:30 (Pier 7&8)
Emotional support conversation (ESC) aims to provide emotional support (ES) to improve one's mental state. Existing works stay at fitting grounded responses and responding strategies (e.g., *question*), which ignore the effect on ES and lack explicit goals to guide emotional positive transition. To this end, we introduce a new paradigm to formalize multi-turn ESC as a process of positive emotion elicitation. Addressing this task requires finely adjusting the elicitation intensity in ES as the conversation progresses while maintaining conversational goals like coherence. In this paper, we propose SUPPORTER, a mixture-of-expert-based reinforcement learning model, and well design ES and dialogue coherence rewards to guide policy's learning for responding. Experiments verify the superiority of SUPPORTER in achieving positive emotion elicitation during responding while maintaining conversational goals including coherence.

Task-Optimized Adapters for an End-to-End Task-Oriented Dialogue System
Namo Bang, Jeehyun Lee and Myoung-Wan Koo   11:00-12:30 (Pier 7&8)
Task-Oriented Dialogue (TOD) systems are designed to carry out specific tasks by tracking dialogue states and generating appropriate responses to help users achieve defined goals. Recently, end-to-end dialogue models pre-trained based on large datasets have shown promising performance in the conversational system. However, they share the same parameters to train tasks of the dialogue system (NLU, DST, NLG), so debugging each task is challenging. Also, they require a lot of effort to fine-tune large parameters to create a task-oriented chatbot, making it difficult for non-experts to handle. Therefore, we intend to train relatively lightweight and fast models compared to PLM. In this paper, we propose an End-to-end TOD system with Task-Optimized Adapters which learn independently per task, adding only small number of parameters after fixed layers of pre-trained network. We also enhance the performance of the DST and NLG modules through reinforcement learning, overcoming the learning curve that has lacked at the adapter learning and enabling the natural and consistent response generation that is appropriate for the goal. Our method is a model-agnostic approach and does not require prompt-tuning as only input data without a prompt. As results of the experiment, our method shows competitive performance on the MultiWOZ benchmark compared to the existing end-to-end models. In particular, we attain state-of-the-art performance on the DST task of 2.2 dataset.

Dialogue Planning via Brownian Bridge Stochastic Process for Goal-directed Proactive Dialogue
Jian Wang, Dongding Lin and Wenjie Li   11:00-12:30 (Pier 7&8)
Goal-directed dialogue systems aim to proactively reach a pre-determined target through multi-turn conversations. The key to achieving this task lies in planning dialogue paths that smoothly and coherently direct conversations towards the target. However, this is a challenging and under-explored task. In this work, we propose a coherent dialogue planning approach that uses a stochastic process to model the temporal dynamics of dialogue paths. We define a latent space that captures the coherence of goal-directed behavior using a Brownian bridge process, which allows us to incorporate user feedback flexibly in dialogue planning. Based on the derived latent trajectories, we generate dialogue paths explicitly using pre-trained language models. We finally employ these paths as natural language prompts to guide dialogue generation. Our experiments show that our approach generates more coherent utterances and achieves the goal with a higher success rate.

Dual Class Knowledge Propagation Network for Multi-label Few-shot Intent Detection
Feng Zhang, Wei Chen, Fei Ding and Tengjiao Wang   11:00-12:30 (Pier 7&8)
Multi-label intent detection aims to assign multiple labels to utterances and attracts increasing attention as a practical task in task-oriented dialogue systems. As dialogue domains change rapidly and new intents emerge fast, the lack of annotated data motivates multi-label few-shot intent detection. However, previous studies are confused by the identical representation of the utterance with multiple labels and overlook the intrinsic intra-class and inter-class interactions. To address these two limitations, we propose a novel dual class knowledge propagation network in this paper. In order to learn well-separated representations for utterances with multiple intents, we first introduce a label-semantic augmentation module incorporating class name information. For better consideration of the inherent intra-class and inter-class relations, an instance-level and a class-level graph neural network are constructed, which not only propagate label information but also propagate feature structure. And we use a simple yet effective method to predict the intent count of each utterance. Extensive experimental results on two multi-label intent datasets have demonstrated that our proposed method outperforms strong baselines by a large margin.

Bridging The Gap: Entailment Fused-T5 for Open-retrieval Conversational Machine Reading Comprehension
Xiao Zhang, Heyan Huang, Zewen Chi and Xian-Ling Mao   11:00-12:30 (Pier 7&8)
Open-retrieval conversational machine reading comprehension (OCMRC) simulates real-life conversational interaction scenes. Machines are required to make a decision of "Yes/No/Inquire" or generate a follow-up question when the decision is "Inquire" based on retrieved rule texts, user scenario, user question and dialogue history. Recent studies try to reduce the information gap between decision-making and question generation, in order to improve the performance of generation. However, the information gap still persists because these methods are still limited in pipeline framework, where decision-making and question generation are performed separately, making it hard to share the entailment reasoning used in decision-making across all stages. To tackle the above problem, we propose a novel one-stage end-to-end framework, called Entailment Fused-T5 (EFT), to bridge the information gap between decision-making and question generation in a global understanding manner. The extensive experimental results demonstrate that our proposed framework achieves new state-of-the-art performance on the OR-ShARC benchmark. Our model and code are publicly available at an anonymous link.

A Cross-Modality Context Fusion and Semantic Refinement Network for Emotion Recognition in Conversation
Xiaoheng Zhang and Yang Li   11:00-12:30 (Pier 7&8)
Emotion recognition in conversation (ERC) has attracted enormous attention for its applications in empathetic dialogue systems. However, most previous researches simply concatenate multimodal representations, leading to an accumulation of redundant information and a limited context interaction between modalities. Furthermore, they only consider simple contextual features ignoring semantic clues, resulting in an insufficient capture of the semantic coherence and consistency in conversations. To address these limitations, we propose a cross-modality context fusion and semantic refinement network (CMCF-SRNet). Specifically, we first design a cross-modal locality-constrained transformer to explore the multimodal interaction. Second, we investigate a graph-based semantic refinement transformer, which solves the limitation of insufficient semantic relationship information between utterances. Extensive experiments on two public benchmark datasets show the effectiveness of our proposed method compared with other state-of-the-art methods, indicating its potential application in emotion recognition. Our model will be available at https://github.com/zxiaohen/CMCF-SRNet.

Multimodal Recommendation Dialog with Subjective Preference: A New Challenge and Benchmark
Yuxing Long, Binyuan Hui, Caixia Yuan, Fei Huang, Yongbin Li and Xiaojie WANG   11:00-12:30 (Pier 7&8)
Existing multimodal task-oriented dialog data fails to demonstrate the diverse expressions of user subjective preferences and recommendation acts in the real-life shopping scenario. This paper introduces a new dataset SURE (Multimodal Recommendation Dialog with Subjective Preference), which contains 12K shopping dialogs in complex store scenes. The data is built in two phases with human annotations to ensure quality and diversity. SURE is well-annotated with subjective preferences and recommendation acts proposed by sales experts. A comprehensive analysis is given to reveal the distinguishing features of SURE. Three benchmark tasks are then proposed on the data to evaluate the capability of multimodal recommendation agents. Basing on the SURE, we propose a baseline model, powered by a state-of-the-art multimodal model, for these tasks.

TREA: Tree-Structure Reasoning Schema for Conversational Recommendation
Wendi Li, Wei Wei, Xiaoye Qu, Xian-Ling Mao, Ye Yuan, Wenfeng Xie and Dangyang Chen   11:00-12:30 (Pier 7&8)
Conversational recommender systems (CRS) aim to timely trace the dynamic interests of users through dialogues and generate relevant responses for item recommendations. Recently, various external knowledge bases (especially knowledge graphs) are incorporated into CRS to enhance the understanding of conversation contexts. However, recent reasoning-based models heavily rely on simplified structures such as linear structures or fixed-hierarchical structures for causality reasoning, hence they cannot fully figure out sophisticated relationships among utterances with external knowledge. To address this, we propose a novel Tree structure Reasoning schEmA named TREA. TREA constructs a multi-hierarchical scalable tree as the reasoning structure to clarify the causal relationships between mentioned entities, and fully utilizes historical conversations to generate more reasonable and suitable responses for recommended results. Extensive experiments on two public CRS datasets have demonstrated the effectiveness of our approach.

SimOAP: Improve Coherence and Consistency in Persona-based Dialogue Generation via Over-sampling and Post-evaluation
Junkai Zhou, Liang Pang, Huawei Shen and Xueqi Cheng   11:00-12:30 (Pier 7&8)
Language models trained on large-scale corpora can generate remarkably fluent results in open-domain dialogue. However, for the persona-based dialogue generation task, consistency and coherence are also key factors, which are great challenges for language models. Existing works mainly focus on valuable data filtering, model structure modifying, or objective function designing, while their improvements are limited and hard to generalize to all types of pre-trained language models. However, we find that language models can produce consistent and coherent responses if we consider enough generations. Thus, the problems lay in large-scale response generation and target response selection. In this work, a simple but effective two-stage SimOAP strategy is proposed, i.e., over-sampling and post-evaluation. The over-sampling stage takes large-scale responses from existing trained models efficiently via off-the-shelf distilling and compressing methods, and the post-evaluation stage selects a good response based on multiple well-designed evaluation metrics from large-scale candidates. Experimental results show that the proposed plug-in SimOAP strategy improves the backbone models and outperforms the baseline strategies in both automatic and human evaluations.

PAL to Lend a Helping Hand: Towards Building an Emotion Adaptive Polite and Empathetic Counseling Conversational Agent
Kshitij Mishra, Priyanshu Priya and Asif Ekbal   11:00-12:30 (Pier 7&8)
The World Health Organization (WHO) has significantly emphasized the need for mental health care. The social stigma associated with mental illness prevents individuals from addressing their issues and getting assistance. In such a scenario, the relevance of online counseling has increased dramatically. The feelings and attitudes that a client and a counselor express towards each other result in a higher or lower counseling experience. A counselor should be friendly and gain clients' trust to make them share their problems comfortably. Thus, it is essential for the counselor to adequately comprehend the client's emotions and ensure client's welfare, i.e. s/he should adapt and deal with the clients politely and empathetically to provide a pleasant, cordial and personalized experience. Motivated by this, in this work, we attempt to build a novel Polite and empAthetic counseLing conversational agent PAL to lay down the counseling support to substance addict and crime victims. To have client's emotion-based polite and empathetic responses, two counseling datasets laying down the counseling support to substance addicts and crime victims are annotated. These annotated datasets are used to build PAL in a reinforcement learning framework. A novel reward function is formulated to ensure correct politeness and empathy preferences as per client's emotions with naturalness and non-repetitiveness in responses. Thorough automatic and human evaluation showcase the usefulness and strength of the designed novel reward function. Our proposed system is scalable and can be easily modified with different modules of preference models as per need.

Imagination is All You Need! Curved Contrastive Learning for Abstract Sequence Modeling Utilized on Long Short-Term Dialogue Planning
Justus-Jonas Erker, Stefan Schaffer and Gerasimos Spanakis   11:00-12:30 (Pier 7&8)
Inspired by the curvature of space-time, we introduce Curved Contrastive Learning (CCL), a novel representation learning technique for learning the relative turn distance between utterance pairs in multi-turn dialogues. The resulting bi-encoder models can guide transformers as a response ranking model towards a goal in a zero-shot fashion by projecting the goal utterance and the corresponding reply candidates into a latent space. Here the cosine similarity indicates the distance/reachability of a candidate utterance toward the corresponding goal. Furthermore, we explore how these forward-entailing language representations can be utilized for assessing the likelihood of sequences by the entailment strength i.e. through the cosine similarity of its individual members (encoded separately) as an emergent property in the curved space. These non-local properties allow us to imagine the likelihood of future patterns in dialogues, specifically by ordering/identifying future goal utterances that are multiple turns away, given a dialogue context. As part of our analysis, we investigate characteristics that make conversations (un)plannable and find strong evidence of planning capability over multiple turns (in 61.56% over 3 turns) in conversations from the DailyDialog dataset. Finally, we show how we achieve higher efficiency in sequence modeling tasks compared to previous work thanks to our relativistic approach, where only the last utterance needs to be encoded and computed during inference.

On the Compositional Generalization in Versatile Open-domain Dialogue
Tingchen Fu, Xueliang Zhao, Lemao Liu and Rui Yan   11:00-12:30 (Pier 7&8)

Previous research has demonstrated the potential of multi-task learning to foster a conversational agent's ability to acquire a variety of skills. However, these approaches either suffer from interference among different datasets (also known as negative transfer), or fail to effectively reuse knowledge and skills learned from other datasets. In contrast to previous works, we develop a sparsely activated modular network: (1) We propose a well-rounded set of operators and instantiate each operator with an independent module; (2) We formulate dialogue generation as the execution of a generated programme which recursively composes and assembles modules. Extensive experiments on 9 datasets verify the efficacy of our methods through automatic evaluation and human evaluation. Notably, our model outperforms state-of-the-art supervised approaches on 4 datasets with only 10% training data thanks to the modular architecture and multi-task learning.

DialoGPS: Dialogue Path Sampling in Continuous Semantic Space for Data Augmentation in Multi-Turn Conversations
Ang Lv, Jinpeng Li, yuhan chen, GAO XING, Ji Zhang and Rui Yan   11:00-12:30 (Pier 7&8)
In open-domain dialogue generation tasks, contexts and responses in most datasets are one-to-one mapped, violating an important many-to-many characteristic: a context leads to various responses, and a response answers multiple contexts. Without such patterns, models poorly generalize and prefer responding safely. Many attempts have been made in either multi-turn settings from a one-to-many perspective or in a many-to-many perspective but limited to single-turn settings. The major challenge to many-to-many augment multi-turn dialogues is that discretely replacing each turn with semantic similarity breaks fragile context coherence. In this paper, we propose DialoGue Path Sampling (DialoGPS) method in continuous semantic space, the first many-to-many augmentation method for multi-turn dialogues. Specifically, we map a dialogue to our extended Brownian Bridge, a special Gaussian process. We sample latent variables to form coherent dialogue paths in the continuous space. A dialogue path corresponds to a new multi-turn dialogue and is used as augmented training data. We show the effect of DialoGPS with both automatic and human evaluation.

XDailyDialog: A Multilingual Parallel Dialogue Corpus
Zeming Liu, Ping Nie, Jie Cai, Haifeng Wang, Zheng-Yu Niu, PENG ZHANG, Mrinmaya Sachan and Kaiping Peng 11:00-12:30 (Pier 7&8)
High-quality datasets are significant to the development of dialogue models. However, most existing datasets for open-domain dialogue modeling are limited to a single language. The absence of multilingual open-domain dialog datasets not only limits the research on multilingual or cross-lingual transfer learning, but also hinders the development of robust open-domain dialog systems that can be deployed in other parts of the world. In this paper, we provide a multilingual parallel open-domain dialog dataset, XDailyDialog, to enable researchers to explore the challenging task of multilingual and cross-lingual open-domain dialog. XDailyDialog includes 13K dialogues aligned across 4 languages (52K dialogues and 410K utterances in total). We then propose a dialog generation model, kNN-Chat, which has a novel kNN-search mechanism to support unified response retrieval for monolingual, multilingual, and cross-lingual dialogue. Experiment results show the effectiveness of this framework. We will make XDailyDialog and kNN-Chat publicly available soon.

On the Correspondence between Compositionality and Imitation in Emergent Neural Communication
Emily Cheng, Mathieu Rita and Thierry Poibeau   11:00-12:30 (Pier 7&8)
Compositionality is a hallmark of human language that not only enables linguistic generalization, but also potentially facilitates acquisition. When simulating language emergence with neural networks, compositionality has been shown to improve communication performance; however, its impact on imitation learning has yet to be investigated. Our work explores the link between compositionality and imitation in a Lewis game played by deep neural agents. Our contributions are twofold: first, we show that the learning algorithm used to imitate is crucial: supervised learning tends to produce more average languages, while reinforcement learning introduces a selection pressure toward more compositional languages. Second, our study reveals that compositional languages are easier to imitate, which may induce the pressure toward compositional languages in RL imitation settings.

StructSP: Efficient Fine-tuning of Task-Oriented Dialog System by Using Structure-aware Boosting and Grammar Constraints
Truong Dinh Do, Phuong Minh Nguyen and Minh Le Nguyen   11:00-12:30 (Pier 7&8)
We have investigated methods utilizing hierarchical structure information representation in the semantic parsing task and have devised a method that reinforces the semantic awareness of a pre-trained language model via a two-step fine-tuning mechanism: hierarchical structure information strengthening and a final specific task. The model used is better than existing ones at learning the contextual representations of utterances embedded within its hierarchical semantic structure and thereby improves system performance. In addition, we created a mechanism using inductive grammar to dynamically prune the unpromising directions in the semantic structure parsing process. Finally, through experimentsOur code will be published when this paper is accepted. on the TOP and TOPv2 (low-resource setting) datasets, we achieved state-of-the-art (SOTA) performance, confirming the effectiveness of our proposed model.

DKAF: KB Arbitration for Learning Task-Oriented Dialog Systems with Dialog-KB Inconsistencies
Vishal Vivek Saley, Rocktim Jyoti Das, Dinesh Raghu and Mausam -   11:00-12:30 (Pier 7&8)
Task-oriented dialog (TOD) agents often ground their responses on external knowledge bases (KBs). These KBs can be dynamic and may be updated frequently. Existing approaches for learning TOD agents assume the KB snapshot contemporary to each individual dialog is available during training. However, in real-world scenarios, only the latest KB snapshot is available during training and as a result, the train dialogs may contain facts conflicting with the latest KB. These dialog-KB inconsistencies in the training data may potentially confuse the TOD agent learning algorithm.

In this work, we define the novel problem of learning a TOD agent with dialog-KB inconsistencies in the training data. We propose a Dialog-KB Arbitration Framework (DKAF) which reduces the dialog-KB inconsistencies by predicting the contemporary KB snapshot for each train dialog. These predicted KB snapshots are then used for training downstream TOD agents. As there are no existing datasets with dialog-KB inconsistencies, we systematically introduce inconsistencies in two publicly available dialog datasets. We show that TOD agents trained with DKAF perform better than existing baselines on both these datasets.

MMDialog: A Large-scale Multi-turn Dialogue Dataset Towards Multi-modal Open-domain Conversation
Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao and Qingwei Lin 11:00-12:30 (Pier 7&8)
Responding with multi-modal content has been recognized as an essential capability for an intelligent conversational agent. In this paper, we introduce the MMDialog dataset to facilitate multi-modal conversation better. MMDialog is composed of a curated set of 1.08 million real-world dialogues with 1.53 million unique images across 4,184 topics. MMDialog has two main and unique advantages. First, it is the largest multi-modal conversation dataset by the number of dialogues by 88x. Second, it contains massive topics to generalize the open domain. To build an engaging dialogue system with this dataset, we propose and normalize two response prediction tasks based on retrieval and generative scenarios. In addition, we build two baselines for the above tasks with state-of-the-art techniques and report their experimental performance. We also propose a novel evaluation metric MM-Relevance to measure the multi-modal responses. Our dataset is available in https://github.com/victorsungo/MMDialog.

Improving Cross-task Generalization of Unified Table-to-text Models with Compositional Task Configurations
Jifan Chen, Yuhao Zhang, Lan Liu, Rui Dong, Xinchi Chen, Patrick Ng, William Yang Wang and zhiheng huang 11:00-12:30 (Pier 7&8)
There has been great progress in unifying various table-to-text tasks using a single encoder-decoder model trained via multi-task learning (Xie et al., 2022). However, existing methods typically encode task information with a simple dataset name as a prefix to the encoder. This not only

limits the effectiveness of multi-task learning, but also hinders the model's ability to generalize to new domains or tasks that were not seen during training, which is crucial for real-world applications. In this paper, we propose compositional task configurations, a set of prompts prepended to the encoder to improve cross-task generalization of unified models. We design the task configurations to explicitly specify the task type, as well as its input and output types. We show that this not only allows the model to better learn shared knowledge across different tasks at training, but also allows us to control the model by composing new configurations that apply novel input-output combinations in a zero-shot manner. We demonstrate via experiments over ten table-to-text tasks that our method outperforms the UnifiedSKG baseline by noticeable margins in both in-domain and zero-shot settings, with average improvements of +0.5 and +12.6 from using a T5-large backbone, respectively.

Optimizing Test-Time Query Representations for Dense Retrieval
Mujeen Sung, Jungsoo Park, Jaewoo Kang, Danqi Chen and Jinhyuk Lee   11:00-12:30 (Pier 7&8)
Recent developments of dense retrieval rely on quality representations of queries and contexts from pre-trained query and context encoders. In this paper, we introduce TOUR (Test-Time Optimization of Query Representations), which further optimizes instance-level query representations guided by signals from test-time retrieval results. We leverage a cross-encoder re-ranker to provide fine-grained pseudo labels over retrieval results and iteratively optimize query representations with gradient descent. Our theoretical analysis reveals that TOUR can be viewed as a generalization of the classical Rocchio algorithm for pseudo relevance feedback, and we present two variants that leverage pseudo-labels as hard binary or soft continuous labels. We first apply TOUR on phrase retrieval with our proposed phrase re-ranker, and also evaluate its effectiveness on passage retrieval with an off-the-shelf re-ranker. TOUR greatly improves end-to-end open-domain question answering accuracy, as well as passage retrieval performance. TOUR also consistently improves direct re-ranking by up to 2.0% while running 1.3–2.4x faster with an efficient implementation.

Phrase Retrieval for Open Domain Conversational Question Answering with Conversational Dependency Modeling via Contrastive Learning
Soyeong Jeong, Jinheon Baek, Sung Ju Hwang and Jong Park   11:00-12:30 (Pier 7&8)
Open-Domain Conversational Question Answering (ODConvQA) aims at answering questions through a multi-turn conversation based on a retriever-reader pipeline, which retrieves passages and then predicts answers with them. However, such a pipeline approach not only makes the reader vulnerable to the errors propagated from the retriever, but also demands additional effort to develop both the retriever and the reader, which further makes it slower since they are not runnable in parallel. In this work, we propose a method to directly predict answers with a phrase retrieval scheme for a sequence of words, reducing the conventional two distinct subtasks into a single one. Also, for the first time, we study its capability for ODConvQA tasks. However, simply adopting it is largely problematic, due to the dependencies between previous and current turns in a conversation. To address this problem, we further introduce a novel contrastive learning strategy, making sure to reflect previous turns when retrieving the phrase for the current context, by maximizing representational similarities of consecutive turns in a conversation while minimizing irrelevant conversational contexts. We validate our model on two ODConvQA datasets, whose experimental results show that it substantially outperforms the relevant baselines with the retriever-reader. Code is available at: https://github.com/starsuzi/PRO-ConvQA.

S3HQA: A Three-Stage Approach for Multi-hop Text-Table Hybrid Question Answering
Fangyu Lei, Xiang Li, Yifan Wei, Shizhu He, Yiming Huang, Jun Zhao and Kang Liu   11:00-12:30 (Pier 7&8)
Answering multi-hop questions over hybrid factual knowledge from the given text and table (TextTableQA) is a challenging task. Existing models mainly adopt a retriever-reader framework, which have several deficiencies, such as noisy labeling in training retriever, insufficient utilization of heterogeneous information over text and table, and deficient ability for different reasoning operations. In this paper, we propose a three-stage TextTableQA framework S3HQA, which comprises of retriever, selector, and reasoner. We use a retriever with refinement training to solve the noisy labeling problem. Then, a hybrid selector considers the linked relationships between heterogeneous data to select the most relevant factual knowledge. For the final stage, instead of adapting a reading comprehension module like in previous methods, we employ a generation-based reasoner to obtain answers. This includes two approaches: a row-wise generator and an LLM prompting generator (first time used in this task). The experimental results demonstrate that our method achieves competitive results in the few-shot setting. When trained on the full dataset, our approach outperforms all baseline methods, ranking first on the HybridQA leaderboard.

DePlot: One-shot visual language reasoning by plot-to-table translation
Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier and Yasemin Altun   11:00-12:30 (Pier 7&8)
Visual language such as charts and plots is ubiquitous in the human world. Comprehending plots and charts requires strong reasoning skills. Prior state-of-the-art (SOTA) models require at least tens of thousands of training examples and their reasoning capabilities are still much limited, especially on complex human-written queries. This paper presents the first one-shot solution to visual language reasoning. We decompose the challenge of visual language reasoning into two steps: (1) plot-to-text translation, and (2) reasoning over the translated text. The key in this method is a modality conversion module, named as DePlot, which translates the image of a plot or chart to a linearized table. The output of DePlot can then be directly used to prompt a pretrained large language model (LLM), exploiting the few-shot reasoning capabilities of LLMs. To obtain DePlot, we standardize the plot-to-table task by establishing unified task formats and metrics, and train DePlot end-to-end on this task. DePlot can then be used off-the-shelf together with LLMs in a plug-and-play fashion. Compared with a SOTA model finetuned on more than thousands of data points, DePlot+LLM with just one-shot prompting achieves a 29.4% improvement over finetuned SOTA on human-written queries from the task of chart QA.

Coupling Large Language Models with Logic Programming for Robust and General Reasoning from Text
Zhun Yang, Adam Ishay and Joohyung Lee   11:00-12:30 (Pier 7&8)
While large language models (LLMs), such as GPT-3, appear to be robust and general, their reasoning ability is not at a level to compete with the best models trained for specific natural language reasoning problems. In this study, we observe that a large language model can serve as a highly effective few-shot semantic parser. It can convert natural language sentences into a logical form that serves as input for answer set programs, a logic-based declarative knowledge representation formalism. The combination results in a robust and general system that can handle multiple question-answering tasks without requiring retraining for each new task. It only needs a few examples to guide the LLM's adaptation to a specific task, along with reusable ASP knowledge modules that can be applied to multiple tasks. We demonstrate that this method achieves state-of-the-art performance on several NLP benchmarks, including bAbI, StepGame, CLUTRR, and gSCAN. Additionally, it successfully tackles robot planning tasks that an LLM alone fails to solve.

Dynamic Heterogeneous-Graph Reasoning with Language Models and Knowledge Representation Learning for Commonsense Question Answering
Yujie Wang, Hu Zhang, Jiye Liang and Ru Li   11:00-12:30 (Pier 7&8)
Recently, knowledge graphs (KGs) have won noteworthy success in commonsense question answering. Existing methods retrieve relevant subgraphs in the KGs through key entities and reason about the answer with language models (LMs) and graph neural networks. However, they ignore (i) optimizing the knowledge representation and structure of subgraphs and (ii) deeply fusing heterogeneous QA context with

subgraphs. In this paper, we propose a dynamic heterogeneous-graph reasoning method with LMs and knowledge representation learning (DHLK), which constructs a heterogeneous knowledge graph (HKG) based on multiple knowledge sources and optimizes the structure and knowledge representation of the HKG using a two-stage pruning strategy and knowledge representation learning (KRL). It then performs joint reasoning by LMs and Relation Mask Self-Attention (RMSA). Specifically, DHLK filters key entities based on the dictionary vocabulary to achieve the first-stage pruning while incorporating the paraphrases in the dictionary into the subgraph to construct the HKG. Then, DHLK encodes and fuses the QA context and HKG using LM, and dynamically removes irrelevant KG entities based on the attention weights of LM for the second-stage pruning. Finally, DHLK introduces KRL to optimize the knowledge representation and perform answer reasoning on the HKG by RMSA.We evaluate DHLK at CommonsenseQA and OpenBookQA, and show its improvement on existing LM and LM+KG methods.

Tab-CoT: Zero-shot Tabular Chain of Thought
Jin Ziqi and Wei Lu 11:00-12:30 (Pier 7&8)
The chain-of-though (CoT) prompting methods were successful in various natural language processing (NLP) tasks thanks to their ability to unveil the underlying complex reasoning processes. Such reasoning processes typically exhibit highly structured steps. Recent efforts also started investigating methods to encourage more structured reasoning procedures to be captured (cite least to most). In this work, we propose Tab-CoT, a novel tabular-format CoT prompting method, which allows the complex reasoning process to be explicitly modeled in a highly structured manner. Despite its simplicity, we show that our approach is capable of performing reasoning across multiple dimensions (i.e., both rows and columns). We demonstrate our approach's strong zero-shot and few-shot capabilities through extensive experiments on a range of reasoning tasks.

World Models for Math Story Problems
Andreas Opedal, Niklas Stoehr, Abulhair Saparov and Mrinmaya Sachan 11:00-12:30 (Pier 7&8)
Solving math story problems is a complex task for students and NLP models alike, requiring them to understand the world as described in the story and reason over it to compute an answer. Recent years have seen impressive performance on automatically solving these problems with large pre-trained language models and innovative techniques to prompt them. However, it remains unclear if these models possess accurate representations of mathematical concepts. This leads to lack of interpretability and trustworthiness which impedes their usefulness in various applications. In this paper, we consolidate previous work on categorizing and representing math story problems and develop MathWorld, which is a graph-based semantic formalism specific for the domain of math story problems. With MathWorld, we can assign world models to math story problems which represent the situations and actions introduced in the text and their mathematical relationships. We combine math story problems from several existing datasets and annotate a corpus of 1,019 problems and 3,204 logical forms with MathWorld. Using this data, we demonstrate the following use cases of MathWorld: (1) prompting language models with synthetically generated question-answer pairs to probe their reasoning and world modeling abilities, and (2) generating new problems by using the world models as a design space.

Reasoning over Hierarchical Question Decomposition Tree for Explainable Question Answering
Jiajie Zhang, Shulin Cao, Tingjian Zhang, Xin Lv, Juanzi Li, Lei Hou, Jiaxin Shi and Qi Tian 11:00-12:30 (Pier 7&8)
Explainable question answering (XQA) aims to answer a given question and provide an explanation why the answer is selected. Existing XQA methods focus on reasoning on a single knowledge source, e.g., structured knowledge bases, unstructured corpora, etc. However, integrating information from heterogeneous knowledge sources is essential to answer complex questions. In this paper, we propose to leverage question decomposing for heterogeneous knowledge integration, by breaking down a complex question into simpler ones, and selecting the appropriate knowledge source for each sub-question. To facilitate reasoning, we propose a novel two-stage XQA framework, Reasoning over Hierarchical Question Decomposition Tree (RoHT). First, we build the Hierarchical Question Decomposition Tree (HQDT) to understand the semantics of a complex question; then, we conduct probabilistic reasoning over HQDT from root to leaves recursively, to aggregate heterogeneous knowledge at different tree levels and search for a best solution considering the decomposing and answering probabilities. The experiments on complex QA datasets KQA Pro and Musique show that our framework outperforms SOTA methods significantly, demonstrating the effectiveness of leveraging question decomposing for knowledge integration and our RoHT framework.

When to Read Documents or QA History: On Unified and Selective Open-domain QA
Kyungjae Lee, Sang-eun Han, Seung-won Hwang and Moontae Lee 11:00-12:30 (Pier 7&8)
This paper studies the problem of open-domain question answering, with the aim of answering a diverse range of questions leveraging knowledge resources. Two types of sources, QA-pair and document corpora, have been actively leveraged with the following complementary strength. The former is highly precise when the paraphrase of given question q was seen and answered during training, often posed as a retrieval problem, while the latter generalizes better for unseen questions. A natural follow-up is thus leveraging both models, while a naive pipelining or integration approaches have failed to bring additional gains over either model alone. Our distinction is interpreting the problem as calibration, which estimates the confidence of predicted answers as an indicator to decide when to use a document or QA-pair corpus. The effectiveness of our method was validated on widely adopted benchmarks such as Natural Questions and TriviaQA.

Distinguish Before Answer: Generating Contrastive Explanation as Knowledge for Commonsense Question Answering
Qianglong Chen, Guohai Xu, Ming Yan, Ji Zhang, Fei Huang, Luo Si and Yin Zhang 11:00-12:30 (Pier 7&8)
Existing knowledge-enhanced methods have achieved remarkable results in certain Q&A tasks via obtaining diverse knowledge from different knowledge bases. However, limited by the properties of retrieved knowledge, they still have trouble benefiting from both the knowledge relevance and distinguishment simultaneously. To address the challenge, we propose **CPACE**, a **C**oncept-centric **P**rompt-b**A**sed **C**ontrastive **E**xplanation Generation model, which aims to convert obtained symbolic knowledge into the contrastive explanation for better distinguishing the differences among given candidates. Firstly, following previous works, we retrieve different types of symbolic knowledge with a concept-centric knowledge extraction module. After that, we generate corresponding contrastive explanation using acquired symbolic knowledge and prompt as guidance for better modeling the knowledge distinguishment and interpretability. Finally, we regard the generated contrastive explanation as external knowledge for downstream task enhancement. We conduct a series of experiments on three widely-used question-answering datasets: CSQA, QASC, and OBQA. Experimental results demonstrate that with the help of generated contrastive explanation, our CPACE model achieves new SOTA on CSQA (89.8% on the testing set, 0.9% higher than human performance), and gains impressive improvement on QASC and OBQA (4.2% and 3.5%, respectively).

Multi-Row, Multi-Span Distant Supervision For Table+Text Question Answering
*Vishwajeet Kumar, Yash Gupta, Saneem Ahmed Chemmengath, Jaydeep Sen, Soumen Chakrabarti, Samarth Bharadwaj and Feifei Pan*11:00-12:30 (Pier 7&8)
Question answering (QA) over tables and linked text, also called TextTableQA, has witnessed significant research in recent years, as tables are often found embedded in documents along with related text. HybridQA and OTT-QA are the two best-known TextTableQA datasets, with questions that are best answered by combining information from both table cells and linked text passages. A common challenge in both datasets, and TextTableQA in general, is that the training instances include just the question and answer, where the gold answer may match not only multiple table cells across table rows but also multiple text spans within the scope of a table row and its associated text. This leads to a noisy multi-instance training regime. We present MITQA, a transformer-based TextTableQA system that is explicitly designed to cope with

distant supervision along both these axes, through a multi-instance loss objective, together with careful curriculum design. Our experiments show that the proposed multi-instance distant supervision approach helps MITQA get sate-of-the-art results beating the baselines for both HybridQA and OTT-QA, putting MITQA at the top of HybridQA leaderboard with best EM and F1 scores on a held out test set.

Product Question Answering in E-Commerce: A Survey
Yang Deng, Wenxuan Zhang, Qian Yu and Wai Lam   11:00-12:30 (Pier 7&8)
Product question answering (PQA), aiming to automatically provide instant responses to customer's questions in E-Commerce platforms, has drawn increasing attention in recent years. Compared with typical QA problems, PQA exhibits unique challenges such as the subjectivity and reliability of user-generated contents in E-Commerce platforms. Therefore, various problem settings and novel methods have been proposed to capture these special characteristics. In this paper, we aim to systematically review existing research efforts on PQA. Specifically, we categorize PQA studies into four problem settings in terms of the form of provided answers. We analyze the pros and cons, as well as present existing datasets and evaluation protocols for each setting. We further summarize the most significant challenges that characterize PQA from general QA applications and discuss their corresponding solutions. Finally, we conclude this paper by providing the prospect on several future directions.

SkillQG: Learning to Generate Question for Reading Comprehension Assessment
Xiaoqiang Wang, Bang Liu, Siliang Tang and Lingfei Wu   11:00-12:30 (Pier 7&8)
We present SkillQG: a question generation framework with controllable comprehension types for assessing and improving machine reading comprehension models. Existing question generation systems widely differentiate questions by literal information such as question words and answer types to generate semantically relevant questions for a given context. However, they rarely consider the comprehension nature of questions, i.e., the different comprehension capabilities embodied by different questions. In comparison, our SkillQG is able to tailor a fine-grained assessment and improvement to the capabilities of questions answering models built on it. Specifically, we first frame the comprehension type of questions based on a hierarchical skill-based schema. We then formulate SkillQG as a skill-conditioned question generator. Furthermore, to improve the controllability of generation, we augment the input text with skill-specific question focus and knowledge, which are constructed by iteratively prompting the pre-trained language models. Empirical results demonstrate that SkillQG outperforms baselines in terms of quality, relevance, and skill-controllability while showing a promising performance boost in downstream question answering task.

A Multi-modal Debiasing Model with Dynamical Constraint for Robust Visual Question Answering
Yu Li, Bojie Hu, Fengshuo Zhang, Yahan Yu, Jian Liu, Yufeng Chen and Jinan Xu   11:00-12:30 (Pier 7&8)
Recent studies have pointed out that many well-developed Visual Question Answering (VQA) systems suffer from bias problem. Despite the remarkable performance gained on In-Distribution (ID) datasets, the VQA model might merely capture the superficial correlation from question to answer rather than showing real reasoning abilities. Therefore, when switching to Out-of-Distribution (OOD) dataset, whose test distribution is unknown or even reversed with the training set, significant drops might be demonstrated. Although efforts have been devoted to easing the negative bias effect brought by language prior and analysing its inherent cause, they are still limited by the following two aspects. First, most current debiasing methods achieve promising OOD generalization ability with a major sacrifice of the ID performance. Second, existing researches are restricted by exploiting comprehensive biases, since weakening the language bias is mainly focused, while only a few works consider vision bias. In this paper, we investigate a straightforward way to mitigate bias problem for VQA task. Specifically, we reduce bias effect by subtracting bias score from standard VQA base score. Based on such a direct strategy, we design two bias learning branches to detect more bias information, which are combined with a dynamical constraint loss to alleviate the problem of over-correction and insufficient debiasing effect. We evaluate our method on the challenging VQA v2.0 and VQA-CP V2,0 datasets and the proposed method achievessignificant improvement.

Re-appraising the Schema Linking for Text-to-SQL
Yujian Gan, Xinyun Chen and Matthew Purver   11:00-12:30 (Pier 7&8)
Most text-to-SQL models, even though based on the same grammar decoder, generate the SQL structure first and then fill in the SQL slots with the correct schema items. This second step depends on schema linking: aligning the entity references in the question with the schema columns or tables. This is generally approached via Exact Match based Schema Linking (EMSL) within a neural network-based schema linking module. EMSL has become standard in text-to-SQL: many state-of-the-art models employ EMSL, with performance dropping significantly when the EMSL component is removed. In this work, however, we show that EMSL reduces robustness, rendering models vulnerable to synonym substitution and typos. Instead of relying on EMSL to make up for deficiencies in question-schema encoding, we show that using a pre-trained language model as an encoder can improve performance without using EMSL, giving a more robust model. We also study the design choice of the schema linking module, finding that a suitable design benefits performance and interoperability. Finally, based on the above study of schema linking, we introduce the grammar linking to help model align grammar references in the question with the SQL keywords.

QAP: A Quantum-Inspired Adaptive-Priority-Learning Model for Multimodal Emotion Recognition
Ziming Li, Yan Zhou, Yaxin Liu, Fuqing Zhu, Chuanpeng Yang and Songlin Hu   11:00-12:30 (Pier 7&8)
Multimodal emotion recognition for video has gained considerable attention in recent years, in which three modalities (*i.e.,* textual, visual and acoustic) are involved. Due to the diverse levels of informational content related to emotion, three modalities typically possess varying degrees of contribution to emotion recognition. More seriously, there might be inconsistencies between the emotion of individual modality and the video. The challenges mentioned above are caused by the inherent uncertainty of emotion. Inspired by the recent advances of quantum theory in modeling uncertainty, we make an initial attempt to design a quantum-inspired adaptive-priority-learning model (QAP) to address the challenges. Specifically, the quantum state is introduced to model modal features, which allows each modality to retain all emotional tendencies until the final classification. Additionally, we design Q-attention to orderly integrate three modalities, and then QAP learns modal priority adaptively so that modalities can provide different amounts of information based on priority. Experimental results on the IEMOCAP and MOSEI datasets show that QAP establishes new state-of-the-art results.

Making Better Use of Training Corpus: Retrieval-based Aspect Sentiment Triplet Extraction via Label Interpolation
Guoxin Yu, Lemao Liu, Haiyun Jiang, Shuming Shi and Xiang Ao   11:00-12:30 (Pier 7&8)
In this paper, we aim to adapt the idea of retrieval-based neural approaches to the Aspect Sentiment Triplet Extraction (ASTE) task. Different from previous studies retrieving semantic similar neighbors, the ASTE task has its specialized challenges when adapting, i.e., the purpose includes predicting the sentiment polarity and it is usually aspect-dependent. Semantic similar neighbors with different polarities will be infeasible even counterproductive. To tackle this issue, we propose a retrieval-based neural ASTE approach, named RLI (Retrieval-based Aspect Sentiment Triplet Extraction via Label Interpolation), which exploits the label information of neighbors. Given an aspect-opinion term pair, we retrieve semantic similar triplets from the training corpus and interpolate their label information into the augmented representation of the target pair. The retriever is jointly trained with the whole ASTE framework, and neighbors with both similar semantics and sentiments can be recalled with the aid of this distant supervision. In addition, we design a simple yet effective pre-train method for the retriever that implicitly encodes the label similarities. Extensive experiments and analysis on two widely-used benchmarks show that the proposed model establishes a new state-of-the-art on ASTE.

Span-level Aspect-based Sentiment Analysis via Table Filling
Mao Zhang, Yongxin Zhu, Zhen Liu, Zhimin Bao, Yunfei Wu, Xing Sun and Linli Xu   11:00-12:30 (Pier 7&8)
In this paper, we propose a novel span-level model for Aspect-Based Sentiment Analysis (ABSA), which aims at identifying the sentiment polarity of the given aspect. In contrast to conventional ABSA models that focus on modeling the word-level dependencies between an aspect and its corresponding opinion expressions, in this paper, we propose Table Filling BERT (TF-BERT), which considers the consistency of multi-word opinion expressions at the span-level. Specially, we learn the span representations with a table filling method, by constructing an upper triangular table for each sentiment polarity, of which the elements represent the sentiment intensity of the specific sentiment polarity for all spans in the sentence. Two methods are then proposed, including table-decoding and table-aggregation, to filter out target spans or aggregate each table for sentiment polarity classification. In addition, we design a sentiment consistency regularizer to guarantee the sentiment consistency of each span for different sentiment polarities. Experimental results on three benchmarks demonstrate the effectiveness of our proposed model.

StoryTrans: Non-Parallel Story Author-Style Transfer with Discourse Representations and Content Enhancing
Xuekai Zhu, Jian Guan, Minlie Huang and Juan Liu   11:00-12:30 (Pier 7&8)
Non-parallel text style transfer is an important task in natural language generation. However, previous studies concentrate on the token or sentence level, such as sentence sentiment and formality transfer, but neglect long style transfer at the discourse level. Long texts usually involve more complicated author linguistic preferences such as discourse structures than sentences. In this paper, we formulate the task of non-parallel story author-style transfer, which requires transferring an input story into a specified author style while maintaining source semantics. To tackle this problem, we propose a generation model, named StoryTrans, which leverages discourse representations to capture source content information and transfer them to target styles with learnable style embeddings. We use an additional training objective to disentangle stylistic features from the learned discourse representation to prevent the model from degenerating to an auto-encoder. Moreover, to enhance content preservation, we design a mask-and-fill framework to explicitly fuse style-specific keywords of source texts into generation. Furthermore, we constructed new datasets for this task in Chinese and English, respectively. Extensive experiments show that our model outperforms strong baselines in overall performance of style transfer and content preservation.

AMR-based Network for Aspect-based Sentiment Analysis
Fukun Ma, Xuming Hu, Aiwei Liu, Yawen Yang, Shuang Li, Philip S. Yu and Lijie Wen   11:00-12:30 (Pier 7&8)
Aspect-based sentiment analysis (ABSA) is a fine-grained sentiment classification task. Many recent works have used dependency trees to extract the relation between aspects and contexts and have achieved significant improvements. However, further improvement is limited due to the potential mismatch between the dependency tree as a syntactic structure and the sentiment classification as a semantic task. To alleviate this gap, we replace the syntactic dependency tree with the semantic structure named Abstract Meaning Representation (AMR) and propose a model called AMR-based Path Aggregation Relational Network (APARN) to take full advantage of semantic structures. In particular, we design the path aggregator and the relation-enhanced self-attention mechanism that complement each other. The path aggregator extracts semantic features from AMRs under the guidance of sentence information, while the relation-enhanced self-attention mechanism in turn improves sentence features with refined semantic information. Experimental results on four public datasets demonstrate 1.13% average F1 improvement of APARN in ABSA when compared with state-of-the-art baselines.

A Unified One-Step Solution for Aspect Sentiment Quad Prediction
Junxian Zhou, Haiqin Yang, Yuxuan He, Hao Mou and JunBo Yang   11:00-12:30 (Pier 7&8)
Aspect sentiment quad prediction (ASQP) is a challenging yet significant subtask in aspectbased sentiment analysis as it provides a complete aspect-level sentiment structure. However, existing ASQP datasets are usually small and low-density, hindering technical advancement. To expand the capacity, in this paper, we release two new datasets for ASQP, which contain the following characteristics: larger size, more words per sample, and higher density. With such datasets, we unveil the shortcomings of existing strong ASQP baselines and therefore propose a unified one-step solution for ASQP, namely One-ASQP, to detect the aspect categories and to identify the aspectopinion-sentiment (AOS) triplets simultaneously. Our One-ASQP holds several unique advantages: (1) by separating ASQP into two subtasks and solving them independently and simultaneously, we can avoid error propagation in pipeline-based methods and overcome slow training and inference in generation-based methods; (2) by introducing sentiment-specific horns tagging schema in a token-pair-based two-dimensional matrix, we can exploit deeper interactions between sentiment elements and efficiently decode the AOS triplets; (3) we design "[NULL]" token can help us effectively identify the implicit aspects or opinions. Experiments on two benchmark datasets and our released two datasets demonstrate the advantages of our One-ASQP. The two new datasets are publicly released at https://www.github.com/Datastory-CN/ASQP-Datasets.

Few-shot Joint Multimodal Aspect-Sentiment Analysis Based on Generative Multimodal Prompt
Xiaocui Yang, Shi Feng, Daling Wang, Qi Sun, Wenfang Wu, Yifei Zhang, Pengfei Hong and Soujanya Poria   11:00-12:30 (Pier 7&8)
We have witnessed the rapid proliferation of multimodal data on numerous social media platforms. Conventional studies typically require massive labeled data to train models for Multimodal Aspect-Based Sentiment Analysis (MABSA). However, collecting and annotating fine-grained multimodal data for MABSA is tough. To alleviate the above issue, we perform three MABSA-related tasks with quite a small number of labeled multimodal samples. We first build diverse and comprehensive multimodal few-shot datasets according to the data distribution. To capture the specific prompt for each aspect term in a few-shot scenario, we propose a novel Generative Multimodal Prompt (GMP) model for MABSA, which includes the Multimodal Encoder module and the N-Stream Decoders module. We further introduce a subtask to predict the number of aspect terms in each instance to construct the multimodal prompt. Extensive experiments on two datasets demonstrate that our approach outperforms strong baselines on two MABSA-related tasks in the few-shot setting.

Zero-shot Approach to Overcome Perturbation Sensitivity of Prompts
Mohna Chakraborty, Adithya Kulkarni and Qi Li   11:00-12:30 (Pier 7&8)
Recent studies have demonstrated that natural-language prompts can help to leverage the knowledge learned by pre-trained language models for the binary sentence-level sentiment classification task. Specifically, these methods utilize few-shot learning settings to fine-tune the sentiment classification model using manual or automatically generated prompts. However, the performance of these methods is sensitive to the perturbations of the utilized prompts. Furthermore, these methods depend on a few labeled instances for automatic prompt generation and prompt ranking. This study aims to find high-quality prompts for the given task in a zero-shot setting. Given a base prompt, our proposed approach automatically generates multiple prompts similar to the base prompt employing positional, reasoning, and paraphrasing techniques and then ranks the prompts using a novel metric. We empirically demonstrate that the top-ranked prompts are high-quality and significantly outperform the base prompt and the prompts generated using few-shot learning for the binary sentence-level sentiment classification task.

A Simple Yet Strong Domain-Agnostic De-bias Method for Zero-Shot Sentiment Classification
Yang Zhao, Tetsuya Nasukawa, Masayasu Muraoka and Bishwaranjan Bhattacharjee   11:00-12:30 (Pier 7&8)
Zero-shot prompt-based learning has made much progress in sentiment analysis, and considerable effort has been dedicated to designing high-performing prompt templates. However, two problems exist; First, large language models are often biased to their pre-training data, leading to poor performance in prompt templates that models have rarely seen. Second, in order to adapt to different domains, re-designing

prompt templates is usually required, which is time-consuming and inefficient. To remedy both shortcomings, we propose a simple yet strong data construction method to de-bias a given prompt template, yielding a large performance improvement in sentiment analysis tasks across different domains, pre-trained language models, and prompt templates. Also, we demonstrate the advantage of using domain-agnostic generic responses over the in-domain ground-truth data.

TransESC: Smoothing Emotional Support Conversation via Turn-Level State Transition

Weixiang Zhao, Yanyan Zhao, Shilong Wang and Bing Qin   11:00-12:30 (Pier 7&8)

Emotion Support Conversation (ESC) is an emerging and challenging task with the goal of reducing the emotional distress of people. Previous attempts fail to maintain smooth transitions between utterances in ESC because they ignoring to grasp the fine-grained transition information at each dialogue turn. To solve this problem, we propose to take into account turn-level state Transitions of ESC (TransESC) from three perspectives, including semantics transition, strategy transition and emotion transition, to drive the conversation in a smooth and natural way. Specifically, we construct the state transition graph with a two-step way, named transit-then-interact, to grasp such three types of turn-level transition information. Finally, they are injected into the transition aware decoder to generate more engaging responses. Both automatic and human evaluations on the benchmark dataset demonstrate the superiority of TransESC to generate more smooth and effective supportive responses. Our source code will be publicly available.

A Dataset of Argumentative Dialogues on Scientific Papers

Federico Ruggeri, Mohsen Mesgar and Iryna Gurevych   11:00-12:30 (Pier 7&8)

With recent advances in question-answering models, various datasets have been collected to improve and study the effectiveness of these models on scientific texts. Questions and answers in these datasets explore a scientific paper by seeking factual information from the paper's content. However, these datasets do not tackle the argumentative content of scientific papers, which is of huge importance in persuasiveness of a scientific discussion. We introduce ArgSciChat, a dataset of 41 argumentative dialogues between scientists on 20 NLP papers. The unique property of our dataset is that it includes both exploratory and argumentative questions and answers in a dialogue discourse on a scientific paper. Moreover, the size of ArgSciChat demonstrates the difficulties in collecting dialogues for specialized domains. Thus, our dataset is a challenging resource to evaluate dialogue agents in low-resource domains, in which collecting training data is costly. We annotate all sentences of dialogues in ArgSciChat and analyze them extensively. The results confirm that dialogues in ArgSciChat include exploratory and argumentative interactions. Furthermore, we use our dataset to fine-tune and evaluate a pre-trained document-grounded dialogue agent. The agent achieves a low performance on our dataset, motivating a need for dialogue agents with a capability to reason and argue about their answers. We publicly release ArgSciChat.

PAED: Zero-Shot Persona Attribute Extraction in Dialogues

Luyao Zhu, Wei Li, Rui Mao, Vlad Pandelea and Erik Cambria   11:00-12:30 (Pier 7&8)

Persona attribute extraction is critical for personalized human-computer interaction. Dialogue is an important medium that communicates and delivers persona information. Although there is a public dataset for triplet-based persona attribute extraction from conversations, its automatically generated labels present many issues, including unspecific relations and inconsistent annotations. We fix such issues by leveraging more reliable text-label matching criteria to generate high-quality data for persona attribute extraction. We also propose a contrastive learning- and generation-based model with a novel hard sampling strategy for generalized zero-shot persona attribute extraction. We benchmark our model with state-of-the-art baselines on our dataset and a public dataset, showing outstanding accuracy gains. Our sampling strategy also exceeds others by a large margin in persona attribute extraction.

Balancing the Effect of Training Dataset Distribution of Multiple Styles for Multi-Style Text Transfer

Debarati Das, David Ma and Dongyeop Kang   11:00-12:30 (Pier 7&8)

Text style transfer is an exciting task within the field of natural language generation that is often plagued by the need for high-quality paired datasets. Furthermore, training a model for multi-attribute text style transfer requires datasets with sufficient support across all combinations of the considered stylistic attributes, adding to the challenges of training a style transfer model. This paper explores the impact of training data input diversity on the quality of the generated text from the multi-style transfer model. We construct a pseudo-parallel dataset by devising heuristics to adjust the style distribution in the training samples. We balance our training dataset using marginal and joint distributions to train our style transfer models. We observe that a balanced dataset produces more effective control effects over multiple styles than an imbalanced or skewed one. Through quantitative analysis, we explore the impact of multiple style distributions in training data on style-transferred output. These findings will better inform the design of style-transfer datasets.

Measuring Your ASTE Models in The Wild: A Diversified Multi-domain Dataset For Aspect Sentiment Triplet Extraction

Ting Xu, Huiyun Yang, Zhen Wu, Jiaze Chen, Fei Zhao and Xinyu Dai   11:00-12:30 (Pier 7&8)

Aspect Sentiment Triplet Extraction (ASTE) is widely used in various applications. However, existing ASTE datasets are limited in their ability to represent real-world scenarios, hindering the advancement of research in this area. In this paper, we introduce a new dataset, named DMASTE, which is manually annotated to better fit real-world scenarios by providing more diverse and realistic reviews for the task. The dataset includes various lengths, diverse expressions, more aspect types, and more domains than existing datasets. We conduct extensive experiments on DMASTE in multiple settings to evaluate previous ASTE approaches. Empirical results demonstrate that DMASTE is a more challenging ASTE dataset. Further analyses of in-domain and cross-domain settings provide some promising directions for future research.

Opinion Tree Parsing for Aspect-based Sentiment Analysis

Xiaoyi Bao, Xiaotong Jiang, Zhongqing Wang, Yue Zhang and Guodong Zhou   11:00-12:30 (Pier 7&8)

Extracting sentiment elements using pre-trained generative models has recently led to large improvements in aspect-based sentiment analysis benchmarks. These models avoid explicit modeling of structure between sentiment elements, which are succinct yet lack desirable properties such as structure well-formedness guarantees or built-in elements alignments. In this study, we propose an opinion tree parsing model, aiming to parse all the sentiment elements from an opinion tree, which can explicitly reveal a more comprehensive and complete aspect-level sentiment structure. In particular, we first introduce a novel context-free opinion grammar to normalize the sentiment structure. We then employ a neural chart-based opinion tree parser to fully explore the correlations among sentiment elements and parse them in the opinion tree form. Extensive experiments show the superiority of our proposed model and the capacity of the opinion tree parser with the proposed context-free opinion grammar. More importantly, our model is much faster than previous models.

MvP: Multi-view Prompting Improves Aspect Sentiment Tuple Prediction

Zhibin Gou, qingyan guo and Yujiu Yang   11:00-12:30 (Pier 7&8)

Generative methods greatly promote aspect-based sentiment analysis via generating a sequence of sentiment elements in a specified format. However, existing studies usually predict sentiment elements in a fixed order, which ignores the effect of the interdependence of the elements in a sentiment tuple and the diversity of language expression on the results. In this work, we propose Multi-view Prompting (MVP) that aggregates sentiment elements generated in different orders, leveraging the intuition of human-like problem-solving processes from different views. Specifically, MVP introduces element order prompts to guide the language model to generate multiple sentiment tuples, each with a different element order, and then selects the most reasonable tuples by voting. MVP can naturally model multi-view and multi-task as permutations

and combinations of elements, respectively, outperforming previous task-specific designed methods on multiple ABSA tasks with a single model. Extensive experiments show that MVP significantly advances the state-of-the-art performance on 10 datasets of 4 benchmark tasks, and performs quite effectively in low-resource settings. Detailed evaluation verified the effectiveness, flexibility, and cross-task transferability of MVP.

Zero-Shot and Few-Shot Stance Detection on Varied Topics via Conditional Generation
Haoyang Wen and Alexander Hauptmann   11:00-12:30 (Pier 7&8)
Zero-shot and few-shot stance detection identify the polarity of text with regard to a certain target when we have only limited or no training resources for the target. Previous work generally formulates the problem into a classification setting, ignoring the potential use of label text. In this paper, we instead utilize a conditional generation framework and formulate the problem as denoising from partially-filled templates, which can better utilize the semantics among input, label, and target texts. We further propose to jointly train an auxiliary task, target prediction, and to incorporate manually constructed incorrect samples with unlikelihood training to improve the representations for both target and label texts. We also verify the effectiveness of target-related Wikipedia knowledge with the generation framework. Experiments show that our proposed method significantly outperforms several strong baselines on VAST, and achieves new state-of-the-art performance.

MetaAdapt: Domain Adaptive Few-Shot Misinformation Detection via Meta Learning
Zhenrui Yue, Huimin Zeng, Yang Zhang, Lanyu Shang and Dong Wang   11:00-12:30 (Pier 7&8)
With emerging topics (e.g., COVID-19) on social media as a source for the spreading misinformation, overcoming the distributional shifts between the original training domain (i.e., source domain) and such target domains remains a non-trivial task for misinformation detection. This presents an elusive challenge for early-stage misinformation detection, where a good amount of data and annotations from the target domain is not available for training. To address the data scarcity issue, we propose MetaAdapt, a meta learning based approach for domain adaptive few-shot misinformation detection. MetaAdapt leverages limited target examples to provide feedback and guide the knowledge transfer from the source to the target domain (i.e., learn to adapt). In particular, we train the initial model with multiple source tasks and compute their similarity scores to the meta task. Based on the similarity scores, we rescale the meta gradients to adaptively learn from the source tasks. As such, MetaAdapt can learn how to adapt the misinformation detection model and exploit the source data for improved performance in the target domain. To demonstrate the efficiency and effectiveness of our method, we perform extensive experiments to compare MetaAdapt with state-of-the-art baselines and large language models (LLMs) such as LLaMA, where MetaAdapt achieves better performance in domain adaptive few-shot misinformation detection with substantially reduced parameters on real-world datasets.

Words as Gatekeepers: Measuring Discipline-specific Terms and Meanings in Scholarly Publications
Li Lucy, Jesse Dodge, David Bamman and Katherine A Keith   11:00-12:30 (Pier 7&8)
Scholarly text is often laden with jargon, or specialized language that can facilitate efficient in-group communication within fields but hinder understanding for out-groups. In this work, we develop and validate an interpretable approach for measuring scholarly jargon from text. Expanding the scope of prior work which focuses on word types, we use word sense induction to also identify words that are widespread but overloaded with different meanings across fields. We then estimate the prevalence of these discipline-specific words and senses across hundreds of subfields, and show that word senses provide a complementary, yet unique view of jargon alongside word types. We demonstrate the utility of our metrics for science of science and computational sociolinguistics by highlighting two key social implications. First, though most fields reduce their use of jargon when writing for general-purpose venues, and some fields (e.g., biological sciences) do so less than others. Second, the direction of correlation between jargon and citation rates varies among fields, but jargon is nearly always negatively correlated with interdisciplinary impact. Broadly, our findings suggest that though multidisciplinary venues intend to cater to more general audiences, some fields' writing norms may act as barriers rather than bridges, and thus impede the dispersion of scholarly ideas.

Dramatic Conversation Disentanglement
Kent K. Chang, Danica Chen and David Bamman   11:00-12:30 (Pier 7&8)
We present a new dataset for studying conversation disentanglement in movies and TV series. While previous work has focused on conversation disentanglement in IRC chatroom dialogues, movies and TV shows provide a space for studying complex pragmatic patterns of floor and topic change in face-to-face multi-party interactions. In this work, we draw on theoretical research in sociolinguistics, sociology, and film studies to operationalize a conversational thread (including the notion of a floor change) in dramatic texts, and use that definition to annotate a dataset of 10,033 dialogue turns (comprising 2,209 threads) from 831 movies. We compare the performance of several disentanglement models on this dramatic dataset, and apply the best-performing model to disentangle 808 movies. We see that, contrary to expectation, average thread lengths do not decrease significantly over the past 40 years, and characters portrayed by actors who are women, while underrepresented, initiate more new conversational threads relative to their speaking time.

Counterspeeches up my sleeve! Intent Distribution Learning and Persistent Fusion for Intent-Conditioned Counterspeech Generation
Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty and Md. Shad Akhtar   11:00-12:30 (Pier 7&8)
Counterspeech has been demonstrated to be an efficacious approach for combating hate speech. While various conventional and controlled approaches have been studied in recent years to generate counterspeech, a counterspeech with a certain intent may not be sufficient in every scenario. Due to the complex and multifaceted nature of hate speech, utilizing multiple forms of counter-narratives with varying intents may be advantageous in different circumstances. In this paper, we explore intent-conditioned counterspeech generation. At first, we develop IntentCONAN, a diversified intent-specific counterspeech dataset with 6831 counterspeeches conditioned on five intents, i.e., informative, denouncing, question, positive, and humour. Subsequently, we propose QUARC, a two-stage framework for intent-conditioned counterspeech generation. QUARC leverages vector-quantized representations learned for each intent category along with PerFuMe, a novel fusion module to incorporate intent-specific information into the model. Our evaluation demonstrates that QUARC outperforms several baselines by an average of 10% across evaluation metrics. An extensive human evaluation supplements our hypothesis of better and more appropriate responses than comparative systems.

Race, Gender, and Age Biases in Biomedical Masked Language Models
Michelle YoungJin Kim, Junghwan Kim and Kristen Johnson   11:00-12:30 (Pier 7&8)
Biases cause discrepancies in healthcare services. Race, gender, and age of a patient affect interactions with physicians and the medical treatments one receives. These biases in clinical practices can be amplified following the release of pre-trained language models trained on biomedical corpora. To bring awareness to such repercussions, we examine social biases present in the biomedical masked language models. We curate prompts based on evidence-based practice and compare generated diagnoses based on biases. For a case study, we measure bias in diagnosing coronary artery disease and using cardiovascular procedures based on bias. Our study demonstrates that biomedical models are less biased than BERT in gender, while the opposite is true for race and age.

Text Augmentation Using Dataset Reconstruction for Low-Resource Classification
Adir Rahamim, Guy Uziel, Esther Goldbraich and Ateret Anaby Tavor   11:00-12:30 (Pier 7&8)
In the deployment of real-world text classification models, label scarcity is a common problem and as the number of classes increases, this

problem becomes even more complex. An approach to addressing this problem is by applying text augmentation methods.

One of the more prominent methods involves using the text-generation capabilities of language models. In this paper, we propose Text AUgmentation by Dataset Reconstruction (TAU-DR), a novel method of data augmentation for text classification. We conduct experiments on several multi-class datasets, showing that our approach improves the current state-of-the-art techniques for data augmentation.

To Copy Rather Than Memorize: A Vertical Learning Paradigm for Knowledge Graph Completion

Rui Li, Xu Chen, Chaozhuo Li, Yanming Shen, Jianan Zhao, Yujing Wang, Weihao Han, Hao Sun, Weiwei Deng, Qi Zhang and Xing Xie
11:00-12:30 (Pier 7&8)

Embedding models have shown great power in knowledge graph completion (KGC) task. By learning structural constraints for each training triple, these methods implicitly memorize intrinsic relation rules to infer missing links. However, this paper points out that the multi-hop relation rules are hard to be reliably memorized due to the inherent deficiencies of such implicit memorization strategy, making embedding models underperform in predicting links between distant entity pairs. To alleviate this problem, we present Vertical Learning Paradigm (VLP), which extends embedding models by allowing to explicitly copy target information from related factual triples for more accurate prediction. Rather than solely relying on the implicit memory, VLP directly provides additional cues to improve the generalization ability of embedding models, especially making the distant link prediction significantly easier. Moreover, we also propose a novel relative distance based negative sampling technique (ReD) for more effective optimization. Experiments demonstrate the validity and generality of our proposals on two standard benchmarks. Our code is available at https://github.com/rui9812/VLP.

LABO: Towards Learning Optimal Label Regularization via Bi-level Optimization

Peng Lu, Ahmad Rashid, Ivan Kobyzev, Mehdi Rezagholizadeh and Phillippe Langlais   11:00-12:30 (Pier 7&8)

Regularization techniques are crucial to improving the generalization performance and training efficiency of deep neural networks. Many deep learning algorithms rely on weight decay, dropout, batch/layer normalization to converge faster and generalize. Label Smoothing (LS) is another simple, versatile and efficient regularization which can be applied to various supervised classification tasks. Conventional LS, however, regardless of the training instance assumes that each non-target class is equally likely. In this work, we present a general framework for training with label regularization, which includes conventional LS but can also model instance-specific variants. Based on this formulation, we propose an efficient way of learning LAbel regularization by devising a Bi-level Optimization (LABO) problem. We derive a deterministic and interpretable solution of the inner loop as the optimal label smoothing without the need to store the parameters or the output of a trained model. Finally, we conduct extensive experiments and demonstrate our LABO consistently yields improvement over conventional label regularization on various fields, including seven machine translation and three image classification tasks across various neural network architectures while maintaining training efficiency.

Structured Pruning for Efficient Generative Pre-trained Language Models

Chaofan Tao, Lu Hou, Haoli Bai, Jiansheng Wei, Xin Jiang, Qun Liu, Ping Luo and Ngai Wong   11:00-12:30 (Pier 7&8)

The increasing sizes of large generative Pre-trained Language Models (PLMs) hinder their deployment in real-world applications. To obtain efficient PLMs, previous studies mostly focus on pruning the attention heads and feed-forward networks (FFNs) of the Transformer. Nevertheless, we find that in generative PLMs, the hidden dimension shared by many other modules (e.g., embedding layer and layer normalization) contains persistent outliers regardless of the network input. This study comprehensively investigates the structured pruning of generative PLMs with all the above compressible components. To identify redundant network structures, we assign learnable masks over compressible components followed by sparse training. Various sizes of PLMs can be flexibly extracted via different thresholds, and are then task-specifically fine-tuned for further improvement. Extensive experiments on language modeling, summarization and machine translation validate the effectiveness of the proposed method. For example, the pruned BART brings 1.51x/6.96x inference speedup on GPU/CPU with 67% size reduction, and can be further combined with quantization for more than $25\times$ compression.

Self-Distilled Quantization: Achieving High Compression Rates in Transformer-Based Language Models

James O'Neill and Sourav Dutta   11:00-12:30 (Pier 7&8)

We investigate the effects of post-training quantization and quantization-aware training on the generalization of Transformer language models. We present a new method called self-distilled quantization (SDQ) that minimizes accumulative quantization errors and outperforms baselines. We apply SDQ to multilingual models XLM-R$_{Base}$ and InfoXLM$_{Base}$ and demonstrate that both models can be reduced from 32-bit floating point weights to 8-bit integer weights while maintaining a high level of performance on the XGLUE benchmark. Our results also highlight the challenges of quantizing multilingual models, which must generalize to languages they were not fine-tuned on.

Multitask Pre-training of Modular Prompt for Chinese Few-Shot Learning

Tianxiang Sun, Zhengfu He, Qin Zhu, Xipeng Qiu and Xuanjing Huang   11:00-12:30 (Pier 7&8)

Prompt tuning is a parameter-efficient approach to adapting pre-trained language models to downstream tasks. Although prompt tuning has been shown to match the performance of full model tuning when training data is sufficient, it tends to struggle in few-shot learning settings. In this paper, we present Multi-task Pre-trained Modular Prompt (MP2) to boost prompt tuning for few-shot learning. MP2 is a set of combinable prompts pre-trained on 38 Chinese tasks. On downstream tasks, the pre-trained prompts are selectively activated and combined, leading to strong compositional generalization to unseen tasks. To bridge the gap between pre-training and fine-tuning, we formulate upstream and downstream tasks into a unified machine reading comprehension task. Extensive experiments under two learning paradigms, i.e., gradient descent and black-box tuning, show that MP2 significantly outperforms prompt tuning, full model tuning, and prior prompt pre-training methods in few-shot settings. In addition, we demonstrate that MP2 can achieve surprisingly fast and strong adaptation to downstream tasks by merely learning 8 parameters to combine the pre-trained modular prompts.

MatSci-NLP: Evaluating Scientific Language Models on Materials Science Language Tasks Using Text-to-Schema Modeling

Yu Song, Santiago Miret and Bang Liu   11:00-12:30 (Pier 7&8)

We present MatSci-NLP, a natural language benchmark for evaluating the performance of natural language processing (NLP) models on materials science text. We construct the benchmark from publicly available materials science text data to encompass seven different NLP tasks, including conventional NLP tasks like named entity recognition and relation classification, as well as NLP tasks specific to materials science, such as synthesis action retrieval which relates to creating synthesis procedures for materials. We study various BERT-based models pretrained on different scientific text corpora on MatSci-NLP to understand the impact of pretraining strategies on understanding materials science text. Given the scarcity of high-quality annotated data in the materials science domain, we perform our fine-tuning experiments with limited training data to encourage the generalize across MatSci-NLP tasks. Our experiments in this low-resource training setting show that language models pretrained on scientific text outperform BERT trained on general text. MatBERT, a model pretrained specifically on materials science journals, generally performs best for most tasks. Moreover, we propose a unified text-to-schema for multitask learning on {pasted macro 'BENCHMARK'} and compare its performance with traditional fine-tuning methods. In our analysis of different training methods, we find that our proposed text-to-schema methods inspired by question-answering consistently outperform single and multitask NLP fine-tuning methods. The code and datasets are publicly available https://github.com/BangLab-UdeM-Mila/NLP4MatSci-ACL23.

HiTIN: Hierarchy-aware Tree Isomorphism Network for Hierarchical Text Classification

He Zhu, Chong Zhang, Junjie Huang, Junran Wu and Ke Xu   11:00-12:30 (Pier 7&8)
Hierarchical text classification (HTC) is a challenging subtask of multi-label classification as the labels form a complex hierarchical structure. Existing dual-encoder methods in HTC achieve weak performance gains with huge memory overheads and their structure encoders heavily rely on domain knowledge. Under such observation, we tend to investigate the feasibility of a memory-friendly model with strong generalization capability that could boost the performance of HTC without prior statistics or label semantics. In this paper, we propose Hierarchy-aware Tree Isomorphism Network (HiTIN) to enhance the text representations with only syntactic information of the label hierarchy. Specifically, we convert the label hierarchy into an unweighted tree structure, termed coding tree, with the guidance of structural entropy. Then we design a structure encoder to incorporate hierarchy-aware information in the coding tree into text representations. Besides the text encoder, HiTIN only contains a few multi-layer perceptions and linear transformations, which greatly saves memory. We conduct experiments on three commonly used datasets and the results demonstrate that HiTIN could achieve better test performance and less memory consumption than state-of-the-art (SOTA) methods.

PAD-Net: An Efficient Framework for Dynamic Networks
Shwai He, Liang Ding, Daize Dong, Boan Liu, Fuqiang Yu and Dacheng Tao   11:00-12:30 (Pier 7&8)
Dynamic networks, e.g., Dynamic Convolution (DY-Conv) and the Mixture of Experts (MoE), have been extensively explored as they can considerably improve the model's representation power with acceptable computational cost. The common practice in implementing dynamic networks is to convert the given static layers into fully dynamic ones where all parameters are dynamic (at least within a single layer) and vary with the input. However, such a fully dynamic setting may cause redundant parameters and high deployment costs, limiting the applicability of dynamic networks to a broader range of tasks and models. The main contributions of our work are challenging the basic commonsense in dynamic networks and proposing a partially dynamic network, namely PAD-Net, to transform the redundant dynamic parameters into static ones. Also, we further design Iterative Mode Partition to partition dynamic and static parameters efficiently. Our method is comprehensively supported by large-scale experiments with two typical advanced dynamic architectures, i.e., DY-Conv and MoE, on both image classification and GLUE benchmarks. Encouragingly, we surpass the fully dynamic networks by $+0.7\%$ top-1 acc with only 30% dynamic parameters for ResNet-50 and $+1.9\%$ average score in language understanding with only 50% dynamic parameters for BERT. Code will be released at: https://github.com/Shwai-He/PAD-Net.

CAME: Confidence-guided Adaptive Memory Efficient Optimization
Yang Luo, Xiaozhe REN, Zangwei Zheng, ZHUO JIANG, Xin Jiang and Yang You   11:00-12:30 (Pier 7&8)
Adaptive gradient methods, such as Adam and LAMB, have demonstrated excellent performance in the training of large language models. Nevertheless, the need for adaptivity requires maintaining second-moment estimates of the per-parameter gradients, which entails a high cost of extra memory overheads. To solve this problem, several memory-efficient optimizers (e.g., Adafactor) have been proposed to obtain a drastic reduction in auxiliary memory usage, but with a performance penalty. In this paper, we first study a confidence-guided strategy to reduce the instability of existing memory efficient optimizers. Based on this strategy, we propose CAME to simultaneously achieve two goals: fast convergence as in traditional adaptive methods, and low memory usage as in memory-efficient methods. Extensive experiments demonstrate the training stability and superior performance of CAME across various NLP tasks such as BERT and GPT-2 training. Notably, for BERT pre-training on the large batch size of 32,768, our proposed optimizer attains faster convergence and higher accuracy compared with the Adam optimizer. The implementation of CAME is publicly available.

Connectivity Patterns are Task Embeddings
Zhiheng Xi, Rui Zheng, Yuansen Zhang, Xuanjing Huang, Zhongyu Wei, Minlong Peng, Mingming Sun, Qi Zhang and Tao Gui   11:00-12:30 (Pier 7&8)
Task embeddings are task-specific vectors designed to construct a semantic space of tasks, which can be used to predict the most transferable source task for a given target task via the similarity between task embeddings. However, existing methods use optimized parameters and representations as task embeddings, resulting in substantial computational complexity and storage requirements. In this work, we draw inspiration from the operating mechanism of deep neural networks (DNNs) and biological brains, where neuronal activations are sparse and task-specific, and we use the connectivity patterns of neurons as a unique identifier associated with the task. The proposed method learns to assign importance masks for sub-structures of DNNs, and accordingly indicate the task-specific connectivity patterns. In addition to the storage advantages brought by the binary masking mechanism and structured sparsity, the early-bird nature of the sparse optimization process can deliver an efficient computation advantage. Experiments show that our method consistently outperforms other baselines in predicting inter-task transferability across data regimes and transfer settings, while keeping high efficiency in computation and storage.

Low-Rank Updates of pre-trained Weights for Multi-Task Learning
Alexandre Daniel AUDIBERT, Massih R Amini, Konstantin Usevich and Marianne Clausel   11:00-12:30 (Pier 7&8)
Multi-Task Learning used with pre-trained models has been quite popular in the field of Natural Language Processing in recent years. This framework remains still challenging due to the complexity of the tasks and the challenges associated with fine-tuning large pre-trained models. In this paper, we propose a new approach for Multi-task learning which is based on stacking the weights of Neural Networks as a tensor. We show that low-rank updates in the canonical polyadic tensor decomposition of this tensor of weights lead to a simple, yet efficient algorithm, which without loss of performance allows to reduce considerably the model parameters. We investigate the interactions between tasks inside the model as well as the inclusion of sparsity to find the best tensor rank and to increase the compression rate. Our strategy is consistent with recent efforts that attempt to use constraints to fine-tune some model components. More precisely, we achieve equivalent performance as the state-of-the-art on the General Language Understanding Evaluation benchmark by training only 0.3 of the parameters per task while not modifying the baseline weights.

LaSQuE: Improved Zero-Shot Classification from Explanations Through Quantifier Modeling and Curriculum Learning
Sayan Ghosh, Rakesh R. Menon and Shashank Srivastava   11:00-12:30 (Pier 7&8)
A hallmark of human intelligence is the ability to learn new concepts purely from language. Several recent approaches have explored training machine learning models via natural language supervision. However, these approaches fall short in leveraging linguistic quantifiers (such as 'always' or 'rarely') and mimicking humans in compositionally learning complex tasks. Here, we present LaSQuE, a method that can learn zero-shot classifiers from language explanations by using three new strategies - (1) modeling the semantics of linguistic quantifiers in explanations (including exploiting ordinal strength relationships, such as 'always' > 'likely'), (2) aggregating information from multiple explanations using an attention-based mechanism, and (3) model training via curriculum learning. With these strategies, LaSQuE outperforms prior work, showing an absolute gain of up to 7% in generalizing to unseen real-world classification tasks.

Which Examples Should be Multiply Annotated? Active Learning When Annotators May Disagree
Connor T Baumler, Anna Sotnikova and Hal Daumé III   11:00-12:30 (Pier 7&8)
Linguistic annotations, especially for controversial topics like hate speech detection, are frequently contested due to annotator backgrounds and positionalities. In such situations, preserving this disagreement through the machine learning pipeline can be important for downstream use cases. However, capturing disagreement can increase annotation time and expense. Fortunately, for many tasks, not all examples are equally controversial; we develop an active learning approach, Disagreement Aware Active Learning (DAAL) that concentrates annotations

on examples where model entropy and annotator entropy are the most different. Because we cannot know the true entropy of annotations on unlabeled examples, we estimate a model that predicts annotator entropy trained using very few multiply-labeled examples. We find that traditional uncertainty-based active learning underperforms simple passive learning on tasks with high levels of disagreement, but that our active learning approach is able to successfully improve on passive and active baselines, reducing the number of annotations required by at least 24% on average across several datasets.

Know Where You're Going: Meta-Learning for Parameter-Efficient Fine-Tuning
Mozhdeh Gheini, Xuezhe Ma and Jonathan May 11:00-12:30 (Pier 7&8)
A recent family of techniques, dubbed lightweight fine-tuning methods, facilitates parameter-efficient transfer by updating only a small set of additional parameters while keeping the parameters of the original model frozen. While proven to be an effective approach, there are no existing studies on if and how such knowledge of the downstream fine-tuning approach calls for complementary measures after pre-training and before fine-tuning. In this work, we show that taking the ultimate choice of fine-tuning into consideration boosts the performance of parameter-efficient fine-tuning. By relying on optimization-based meta-learning using MAML with certain modifications for our distinct purpose, we prime the pre-trained model specifically for parameter-efficient fine-tuning, resulting in gains of up to 4.96 points on cross-lingual NER fine-tuning. Our ablation settings and analyses further reveal that the specific approach we take to meta-learning is crucial for the attained gains.

Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes
Cheng-Yu Hsieh, Chun-Liang Li, CHIH-KUAN YEH, Hootan Nakhost, Yasuhisa Fujii, Alex Jason Ratner, Ranjay Krishna, Chen-Yu Lee and Tomas Pfister 11:00-12:30 (Pier 7&8)
Deploying large language models (LLMs) is challenging because they are memory inefficient and compute-intensive for practical applications. In reaction, researchers train smaller task-specific models by either finetuning with human labels or distilling using LLM-generated labels. However, finetuning and distillation require large amounts of training data to achieve comparable performance to LLMs. We introduce Distilling step-by-step, a new mechanism that (a) trains smaller models that outperform LLMs, and (b) achieves so by leveraging less training data needed by finetuning or distillation. Our method extracts LLM rationales as additional supervision for training small models within a multi-task framework. We present three findings across 4 NLP benchmarks: First, compared to both finetuning and distillation, our mechanism achieves better performance with much fewer labeled/unlabeled training examples. Second, compared to few-shot prompted LLMs, we achieve better performance using substantially smaller model sizes. Third, we reduce both the model size and the amount of data required to outperform LLMs; our finetuned 770M T5 model outperforms the few-shot prompted 540B PaLM model using only 80% of available data on a benchmark, whereas standard finetuning the same T5 model struggles to match even by using 100% of the dataset.

TADA: Efficient Task-Agnostic Domain Adaptation for Transformers
Chia-Chien Hung, Lukas Lange and Jannik Strötgen 11:00-12:30 (Pier 7&8)
Intermediate training of pre-trained transformer-based language models on domain-specific data leads to substantial gains for downstream tasks. To increase efficiency and prevent catastrophic forgetting alleviated from full domain-adaptive pre-training, approaches such as adapters have been developed. However, these require additional parameters for each layer, and are criticized for their limited expressiveness. In this work, we introduce TADA, a novel task-agnostic domain adaptation method which is modular, parameter-efficient, and thus, data-efficient. Within TADA, we retrain the embeddings to learn domain-aware input representations and tokenizers for the transformer encoder, while freezing all other parameters of the model. Then, task-specific fine-tuning is performed. We further conduct experiments with meta-embeddings and newly introduced meta-tokenizers, resulting in one model per task in multi-domain use cases. Our broad evaluation in 4 downstream tasks for 14 domains across single- and multi-domain setups and high- and low-resource scenarios reveals that TADA is an effective and efficient alternative to full domain-adaptive pre-training and adapters for domain adaptation, while not introducing additional parameters or complex training steps.

Exclusive Supermask Subnetwork Training for Continual Learning
Prateek Yadav and Mohit Bansal 11:00-12:30 (Pier 7&8)
Continual Learning (CL) methods focus on accumulating knowledge over time while avoiding catastrophic forgetting. Recently, Wortsman et al. (2020) proposed a CL method, SupSup, which uses a randomly initialized, fixed base network (model) and finds a supermask for each new task that selectively keeps or removes each weight to produce a subnetwork. They prevent forgetting as the network weights are not being updated. Although there is no forgetting, the performance of SupSup is sub-optimal because fixed weights restrict its representational power. Furthermore, there is no accumulation or transfer of knowledge inside the model when new tasks are learned. Hence, we propose ExSSNeT (Exclusive Supermask SubNetwork Training), that performs exclusive and non-overlapping subnetwork weight training. This avoids conflicting updates to the shared weights by subsequent tasks to improve performance while still preventing forgetting. Furthermore, we propose a novel KNN-based Knowledge Transfer (KKT) module that utilizes previously acquired knowledge to learn new tasks better and faster. We demonstrate that ExSSNeT outperforms strong previous methods on both NLP and Vision domains while preventing forgetting. Moreover, ExSSNeT is particularly advantageous for sparse masks that activate 2-10% of the model parameters, resulting in an average improvement of 8.3% over SupSup. Furthermore, ExSSNeT scales to a large number of tasks (100).

History repeats: Overcoming catastrophic forgetting for event-centric temporal knowledge graph completion
Mehrnoosh Mirtaheri, Mohammad Rostami and Aram Galstyan 11:00-12:30 (Pier 7&8)
Temporal knowledge graph (TKG) completion models typically rely on having access to the entire graph during training. However, in real-world scenarios, TKG data is often received incrementally as events unfold, leading to a dynamic non-stationary data distribution over time. While one could incorporate fine-tuning to existing methods to allow them to adapt to evolving TKG data, this can lead to forgetting previously learned patterns. Alternatively, retraining the model with the entire updated TKG can mitigate forgetting but is computationally burdensome. To address these challenges, we propose a general continual training framework that is applicable to any TKG completion method, and leverages two key ideas: (i) a temporal regularization that encourages repurposing of less important model parameters for learning new knowledge, and (ii) a clustering-based experience replay that reinforces the past knowledge by selectively preserving only a small portion of the past data. Our experimental results on widely used event-centric TKG datasets demonstrate the effectiveness of our proposed continual training framework in adapting to new events while reducing catastrophic forgetting. Further, we perform ablation studies to show the effectiveness of each component of our proposed framework. Finally, we investigate the relation between the memory dedicated to experience replay and the benefit gained from our clustering-based sampling strategy.

Grokking of Hierarchical Structure in Vanilla Transformers
Shikhar Murty, Pratyusha Sharma, Jacob Andreas and Christopher D. Manning 11:00-12:30 (Pier 7&8)
For humans, language production and comprehension is sensitive to the hierarchical structure of sentences. In natural language processing, past work has questioned how effectively neural sequence models like transformers capture this hierarchical structure when generalizing to structurally novel inputs. We show that transformer language models can learn to generalize hierarchically after training for extremely long periods—far beyond the point when in-domain accuracy has saturated. We call this phenomenon structural grokking. On multiple datasets, structural grokking exhibits inverted U-shaped scaling in model depth: intermediate-depth models generalize better than both very deep and

very shallow transformers. When analyzing the relationship between model-internal properties and grokking, we find that optimal depth for grokking can be identified using the tree-structuredness metric of CITATION. Overall, our work provides strong evidence that, with extended training, vanilla transformers discover and use hierarchical structure.

Prototype-Guided Pseudo Labeling for Semi-Supervised Text Classification
Weiyi Yang, Richong Zhang, Junfan Chen, Lihong Wang and Jaein Kim   11:00-12:30 (Pier 7&8)
Semi-supervised text classification (SSTC) aims at text classification with few labeled data and massive unlabeled data. Recent works achieve this task by pseudo-labeling methods, with the belief that the unlabeled and labeled data have identical data distribution, and assign the unlabeled data with pseudo-labels as additional supervision. However, existing pseudo-labeling methods usually suffer from ambiguous categorical boundary issues when training the pseudo-labeling phase, and simply select pseudo-labels without considering the unbalanced categorical distribution of the unlabeled data, making it difficult to generate reliable pseudo-labels for each category. We propose a novel semi-supervised framework, namely ProtoS2̂, with prototypical cluster separation (PCS) and prototypical-center data selection (CDS) technology to address the issue. Particularly, PCS exploits categorical prototypes to assimilate instance representations within the same category, thus emphasizing low-density separation for the pseudo-labeled data to alleviate ambiguous boundaries. Besides, CDS selects central pseudo-labeled data considering the categorical distribution, avoiding the model from biasing on dominant categories. Empirical studies and extensive analysis with four benchmarks demonstrate the effectiveness of the proposed model.

When and how to paraphrase for named entity recognition?
Saket Sharma, Aviral Joshi, Yiyun Zhao, Namrata Mukhija, Hanoz Bhathena, Prateek Singh and Sashank Santhanam 11:00-12:30 (Pier 7&8)
While paraphrasing is a promising approach for data augmentation in classification tasks, its effect on named entity recognition (NER) is not investigated systematically due to the difficulty of span-level label preservation. In this paper, we utilize simple strategies to annotate entity spans in generations and compare established and novel methods of paraphrasing in NLP such as back translation, specialized encoder-decoder models such as Pegasus, and GPT-3 variants for their effectiveness in improving downstream performance for NER across different levels of gold annotations and paraphrasing strength on 5 datasets. We thoroughly explore the influence of paraphrasers, and dynamics between paraphrasing strength and gold dataset size on the NER performance with visualizations and statistical testing. We find that the choice of the paraphraser greatly impacts NER performance, with one of the larger GPT-3 variants exceedingly capable of generating high quality paraphrases, yielding statistically significant improvements in NER performance with increasing paraphrasing strength, while other paraphrasers show more mixed results. Additionally, inline gold annotations generated by larger GPT-3 are strictly better than heuristic based annotations. We also find diminishing benefits of paraphrasing as gold annotations increase for most datasets. Furthermore, while most paraphrasers promote entity memorization in NER, the proposed GPT-3 configuration performs most favorably among the compared paraphrasers when tested on unseen entities, with memorization reducing further with paraphrasing strength. Finally, we explore mention replacement using GPT-3, which provides additional benefits over base paraphrasing for specific datasets.

A Universal Discriminator for Zero-Shot Generalization
Haike Xu, Zongyu Lin, Jing Zhou, Yanan Zheng and Zhilin Yang   11:00-12:30 (Pier 7&8)
Generative modeling has been the dominant approach for large-scale pretraining and zero-shot generalization. In this work, we challenge this convention by showing that discriminative approaches perform substantially better than generative ones on a large number of NLP tasks. Technically, we train a single discriminator to predict whether a text sample comes from the true data distribution, similar to GANs. Since many NLP tasks can be formulated as selecting from a few options, we use this discriminator to predict the concatenation of input and which option has the highest probability of coming from the true data distribution. This simple formulation achieves state-of-the-art zero-shot results on the T0 benchmark, outperforming T0 by 16.0%, 7.8%, and 11.5% respectively on different scales. In the finetuning setting, our approach also achieves new state-of-the-art results on a wide range of NLP tasks, with only 1/4 parameters of previous methods. Meanwhile, our approach requires minimal prompting efforts, which largely improves robustness and is essential for real-world applications. Furthermore, we also jointly train a generalized UD in combination with generative tasks, which maintains its advantage on discriminative tasks and simultaneously works on generative tasks.

On Dataset Transferability in Active Learning for Transformers
Fran Jelenić, Josip Jukić, Nina Drobac and Jan Snajder   11:00-12:30 (Pier 7&8)
Active learning (AL) aims to reduce labeling costs by querying the examples most beneficial for model learning. While the effectiveness of AL for fine-tuning transformer-based pre-trained language models (PLMs) has been demonstrated, it is less clear to what extent the AL gains obtained with one model transfer to others. We consider the problem of transferability of actively acquired datasets in text classification and investigate whether AL gains persist when a dataset built using AL coupled with a specific PLM is used to train a different PLM. We link the AL dataset transferability to the similarity of instances queried by the different PLMs and show that AL methods with similar acquisition sequences produce highly transferable datasets regardless of the models used. Additionally, we show that the similarity of acquisition sequences is influenced more by the choice of the AL method than the choice of the model.

Rethinking Semi-supervised Learning with Language Models
Zhengxiang Shi, Francesco Tonolini, Nikolaos Aletras, Emine Yilmaz, Gabriella Kazai and Yunlong Jiao   11:00-12:30 (Pier 7&8)
Semi-supervised learning (SSL) is a popular setting aiming to effectively utilize unlabelled data to improve model performance in downstream natural language processing (NLP) tasks. Currently, there are two popular approaches to make use of the unlabelled data: Self-training (ST) and Task-adaptive pre-training (TAPT). ST uses a teacher model to assign pseudo-labels to the unlabelled data, while TAPT continues pre-training on the unlabelled data before fine-tuning. To the best of our knowledge, the effectiveness of TAPT in SSL tasks has not been systematically studied, and no previous work has directly compared TAPT and ST in terms of their ability to utilize the pool of unlabelled data. In this paper, we provide an extensive empirical study comparing five state-of-the-art ST approaches and TAPT across various NLP tasks and data sizes, including in- and out-of-domain settings. Surprisingly, we find that TAPT is a strong and more robust SSL learner, even when using just a few hundred unlabelled samples or in the presence of domain shifts, compared to more sophisticated ST approaches, and tends to bring greater improvements in SSL than in fully-supervised settings. Our further analysis demonstrates the risks of using ST approaches when the size of labelled or unlabelled data is small or when domain shifts exist, and highlights TAPT as a potential solution.

Leveraging Training Data in Few-Shot Prompting for Numerical Reasoning
Zhanming Jie and Wei Lu   11:00-12:30 (Pier 7&8)
Chain-of-thought (CoT) prompting with large language models has proven effective in numerous natural language process tasks, but designing prompts that generalize well to diverse problem types can be challenging CITATION, especially in the context of math word problem solving. Additionally, it is common to have a large amount of training data that have a better diversity coverage but CoT annotations are not available, which limits the use of supervised learning techniques. To address these issues, we investigate two approaches to leverage the training data in few-shot prompting scenario: *dynamic program prompting* and *program distillation*. Our approach is largely inspired by CITATION where they proposed to replace the CoT with the programs as the intermediate reasoning step. Such a prompting strategy allows us to accurately verify the answer correctness through program execution in MWP solving. Our dynamic program prompting involves annotating the training data by sampling correct programs from a large language model, while program distillation involves adapting a smaller model

to the program-annotated training data. Our experiments on three standard MWP datasets demonstrate the effectiveness of these approaches, yielding significant improvements over previous baselines for prompting and fine-tuning. Our results suggest that leveraging a large amount of training data can improve the generalization ability of prompts and boost the performance of fine-tuned smaller models in MWP solving.

The Larger they are, the Harder they Fail: Language Models do not Recognize Identifier Swaps in Python
Antonio Valerio Miceli Barone, Fazl Barez, Shay B. Cohen and Ioannis Konstas　　　　　　　　11:00-12:30 (Pier 7&8)

Large Language Models (LLMs) have successfully been applied to code generation tasks, raising the question of how well these models understand programming. Typical programming languages have invariances and equivariances in their semantics that human programmers intuitively understand and exploit, such as the (near) invariance to the renaming of identifiers. We show that LLMs not only fail to properly generate correct Python code when default function names are swapped, but some of them even become more confident in their incorrect predictions as the model size increases, an instance of the recently discovered phenomenon of Inverse Scaling, which runs contrary to the commonly observed trend of increasing prediction quality with increasing model size. Our findings indicate that, despite their astonishing typical-case performance, LLMs still lack a deep, abstract understanding of the content they manipulate, making them unsuitable for tasks that statistically deviate from their training data, and that mere scaling is not enough to achieve such capability.

Evaluating the Factual Consistency of Large Language Models Through News Summarization
Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal and Colin Raffel　　　　　11:00-12:30 (Pier 7&8)

While large language models (LLMs) have proven to be effective on a large variety of tasks, they are also known to hallucinate information. To measure whether an LLM prefers factually consistent continuations of its input, we propose a new benchmark called FIB (Factual Inconsistency Benchmark) that focuses on the task of summarization. Specifically, our benchmark involves comparing the scores an LLM assigns to a factually consistent versus a factually inconsistent summary for an input news article. For factually consistent summaries, we use human-written reference summaries that we manually verify as factually consistent. To generate summaries that are factually inconsistent, we generate summaries from a suite of summarization models that we have manually annotated as factually inconsistent. A model's factual consistency is then measured according to its accuracy, i.e. the proportion of documents where it assigns a higher score to the factually consistent summary. To validate the usefulness of {pasted macro 'BENCHMARK'}, we evaluate 23 large language models ranging from 1B to 176B parameters from six different model families including BLOOM and OPT. We find that existing LLMs generally assign a higher score to factually consistent summaries than to factually inconsistent summaries. However, if the factually inconsistent summaries occur verbatim in the document, then LLMs assign a higher score to these factually inconsistent summaries than factually consistent summaries. We validate design choices in our benchmark including the scoring method and source of distractor summaries.

A Length-Extrapolatable Transformer
Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song and Furu Wei　　11:00-12:30 (Pier 7&8)

Position modeling plays a critical role in Transformers. In this paper, we focus on length extrapolation, i.e., training on short texts while evaluating longer sequences. We define *attention resolution* as an indicator of extrapolation. Then we propose two designs to improve the above metric of Transformers. Specifically, we introduce a relative position embedding to explicitly maximize attention resolution. Moreover, we use blockwise causal attention during inference for better resolution. We evaluate different Transformer variants with language modeling. Experimental results show that our model achieves strong performance in both interpolation and extrapolation settings. The code will be available at https://aka.ms/LeX-Transformer.

Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models
Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee and Ee-Peng Lim　　　　　11:00-12:30 (Pier 7&8)

Large language models (LLMs) have recently been shown to deliver impressive performance in various NLP tasks. To tackle multi-step reasoning tasks, Few-shot chain-of-thought (CoT) prompting includes a few manually crafted step-by-step reasoning demonstrations which enable LLMs to explicitly generate reasoning steps and improve their reasoning task accuracy. To eliminate the manual efforts, Zero-shot-CoT concatenates the target problem statement with "*Let's think step by step*" as an input prompt to LLMs. Despite the success of Zero-shot-CoT, it still suffers from three pitfalls: calculation errors, missing-step errors, and semantic misunderstanding errors. To address the missing-step errors, we propose Plan-and-Solve (PS) Prompting. It consists of two components: first, devising a plan to divide the entire task into smaller subtasks, and then carrying out the subtasks according to the plan. To address the calculation errors and improve the quality of generated reasoning steps, we extend PS prompting with more detailed instructions and derive PS+ prompting. We evaluate our proposed prompting strategy on ten datasets across three reasoning problems. The experimental results over GPT-3 show that our proposed zero-shot prompting consistently outperforms Zero-shot-CoT across all datasets by a large margin, is comparable to or exceeds Zero-shot-Program-of-Thought Prompting, and has comparable performance with 8-shot CoT prompting on the math reasoning problem. The code can be found at https://github.com/AGI-Edgerunners/Plan-and-Solve-Prompting.

Revisiting Automated Prompting: Are We Actually Doing Better?
Yulin Zhou, Yiren Zhao, Ilia Shumailov, Robert Mullins and Yarin Gal　　　　　　　　　　　11:00-12:30 (Pier 7&8)

Current literature demonstrates that Large Language Models (LLMs) are great few-shot learners, and prompting significantly increases their performance on a range of downstream tasks in a few-shot learning setting. An attempt to automate human-led prompting followed, with some progress achieved. In particular, subsequent work demonstrates that automation can outperform fine-tuning in certain K-shot learning scenarios. In this paper, we revisit techniques for automated prompting on six different downstream tasks and a larger range of K-shot learning settings. We find that automated prompting does not consistently outperform simple manual prompting. Our work suggests that, in addition to fine-tuning, manual prompting should be used as a baseline in this line of research.

Pre-training Language Model as a Multi-perspective Course Learner
Beiduo Chen, Shaohan Huang, Zihan Zhang, Wu Guo, Zhenhua Ling, Haizhen Huang, Furu Wei, Weiwei Deng and Qi Zhang　　11:00-12:30 (Pier 7&8)

ELECTRA, the generator-discriminator pre-training framework, has achieved impressive semantic construction capability among various downstream tasks. Despite the convincing performance, ELECTRA still faces the challenges of monotonous training and deficient interaction. Generator with only masked language modeling (MLM) leads to biased learning and label imbalance for discriminator, decreasing learning efficiency; no explicit feedback loop from discriminator to generator results in the chasm between these two components, underutilizing the course learning. In this study, a multi-perspective course learning (MCL) method is proposed to fetch a many degrees and visual angles for sample-efficient pre-training, and to fully leverage the relationship between generator and discriminator. Concretely, three self-supervision courses are designed to alleviate inherent flaws of MLM and balance the label in a multi-perspective way. Besides, two self-correction courses are proposed to bridge the chasm between the two encoders by creating a "correction notebook" for secondary-supervision. Moreover, a course soups trial is conducted to solve the "tug-of-war" dynamics problem of MCL, evolving a stronger pre-trained model. Experimental results show that our method significantly improves ELECTRA's average performance by 2.8% and 3.2% absolute points respectively on GLUE and SQuAD 2.0 benchmarks, and overshadows recent advanced ELECTRA-style models under the same settings. The pre-trained MCL model is available at https://huggingface.co/McmanusChen/MCL-base.

Scaling Laws for BERT in Low-Resource Settings

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri and Aitor Soroa   11:00-12:30 (Pier 7&8)

Large language models are very resource intensive, both financially and environmentally, and require an amount of training data which is simply unobtainable for the majority of NLP practitioners. Previous work has researched the scaling laws of such models, but optimal ratios of model parameters, dataset size, and computation costs focused on the large scale. In contrast, we analyze the effect those variables have on the performance of language models in constrained settings, by building three lightweight BERT models (16M/51M/124M parameters) trained over a set of small corpora (5M/25M/125M words). We experiment on four languages of different linguistic characteristics (Basque, Spanish, Swahili and Finnish), and evaluate the models on MLM and several NLU tasks. We conclude that the power laws for parameters, data and compute for low-resource settings differ from the optimal scaling laws previously inferred, and data requirements should be higher. Our insights are consistent across all the languages we study, as well as across the MLM and downstream tasks. Furthermore, we experimentally establish when the cost of using a Transformer-based approach is worth taking, instead of favouring other computationally lighter solutions.

Large Language Models with Controllable Working Memory

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu and Sanjiv Kumar 11:00-12:30 (Pier 7&8)

Large language models (LLMs) have led to a series of breakthroughs in natural language processing (NLP), partly owing to the massive amounts of world knowledge they memorize during pretraining. While many downstream applications provide the model with an informational context to aid its underlying task, how the model's world knowledge interacts with the factual information presented in the context remains under explored. As a desirable behavior, an LLM should give precedence to the context whenever it contains task-relevant information that conflicts with the model's memorized knowledge. This enables model predictions to be grounded in the context, which then facilitates updating specific model predictions without frequently retraining the model. By contrast, when the context is irrelevant to the task, the model should ignore it and fall back on its internal knowledge. In this paper, we undertake a first joint study of the aforementioned two properties, namely controllability and robustness, in the context of LLMs. We demonstrate that state-of-the-art T5 and PaLM models (both pretrained and finetuned) could exhibit low controllability and robustness that does not improve with increasing the model size. As a solution, we propose a simple yet effective method – knowledge aware finetuning (KAFT) – to strengthen both controllability and robustness by injecting counterfactual and irrelevant contexts to standard supervised datasets. Our comprehensive evaluation showcases the utility of KAFT across model architectures and sizes.

Data-Efficient Finetuning Using Cross-Task Nearest Neighbors

Hamish Ivison, Noah A. Smith, Hannaneh Hajishirzi and Pradeep Dasigi   11:00-12:30 (Pier 7&8)

Obtaining labeled data to train a model for a task of interest is often expensive. Prior work shows training models on multitask data augmented with task descriptions (prompts) effectively transfers knowledge to new tasks. Towards efficiently building task-specific models, we assume access to a small number (32-1000) of unlabeled target-task examples and use those to retrieve the most similar labeled examples from a large pool of multitask data augmented with prompts. Compared to the current practice of finetuning models on uniformly sampled prompted multitask data (e.g.: FLAN, T0), our approach of finetuning on cross-task nearest neighbors is significantly more data-efficient. Using only 2% of the data from the P3 pool without any labeled target-task data, our models outperform strong baselines trained on all available data by 3-30% on 12 out of 14 datasets representing held-out tasks including legal and scientific document QA. Similarly, models trained on cross-task nearest neighbors from SuperNaturalInstructions, representing about 5% of the pool, obtain comparable performance to state-of-the-art models on 12 held-out tasks from that pool. Moreover, the models produced by our approach also provide a better initialization than single multitask finetuned models for few-shot finetuning on target-task data, as shown by a 2-23

Towards Adaptive Prefix Tuning for Parameter-Efficient Language Model Fine-tuning

Zhen-Ru Zhang, Chuanqi Tan, Haiyang Xu, Chengyu Wang, jun huang and Songfang Huang   11:00-12:30 (Pier 7&8)

Fine-tuning large pre-trained language models on various downstream tasks with whole parameters is prohibitively expensive. Hence, Parameter-efficient fine-tuning has attracted attention that only optimizes a few task-specific parameters with the frozen pre-trained model. In this work, we focus on prefix tuning, which only optimizes continuous prefix vectors (i.e. pseudo tokens) inserted into Transformer layers. Based on the observation that the learned syntax and semantics representation varies a lot at different layers, we argue that the adaptive prefix will be further tailored to each layer than the fixed one, enabling the fine-tuning more effective and efficient. Thus, we propose Adaptive Prefix Tuning (APT) to adjust the prefix in terms of both fine-grained token level and coarse-grained layer level with a gate mechanism. Experiments on the SuperGLUE and NER datasets show the effectiveness of APT. In addition, taking the gate as a probing, we validate the efficiency and effectiveness of the variable prefix.

Multi-target Backdoor Attacks for Code Pre-trained Models

Yanzhou Li, Shangqing Liu, Kangjie Chen, Xiaofei Xie, Tianwei Zhang and Yang Liu   11:00-12:30 (Pier 7&8)

Backdoor attacks for neural code models have gained considerable attention due to the advancement of code intelligence. However, most existing works insert triggers into task-specific data for code-related downstream tasks, thereby limiting the scope of attacks. Moreover, the majority of attacks for pre-trained models are designed for understanding tasks. In this paper, we propose task-agnostic backdoor attacks for code pre-trained models. Our backdoored model is pre-trained with two learning strategies (i.e., Poisoned Seq2Seq learning and token representation learning) to support the multi-target attack of downstream code understanding and generation tasks. During the deployment phase, the implanted backdoors in the victim models can be activated by the designed triggers to achieve the targeted attack. We evaluate our approach on two code understanding tasks and three code generation tasks over seven datasets. Extensive experimental results demonstrate that our approach effectively and stealthily attacks code-related downstream tasks.

AraMUS: Pushing the Limits of Data and Model Scale for Arabic Natural Language Processing

Asaad Alghamdi, Xinyu Duan, Wei Jiang, Zhenhai Wang, Yimeng Wu, Qingrong Xia, Zhefeng Wang, Yi ZHENG, Mehdi Rezagholizadeh, baoxing Huai, Peilun Cheng and Abbas Ghaddar   11:00-12:30 (Pier 7&8)

Developing monolingual large Pre-trained Language Models (PLMs) is shown to be very successful in handling different tasks in Natural Language Processing (NLP). In this work, we present AraMUS, the largest Arabic PLM with 11B parameters trained on 529GB of high-quality Arabic textual data. AraMUS achieves state-of-the-art performances on a diverse set of Arabic classification and generative tasks. Moreover, AraMUS shows impressive few-shot learning abilities compared with the best existing Arabic PLMs.

Do Large Language Models Know What They Don't Know?

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu and Xuanjing Huang   11:00-12:30 (Pier 7&8)

Large language models (LLMs) have a wealth of knowledge that allows them to excel in various Natural Language Processing (NLP) tasks. Current research focuses on enhancing their performance within their existing knowledge. Despite their vast knowledge, LLMs are still limited by the amount of information they can accommodate and comprehend. Therefore, the ability to understand their own limitations on the unknows, referred to as self-knowledge, is of paramount importance. This study aims to evaluate LLMs' self-knowledge by assessing their ability to identify unanswerable or unknowable questions. We introduce an automated methodology to detect uncertainty in the responses of these models, providing a novel measure of their self-knowledge. We further introduce a unique dataset, SelfAware, consisting of unanswerable questions from five diverse categories and their answerable counterparts. Our extensive analysis, involving 20 LLMs including

GPT-3, InstructGPT, and LLaMA, discovering an intrinsic capacity for self-knowledge within these models. Moreover, we demonstrate that in-context learning and instruction tuning can further enhance this self-knowledge. Despite this promising insight, our findings also highlight a considerable gap between the capabilities of these models and human proficiency in recognizing the limits of their knowledge.

Rehearsal-free Continual Language Learning via Efficient Parameter Isolation
Zhicheng Wang, Yufang Liu, Tao Ji, xiaoling Wang, Yuanbin Wu, congcong jiang, ye chao, zhencong han, ling wang, xu shao and wenqiu zeng
11:00-12:30 (Pier 7&8)
We study the problem of defying catastrophic forgetting when learning a series of language processing tasks. Compared with previous methods, we emphasize the importance of not caching history tasks' data, which makes the problem more challenging. Our proposed method applies the parameter isolation strategy. For each task, it allocates a small portion of private parameters and learns them with a shared pre-trained model. To load correct parameters at testing time, we introduce a simple yet effective non-parametric method. Experiments on continual language learning benchmarks show that our method is significantly better than all existing no-data-cache methods, and is comparable (or even better) than those using historical data.

Reducing Sensitivity on Speaker Names for Text Generation from Dialogues
Qi Jia, Haifeng Tang and Kenny Q. Zhu
11:00-12:30 (Pier 7&8)
Changing speaker names consistently throughout a dialogue should not affect its meaning and corresponding outputs for text generation from dialogues. However, pre-trained language models, serving as the backbone for dialogue-processing tasks, have shown to be sensitive to nuances. This may result in unfairness in real-world applications. No comprehensive analysis of this problem has been done in the past. In this work, we propose to quantitatively measure a model's sensitivity on speaker names, and comprehensively evaluate a number of known methods for reducing speaker name sensitivity, including a novel approach of our own. Extensive experiments on multiple datasets provide a benchmark for this problem and show the favorable performance of our approach in sensitivity reduction and quality of generation.

Multi-source Semantic Graph-based Multimodal Sarcasm Explanation Generation
Liqiang Jing, Xuemeng Song, Kun Ouyang, Mengzhao Jia and Liqiang Nie
11:00-12:30 (Pier 7&8)
Multimodal Sarcasm Explanation (MuSE) is a new yet challenging task, which aims to generate a natural language sentence for a multimodal social post (an image as well as its caption) to explain why it contains sarcasm. Although the existing pioneer study has achieved great success with the BART backbone, it overlooks the gap between the visual feature space and the decoder semantic space, the object-level metadata of the image, as well as the potential external knowledge. To solve these limitations, in this work, we propose a novel mulTi-source sEmantic grAph-based Multimodal sarcasm explanation scheme, named TEAM. In particular, TEAM extracts the object-level semantic meta-data instead of the traditional global visual features from the input image. Meanwhile, TEAM resorts to ConceptNet to obtain the external related knowledge concepts for the input text and the extracted object meta-data. Thereafter, TEAM introduces a multi-source semantic graph that comprehensively characterize the multi-source (i.e., caption, object meta-data, external knowledge) semantic relations to facilitate the sarcasm reasoning. Extensive experiments on a public released dataset MORE verify the superiority of our model over cutting-edge methods.

Context-Aware Document Simplification
Liam Cripwell, Joël Legrand and Claire Gardent
11:00-12:30 (Pier 7&8)
To date, most work on text simplification has focused on sentence-level inputs. Early attempts at document simplification merely applied these approaches iteratively over the sentences of a document. However, this fails to coherently preserve the discourse structure, leading to suboptimal output quality. Recently, strategies from controllable simplification have been leveraged to achieve state-of-the-art results on document simplification by first generating a document-level plan (a sequence of sentence-level simplification operations) and using this plan to guide sentence-level simplification downstream. However, this is still limited in that the simplification model has no direct access to the local inter-sentence document context, likely having a negative impact on surface realisation. We explore various systems that use document context within the simplification process itself, either by iterating over larger text units or by extending the system architecture to attend over a high-level representation of document context. In doing so, we achieve state-of-the-art performance on the document simplification task, even when not relying on plan-guidance. Further, we investigate the performance and efficiency tradeoffs of system variants and make suggestions of when each should be preferred.

Revisiting the Architectures like Pointer Networks to Efficiently Improve the Next Word Distribution, Summarization Factuality, and Beyond
Haw-Shiuan Chang, Zonghai Yao, Alolika Gon, hong yu and Andrew McCallum
11:00-12:30 (Pier 7&8)
Is the output softmax layer, which is adopted by most language models (LMs), always the best way to compute the next word probability? Given so many attention layers in a modern transformer-based LM, are the pointer networks redundant nowadays? In this study, we discover that the answers to both questions are no. This is because the softmax bottleneck sometimes prevents the LMs from predicting the desired distribution and the pointer networks can be used to break the bottleneck efficiently. Based on the finding, we propose several softmax alternatives by simplifying the pointer networks and accelerating the word-by-word rerankers. In GPT-2, our proposals are significantly better and more efficient than mixture of softmax, a state-of-the-art softmax alternative. In summarization experiments, without very significantly decreasing its training/testing speed, our best method based on T5-Small improves factCC score by 2 points in CNN/DM and XSUM dataset, and improves MAUVE scores by 30% in BookSum paragraph-level dataset.

Focus-aware Response Generation in Inquiry Conversation
Yiquan Wu, Weiming Lu, Yating Zhang, Adam Jatowt, Jun Feng, Changlong Sun, Fei Wu and Kun Kuang
11:00-12:30 (Pier 7&8)
Inquiry conversation is a common form of conversation that aims to complete the investigation (e.g., court hearing, medical consultation and police interrogation) during which a series of focus shifts occurs. While many models have been proposed to generate a smooth response to a given conversation history, neglecting the focus can limit performance in inquiry conversation where the order of the focuses plays there a key role. In this paper, we investigate the problem of response generation in inquiry conversation by taking the focus into consideration. We propose a novel Focus-aware Response Generation (FRG) method by jointly optimizing a multi-level encoder and a set of focal decoders to generate several candidate responses that correspond to different focuses. Additionally, a focus ranking module is proposed to predict the next focus and rank the candidate responses. Experiments on two orthogonal inquiry conversation datasets (judicial, medical domain) demonstrate that our method generates results significantly better in automatic metrics and human evaluation compared to the state-of-the-art approaches.

DiffusionBERT: Improving Generative Masked Language Models with Diffusion Models
Zhengfu He, Tianxiang Sun, Qiong Tang, Kuanning Wang, Xuanjing Huang and Xipeng Qiu
11:00-12:30 (Pier 7&8)
We present DiffusionBERT, a new generative masked language model based on discrete dif- fusion models. Diffusion models and many pre-trained language models have a shared training objective, i.e., denoising, making it possible to combine the two powerful models and enjoy the best of both worlds. On the one hand, dif- fusion models offer a promising training strat- egy that helps improve the generation quality. On the other hand, pre-trained denoising lan- guage models (e.g., BERT) can be used as a good initialization that accelerates convergence. We explore training BERT to learn the reverse process of a discrete diffusion process with an absorbing state and elucidate several designs to improve it. First, we propose a new noise schedule for the forward diffusion process that controls the degree of noise added at each step

based on the information of each token. Sec- ond, we investigate several designs of incorpo- rating the time step into BERT. Experiments on unconditional text generation demonstrate that DiffusionBERT achieves significant improve- ment over existing diffusion models for text (e.g., D3PM and Diffusion-LM) and previous generative masked language models in terms of perplexity and BLEU score. Promising re- sults in conditional generation tasks show that DiffusionBERT can generate texts of compa- rable quality and more diverse than a series of established baselines.

NLG Evaluation Metrics Beyond Correlation Analysis: An Empirical Metric Preference Checklist

Iftitahu Nimah, Meng Fang, Vlado Menkovski and Mykola Pechenizkiy   11:00-12:30 (Pier 7&8)

In this study, we analyze automatic evaluation metrics for Natural Language Generation (NLG), specifically task-agnostic metrics and human-aligned metrics. Task-agnostic metrics, such as Perplexity, BLEU, BERTScore, are cost-effective and highly adaptable to diverse NLG tasks, yet they have a weak correlation with human. Human-aligned metrics (CTC, CtrlEval, UniEval) improves correlation level by incorporating desirable human-like qualities as training objective. However, their effectiveness at discerning system-level performance and quality of system outputs remain unclear.

We present metric preference checklist as a framework to assess the effectiveness of automatic metrics in three NLG tasks: Text Summa- rization, Dialogue Response Generation, and Controlled Generation. Our proposed framework provides access: (i) for verifying whether automatic metrics are faithful to human preference, regardless of their correlation level to human; and (ii) for inspecting the strengths and limitations of NLG systems via pairwise evaluation. We show that automatic metrics provide a better guidance than human on discrimi- nating system-level performance in Text Summarization and Controlled Generation tasks. We also show that multi-aspect human-aligned metric (UniEval) is not necessarily dominant over single-aspect human-aligned metrics (CTC, CtrlEval) and task-agnostic metrics (BLEU, BERTScore), particularly in Controlled Generation tasks.

DecompEval: Evaluating Generated Texts as Unsupervised Decomposed Question Answering

Pei Ke, Fei Huang, Fei Mi, Yasheng Wang, Qun Liu, Xiaoyan Zhu and Minlie Huang   11:00-12:30 (Pier 7&8)

Existing evaluation metrics for natural language generation (NLG) tasks face the challenges on generalization ability and interpretability. Specifically, most of the well-performed metrics are required to train on evaluation datasets of specific NLG tasks and evaluation dimensions, which may cause over-fitting to task-specific datasets. Furthermore, existing metrics only provide an evaluation score for each dimension without revealing the evidence to interpret how this score is obtained. To deal with these challenges, we propose a simple yet effective metric called DecompEval. This metric formulates NLG evaluation as an instruction-style question answering task and utilizes instruction-tuned pre- trained language models (PLMs) without training on evaluation datasets, aiming to enhance the generalization ability. To make the evaluation process more interpretable, we decompose our devised instruction-style question about the quality of generated texts into the subquestions that measure the quality of each sentence. The subquestions with their answers generated by PLMs are then recomposed as evidence to obtain the evaluation result. Experimental results show that DecompEval achieves state-of-the-art performance in untrained metrics for evaluating text summarization and dialogue generation, which also exhibits strong dimension-level / task-level generalization ability and interpretability.

AlignScore: Evaluating Factual Consistency with A Unified Alignment Function

Yuheng Zha, Yichi Yang, Ruichen Li and Zhiting Hu   11:00-12:30 (Pier 7&8)

Many text generation applications require the generated text to be factually consistent with input information. Automatic evaluation of factual consistency is challenging. Previous work has developed various metrics that often depend on specific functions, such as natural language inference (NLI) or question answering (QA), trained on limited data. Those metrics thus can hardly assess diverse factual inconsistencies (e.g., contradictions, hallucinations) that occur in varying inputs/outputs (e.g., sentences, documents) from different tasks. In this paper, we propose AlignScore, a new holistic metric that applies to a variety of factual inconsistency scenarios as above. AlignScore is based on a general function of information alignment between two arbitrary text pieces. Crucially, we develop a unified training framework of the alignment function by integrating a large diversity of data sources, resulting in 4.7M training examples from 7 well-established tasks (NLI, QA, paraphrasing, fact verification, information retrieval, semantic similarity, and summarization). We conduct extensive experiments on large-scale benchmarks including 22 evaluation datasets, where 19 of the datasets were never seen in the alignment training. AlignScore achieves substantial improvement over a wide range of previous metrics. Moreover, AlignScore (355M parameters) matches or even outper- forms metrics based on ChatGPT and GPT-4 that are orders of magnitude larger.

Tailor: A Soft-Prompt-Based Approach to Attribute-Based Controlled Text Generation

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen and Jun Xie   11:00-12:30 (Pier 7&8)

Attribute-based Controlled Text Generation (CTG) refers to generating sentences that satisfy desirable attributes (e.g., emotions and topics). Existing work usually utilize fine-tuning or resort to extra attribute classifiers, yet suffer from increases in storage and inference time. To ad- dress these concerns, we explore attribute-based CTG in a parameter-efficient manner. In short, the proposed Tailor represents each attribute as a pre-trained continuous vector i.e., single-attribute prompt), which guides the generation of a fixed pre-trained language model (PLM) to satisfy a pre-specified attribute. These prompts can be simply concatenated as a whole for multi-attribute CTG without any re-training. Nevertheless, this may raise problems of fluency downgrading and position sensitivity. To solve this, Tailor provides two solutions to enhance the combination. The former contains a multi-attribute prompt mask and a re-indexing position sequence to bridge the gap between the training (one single-attribute prompt for each task) and the testing stage (concatenating two prompts). The latter introduces a trainable prompt connector to further enhance the combinations. Experiments demonstrate that, only requiring 0.08% extra training parameters of the GPT-2, Tailor can achieve effective and general improvements on eleven attribute-specific generation tasks.

Unsupervised Graph-Text Mutual Conversion with a Unified Pretrained Language Model

Yi Xu, Shuqian Sheng, Jiexing Qi, Luoyi Fu, Zhouhan Lin, Xinbing Wang and Chenghu Zhou   11:00-12:30 (Pier 7&8)

Graph-to-text (G2T) generation and text-to-graph (T2G) triple extraction are two essential tasks for knowledge graphs. Existing unsupervised approaches become suitable candidates for jointly learning the two tasks due to their avoidance of using graph-text parallel data. However, they adopt multiple complex modules and still require entity information or relation type for training. To this end, we propose INFINITY, a simple yet effective unsupervised method with a unified pretrained language model that does not introduce external annotation tools or addi- tional parallel information. It achieves fully unsupervised graph-text mutual conversion for the first time. Specifically, INFINITY treats both G2T and T2G as a bidirectional sequence generation task by fine-tuning only one pretrained seq2seq model. A novel back-translation-based framework is then designed to generate synthetic parallel data automatically. Besides, we investigate the impact of graph linearization and introduce the structure-aware fine-tuning strategy to alleviate possible performance deterioration via retaining structural information in graph sequences. As a fully unsupervised framework, INFINITY is empirically verified to outperform state-of-the-art baselines for G2T and T2G tasks. Additionally, we also devise a new training setting called cross learning for low-resource unsupervised information extraction.

DivHSK: Diverse Headline Generation using Self-Attention based Keyword Selection

Venkatesh E, Kaushal Kumar Maurya, Deepak Kumar and Maunendra Sankar Desarkar   11:00-12:30 (Pier 7&8)

Diverse headline generation is an NLP task where given a news article, the goal is to generate multiple headlines that are true to the content of the article but are different among themselves. This task aims to exhibit and exploit semantically similar one-to-many relationships between a source news article and multiple target headlines. Toward this, we propose a novel model called DIVHSK. It has two components: KEYSE-

LECT for selecting the important keywords, and SEQGEN, for finally generating the multiple diverse headlines. In KEYSELECT, we cluster the self-attention heads of the last layer of the pre-trained encoder and select the most-attentive theme and general keywords from the source article. Then, cluster-specific keyword sets guide the SEQGEN, a pre-trained encoder-decoder model, to generate diverse yet semantically similar headlines. The proposed model consistently outperformed existing literature and our strong baselines and emerged as a state-of-the-art model. We have also created a high-quality multi-reference headline dataset from news articles.

Coarse-to-Fine Contribution Network for Multimodal Summarization
Min Xiao, Junnan Zhu, Haitao Lin, Yu Zhou and Chengqing Zong   11:00-12:30 (Pier 7&8)
Multimodal summarization usually suffers from the problem that the contribution of the visual modality is unclear. Existing multimodal summarization approaches focus on designing the fusion methods of different modalities, while ignoring the adaptive conditions under which visual modalities are useful. Therefore, we propose a novel Coarse-to-Fine contribution network for multimodal Summarization (CFSum) to consider different contributions of images for summarization. First, to eliminate the interference of useless images, we propose a pre-filter module to abandon useless images. Second, to make accurate use of useful images, we propose two levels of visual complement modules, word level and phrase level. Specifically, image contributions are calculated and are adopted to guide the attention of both textual and visual modalities. Experimental results have shown that CFSum significantly outperforms multiple strong baselines on the standard benchmark. Furthermore, the analysis verifies that useful images can even help generate non-visual words which are implicitly represented in the image.

Incorporating Distributions of Discourse Structure for Long Document Abstractive Summarization
Dongqi Pu, Yifan Wang and Vera Demberg   11:00-12:30 (Pier 7&8)
For text summarization, the role of discourse structure is pivotal in discerning the core content of a text. Regrettably, prior studies on incorporating Rhetorical Structure Theory (RST) into transformer-based summarization models only consider the nuclearity annotation, thereby overlooking the variety of discourse relation types. This paper introduces the 'RSTformer', a novel summarization model that comprehensively incorporates both the types and uncertainty of rhetorical relations. Our RST-attention mechanism, rooted in document-level rhetorical structure, is an extension of the recently devised Longformer framework. Through rigorous evaluation, the model proposed herein exhibits significant superiority over state-of-the-art models, as evidenced by its notable performance on several automatic metrics and human evaluation.

Improving Radiology Summarization with Radiograph and Anatomy Prompts
jinpeng hu, Zhihong Chen, Yang Liu, Xiang Wan and Tsung-Hui Chang   11:00-12:30 (Pier 7&8)
The impression is crucial for the referring physicians to grasp key information since it is concluded from the findings and reasoning of radiologists. To alleviate the workload of radiologists and reduce repetitive human labor in impression writing, many researchers have focused on automatic impression generation. However, recent works on this task mainly summarize the corresponding findings and pay less attention to the radiology images. In clinical, radiographs can provide more detailed valuable observations to enhance radiologists' impression writing, especially for complicated cases. Besides, each sentence in findings usually focuses on single anatomy, such that they only need to be matched to corresponding anatomical regions instead of the whole image, which is beneficial for textual and visual features alignment. Therefore, we propose a novel anatomy-enhanced multimodal model to promote impression generation. In detail, we first construct a set of rules to extract anatomies and put these prompts into each sentence to highlight anatomy characteristics. Then, two separate encoders are applied to extract features from the radiograph and findings. Afterward, we utilize a contrastive learning module to align these two representations at the overall level and use a co-attention to fuse them at the sentence level with the help of anatomy-enhanced sentence representation. The experimental results on two benchmark datasets confirm the effectiveness of the proposed method, which achieves state-of-the-art results.

Towards Argument-Aware Abstractive Summarization of Long Legal Opinions with Summary Reranking
Mohamed Elaraby, Yang Zhong and Diane Litman   11:00-12:30 (Pier 7&8)
We propose a simple approach for the abstractive summarization of long legal opinions that takes into account the argument structure of the document. Legal opinions often contain complex and nuanced argumentation, making it challenging to generate a concise summary that accurately captures the main points of the legal opinion. Our approach involves using argument role information to generate multiple candidate summaries, then reranking these candidates based on alignment with the document's argument structure. We demonstrate the effectiveness of our approach on a dataset of long legal opinions and show that it outperforms several strong baselines.

Multi-Dimensional Evaluation of Text Summarization with In-Context Learning
Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig and Chunting Zhou 11:00-12:30 (Pier 7&8)
Evaluation of natural language generation (NLG) is complex and multi-dimensional. Generated text can be evaluated for fluency, coherence, factuality, or any other dimensions of interest. Most frameworks that perform such multi-dimensional evaluation require training on large manually or synthetically generated datasets. In this paper, we study the efficacy of large language models as multi-dimensional evaluators using in-context learning, obviating the need for large training datasets. Our experiments show that in-context learning-based evaluators are competitive with learned evaluation frameworks for the task of text summarization, establishing state-of-the-art on dimensions such as relevance and factual consistency. We then analyze the effects of factors such as the selection and number of in-context examples on performance. Finally, we study the efficacy of in-context learning-based evaluators in evaluating zero-shot summaries written by large language models such as GPT-3.

Improving Long Dialogue Summarization with Semantic Graph Representation
Yilun Hua, Zhaoyuan Deng and Kathleen McKeown   11:00-12:30 (Pier 7&8)
Although Large Language Models (LLMs) are successful in abstractive summarization of short dialogues, summarization of long dialogues remains challenging. To address this challenge, we propose a novel algorithm that processes complete dialogues comprising thousands of tokens into topic-segment-level Abstract Meaning Representation (AMR) graphs, which explicitly capture the dialogue structure, highlight salient semantics, and preserve high-level information. We also develop a new text-graph attention to leverage both graph semantics and a pretrained LLM that exploits the text. Finally, we propose an AMR node selection loss used jointly with conventional cross-entropy loss, to create additional training signals that facilitate graph feature encoding and content selection. Experiments show that our system outperforms the state-of-the-art models on multiple long dialogue summarization datasets, especially in low-resource settings, and generalizes well to out-of-domain data.

Aspect-aware Unsupervised Extractive Opinion Summarization
Haoyuan Li, Somnath Basu Roy Chowdhury and Snigdha Chaturvedi   11:00-12:30 (Pier 7&8)
Extractive opinion summarization extracts sentences from users' reviews to represent the prevalent opinions about a product or service. However, the extracted sentences can be redundant and may miss some important aspects, especially for centroid-based extractive summarization models (Radev et al., 2004). To alleviate these issues, we introduce TokenCluster– a method for unsupervised extractive opinion summarization that automatically identifies the aspects described in the review sentences and then extracts sentences based on their aspects. It identifies the underlying aspects of the review sentences using roots of noun phrases and adjectives appearing in them. Empirical evaluation shows that

TokenCluster improves aspect coverage in summaries and achieves strong performance on multiple opinion summarization datasets, for both general and aspect-specific summarization. We also perform extensive ablation and human evaluation studies to validate the design choices of our method. The implementation of our work is available at https://github.com/leehaoyuan/TokenCluster

Revisiting Cross-Lingual Summarization: A Corpus-based Study and A New Benchmark with Improved Annotation
Yulong Chen, Huajian Zhang, Yijie Zhou, Xuefeng Bai, Yueguan Wang, Ming Zhong, Jianhao Yan, Yafu Li, Judy Li, Xianchao Zhu and Yue Zhang 11:00-12:30 (Pier 7&8)
Most existing cross-lingual summarization (CLS) work constructs CLS corpora by simply and directly translating pre-annotated summaries from one language to another, which can contain errors from both summarization and translation processes. To address this issue, we propose ConvSumX, a cross-lingual conversation summarization benchmark, through a new annotation schema that explicitly considers source input context. ConvSumX consists of 2 sub-tasks under different real-world scenarios, with each covering 3 language directions. We conduct thorough analysis on ConvSumX and 3 widely-used manually annotated CLS corpora and empirically find that ConvSumX is more faithful towards input text. Additionally, based on the same intuition, we propose a 2-Step method, which takes both conversation and summary as input to simulate human annotation process. Experimental results show that 2-Step method surpasses strong baselines on ConvSumX under both automatic and human evaluation. Analysis shows that both source input text and summary are crucial for modeling cross-lingual summaries.

Dialogue Summarization with Static-Dynamic Structure Fusion Graph
Shen Gao, Xin Cheng, Mingzhe Li, Xiuying Chen, Jinpeng Li, Dongyan Zhao and Rui Yan 11:00-12:30 (Pier 7&8)
Dialogue, the most fundamental and specially privileged arena of language, gains increasing ubiquity across the Web in recent years. Quickly going through the long dialogue context and capturing salient information scattered over the whole dialogue session benefit users in many real-world Web applications such as email thread summarization and meeting minutes draft. Dialogue summarization is a challenging task in that dialogue has dynamic interaction nature and presumably inconsistent information flow among various speakers. Many researchers address this task by modeling dialogue with pre-computed static graph structure using external linguistic toolkits. However, such methods heavily depend on the reliability of external tools and the static graph construction is disjoint with the graph representation learning phase, which makes the graph can't be dynamically adapted for the downstream summarization task. In this paper, we propose a Static-Dynamic graph-based Dialogue Summarization model (SDDS), which fuses prior knowledge from human expertise and adaptively learns the graph structure in an end-to-end learning fashion. To verify the effectiveness of SDDS, we conduct experiments on three benchmark datasets (SAM-Sum, MediaSum, and DialogSum) and the results verify the superiority of SDDS.

Detecting and Mitigating Hallucinations in Machine Translation: Model Internal Workings Alone Do Well, Sentence Similarity Even Better
David Dale, Elena Voita, Loic Barrault and Marta R. Costa-jussà 11:00-12:30 (Pier 7&8)
While the problem of hallucinations in neural machine translation has long been recognized, so far the progress on its alleviation is very little. Indeed, recently it turned out that without artificially encouraging models to hallucinate, previously existing methods fall short and even the standard sequence log-probability is more informative. It means that internal characteristics of the model can give much more information than we expect, and before using external models and measures, we first need to ask: how far can we go if we use nothing but the translation model itself ? We propose to use a method that evaluates the percentage of the source contribution to a generated translation. Intuitively, hallucinations are translations "detached" from the source, hence they can be identified by low source contribution. This method improves detection accuracy for the most severe hallucinations by a factor of 2 and is able to alleviate hallucinations at test time on par with the previous best approach that relies on external models. Next, if we move away from internal model characteristics and allow external tools, we show that using sentence similarity from cross-lingual embeddings further improves these results. We release the code of our experiments.

A Formal Perspective on Byte-Pair Encoding
Vilém Zouhar, Clara Meister, Juan Luis Gastaldi, Li Du, Tim Vieira, Mrinmaya Sachan and Ryan Cotterell 11:00-12:30 (Pier 7&8)
Byte-Pair Encoding (BPE) is a popular algorithm used for tokenizing data in NLP, despite being devised initially as a compression method. BPE appears to be a greedy algorithm at face value, but the underlying optimization problem that BPE seeks to solve has not yet been laid down. We formalize BPE as a combinatorial optimization problem. Via submodular functions, we prove that the iterative greedy version is a $1/\sigma*(1-e^{(-\sigma)})$-approximation of an optimal merge sequence, where sigma is the total backward curvature with respect to the optimal merge sequence. Empirically the lower bound of the approximation is approx 0.37.
We provide a faster implementation of BPE which improves the runtime complexity from $O(NM)$ to $O(N \log M)$, where N is the sequence length and M is the merge count. Finally, we optimize the brute-force algorithm for optimal BPE using memoization.

Revisiting Commonsense Reasoning in Machine Translation: Training, Evaluation and Challenge
Xuebo Liu, Yutong Wang, Derek F. Wong, Runzhe Zhan, Liangxuan Yu and Min Zhang 11:00-12:30 (Pier 7&8)
The ability of commonsense reasoning (CR) decides whether a neural machine translation (NMT) model can move beyond pattern recognition. Despite the rapid advancement of NMT and the use of pretraining to enhance NMT models, research on CR in NMT is still in its infancy, leaving much to be explored in terms of effectively training NMT models with high CR abilities and devising accurate automatic evaluation metrics. This paper presents a comprehensive study aimed at expanding the understanding of CR in NMT. For the training, we confirm the effectiveness of incorporating pretrained knowledge into NMT models and subsequently utilizing these as robust testbeds for investigating CR in NMT. For the evaluation, we propose a novel entity-aware evaluation method that takes into account both the NMT candidate and important entities in the candidate, which is more aligned with human judgement. Based on the strong testbed and evaluation methods, we identify challenges in training NMT models with high CR abilities and suggest directions for further unlabeled data utilization and model design. We hope that our methods and findings will contribute to advancing the research of CR in NMT. Source data, code and scripts are freely available at https://github.com/YutongWang1216/CR-NMT.

What Knowledge Is Needed? Towards Explainable Memory for kNN-MT Domain Adaptation
Wenhao Zhu, Shujian Huang, Yunzhe Lv, Xin Zheng and Jiajun CHEN 11:00-12:30 (Pier 7&8)
kNN-MT presents a new paradigm for domain adaptation by building an external datastore, which usually saves all target language token occurrences in the parallel corpus. As a result, the constructed datastore is usually large and possibly redundant. In this paper, we investigate the interpretability issue of this approach: what knowledge does the NMT model need? We propose the notion of local correctness (LAC) as a new angle, which describes the potential translation correctness for a single entry and for a given neighborhood. Empirical study shows that our investigation successfully finds the conditions where the NMT model could easily fail and need related knowledge. Experiments on six diverse target domains and two language-pairs show that pruning according to local correctness brings a light and more explainable memory for kNN-MT domain adaptation.

Token-Level Self-Evolution Training for Sequence-to-Sequence Learning
Keqin Peng, Liang Ding, Qihuang Zhong, Yuanxin Ouyang, Wenge Rong, Zhang Xiong and Dacheng Tao 11:00-12:30 (Pier 7&8)
Adaptive training approaches, widely used in sequence-to-sequence models, commonly reweigh the losses of different target tokens based

on priors, e.g. word frequency. However, most of them do not consider the variation of learning difficulty in different training steps, and overly emphasize the learning of difficult one-hot labels, making the learning deterministic and sub-optimal. In response, we present Token-Level Self-Evolution Training (SE), a simple and effective dynamic training method to fully and wisely exploit the knowledge from data. SE focuses on dynamically learning the under-explored tokens for each forward pass and adaptively regularizes the training by introducing a novel token-specific label smoothing approach. Empirically, SE yields consistent and significant improvements in three tasks, i.e. machine translation, summarization, and grammatical error correction. Encouragingly, we achieve averaging +0.93 BLEU improvement on three machine translation tasks. Analyses confirm that, besides improving lexical accuracy, SE enhances generation diversity and model generalization.

PEIT: Bridging the Modality Gap with Pre-trained Models for End-to-End Image Translation
Shaolin Zhu, Shangjie Li, Yikun Lei and Deyi Xiong   11:00-12:30 (Pier 7&8)
Image translation is a task that translates an image containing text in the source language to the target language. One major challenge with image translation is the modality gap between visual text inputs and textual inputs/outputs of machine translation (MT). In this paper, we propose PEIT, an end-to-end image translation framework that bridges the modality gap with pre-trained models. It is composed of four essential components: a visual encoder, a shared encoder-decoder backbone network, a vision-text representation aligner equipped with the shared encoder and a cross-modal regularizer stacked over the shared decoder. Both the aligner and regularizer aim at reducing the modality gap. To train PEIT, we employ a two-stage pre-training strategy with an auxiliary MT task: (1) pre-training the MT model on the MT training data to initialize the shared encoder-decoder backbone network; and (2) pre-training PEIT with the aligner and regularizer on a synthesized dataset with rendered images containing text from the MT training data. In order to facilitate the evaluation of PEIT and promote research on image translation, we create a large-scale image translation corpus ECOIT containing 480K image-translation pairs via crowd-sourcing and manual post-editing from real-world images in the e-commerce domain. Experiments on the curated ECOIT benchmark dataset demonstrate that PEIT substantially outperforms both cascaded image translation systems (OCR+MT) and previous strong end-to-end image translation model, with fewer parameters and faster decoding speed.

MTCue: Learning Zero-Shot Control of Extra-Textual Attributes by Leveraging Unstructured Context in Neural Machine Translation
Sebastian T. Vincent, Robert James Flynn and Carolina Scarton   11:00-12:30 (Pier 7&8)
Efficient utilisation of both intra- and extra-textual context remains one of the critical gaps between machine and human translation. Existing research has primarily focused on providing individual, well-defined types of context in translation, such as the surrounding text or discrete external variables like the speaker's gender. This work introduces MTCue, a novel neural machine translation (NMT) framework that interprets all context (including discrete variables) as text. MTCue learns an abstract representation of context, enabling transferability across different data settings and leveraging similar attributes in low-resource scenarios. With a focus on a dialogue domain with access to document and metadata context, we extensively evaluate MTCue in four language pairs in both translation directions. Our framework demonstrates significant improvements in translation quality over a parameter-matched non-contextual baseline, as measured by BLEU (+0.88) and Comet (+1.58). Moreover, MTCue significantly outperforms a "tagging" baseline at translating English text. Analysis reveals that the context encoder of MTCue learns a representation space that organises context based on specific attributes, such as formality, enabling effective zero-shot control. Pre-training on context embeddings also improves MTCue's few-shot performance compared to the "tagging" baseline. Finally, an ablation study conducted on model components and contextual variables further supports the robustness of MTCue for context-based NMT.

Towards Accurate Translation via Semantically Appropriate Application of Lexical Constraints
Yujin Baek, Koanho Lee, Dayeon Ki, Cheonbok Park, Hyoung-Gyu Lee and Jaegul Choo   11:00-12:30 (Pier 7&8)
Lexically-constrained NMT (LNMT) aims to incorporate user-provided terminology into translations. Despite its practical advantages, existing work has not evaluated LNMT models under challenging real-world conditions. In this paper, we focus on two important but understudied issues that lie in the current evaluation process of LNMT studies. The model needs to cope with challenging lexical constraints that are "homographs" or "unseen" during training. To this end, we first design a homograph disambiguation module to differentiate the meanings of homographs. Moreover, we propose PLUMCOT which integrates contextually rich information about unseen lexical constraints from pre-trained language models and strengthens a copy mechanism of the pointer network via direct supervision of a copying score. We also release HOLLY, an evaluation benchmark for assessing the ability of model to cope with "homographic" and "unseen" lexical constraints. Experiments on HOLLY and the previous test setup show the effectiveness of our method. The effects of PLUMCOT are shown to be remarkable in "unseen" constraints.Our dataset is available at https://github.com/papago-lab/HOLLY-benchmark.

Pretrained Bidirectional Distillation for Machine Translation
Yimeng Zhuang and Mei Tu   11:00-12:30 (Pier 7&8)
Knowledge transfer can boost neural machine translation (NMT), for example, by finetuning a pretrained masked language model (LM). However, it may suffer from the forgetting problem and the structural inconsistency between pretrained LMs and NMT models. Knowledge distillation (KD) may be a potential solution to alleviate these issues, but few studies have investigated language knowledge transfer from pretrained language models to NMT models through KD. In this paper, we propose Pretrained Bidirectional Distillation (PBD) for NMT, which aims to efficiently transfer bidirectional language knowledge from masked language pretraining to NMT models. Its advantages are reflected in efficiency and effectiveness through a globally defined and bidirectional context-aware distillation objective. Bidirectional language knowledge of the entire sequence is transferred to an NMT model concurrently during translation training. Specifically, we propose self-distilled masked language pretraining to obtain the PBD objective. We also design PBD losses to efficiently distill the language knowledge, in the form of token probabilities, to the encoder and decoder of an NMT model using the PBD objective. Extensive experiments reveal that pretrained bidirectional distillation can significantly improve machine translation performance and achieve competitive or even better results than previous pretrain-finetune or unified multilingual translation methods in supervised, unsupervised, and zero-shot scenarios. Empirically, it is concluded that pretrained bidirectional distillation is an effective and efficient method for transferring language knowledge from pretrained language models to NMT models.

Encoder and Decoder, Not One Less for Pre-trained Language Model Sponsored NMT
Sufeng Duan and Hai Zhao   11:00-12:30 (Pier 7&8)
Well pre-trained contextualized representations from pre-trained language model (PLM) have been shown helpful for enhancing various natural language processing tasks, surely including neural machine translation (NMT). However, existing methods either consider encoder-only enhancement or rely on specific multilingual PLMs, which leads to a much larger model or give up potentially helpful knowledge from target PLMs. In this paper, we propose a new monolingual PLM-sponsored NMT model to let both encoder and decoder enjoy PLM enhancement to alleviate such obvious inconvenience. Especially, incorporating a newly proposed frequency-weighted embedding transformation algorithm, PLM embeddings can be effectively exploited in terms of the representations of the NMT decoder. We evaluate our model on IWSLT14 En-De, De-En, WMT14 En-De, and En-Fr tasks, and the results show that our proposed PLM enhancement gives significant improvement and even helps achieve new state-of-the-art.

How effective is machine translation on low-resource code-switching? A case study comparing human and automatic metrics
Li Nguyen, Christopher Bryant, Oliver Mayeux and Zheng Yuan   11:00-12:30 (Pier 7&8)

This paper presents an investigation into the differences between processing monolingual input and code-switching (CSW) input in the context of machine translation (MT). Specifically, we compare the performance of three MT systems (Google, mBART-50 and M2M-100-big) in terms of their ability to translate monolingual Vietnamese, a low-resource language, and Vietnamese-English CSW respectively. To our knowledge, this is the first study to systematically analyse what might happen when multilingual MT systems are exposed to CSW data using both automatic and human metrics. We find that state-of-the-art neural translation systems not only achieve higher scores on automatic metrics when processing CSW input (compared to monolingual input), but also produce translations that are consistently rated as more semantically faithful by humans. We further suggest that automatic evaluation alone is insufficient for evaluating the translation of CSW input. Our findings establish a new benchmark that offers insights into the relationship between MT and CSW.

TranSFormer: Slow-Fast Transformer for Machine Translation
Bei Li, Yi Jing, Xu Tan, Zhen Xing, Tong Xiao and Jingbo Zhu   11:00-12:30 (Pier 7&8)
Learning multiscale Transformer models has been evidenced as a viable approach to augmenting machine translation systems. Prior research has primarily focused on treating subwords as basic units in developing such models. However, the incorporation of fine-grained character-level features into multiscale Transformer has not yet been explored. In this work, we present a **S**low-**F**ast two-stream learning model, referred to as Tran**SF**ormer, which utilizes a "slow" branch to deal with subword sequences and a "fast" branch to deal with longer character sequences. This model is efficient since the fast branch is very lightweight by reducing the model width, and yet provides useful fine-grained features for the slow branch. Our TranSFormer shows consistent BLEU improvements (larger than 1 BLEU point) on several machine translation benchmarks.

Neural Machine Translation Methods for Translating Text to Sign Language Glosses
Dele Zhu, Vera Czehmann and Eleftherios Avramidis   11:00-12:30 (Pier 7&8)
State-of-the-art techniques common to low resource Machine Translation (MT) are applied to improve MT of spoken language text to Sign Language (SL) glosses. In our experiments, we improve the performance of the transformer-based models via (1) data augmentation, (2) semi-supervised Neural Machine Translation (NMT), (3) transfer learning and (4) multilingual NMT. The proposed methods are implemented progressively on two German SL corpora containing gloss annotations. Multilingual NMT combined with data augmentation appear to be the most successful setting, yielding statistically significant improvements as measured by three automatic metrics (up to over 6 points BLEU), and confirmed via human evaluation. Our best setting outperforms all previous work that report on the same test-set and is also confirmed on a corpus of the American Sign Language (ASL).

Dual-Gated Fusion with Prefix-Tuning for Multi-Modal Relation Extraction
Qian Li, Shu Guo, Cheng Ji, Xutan Peng, Shiyao Cui, Jianxin Li and Lihong Wang   11:00-12:30 (Pier 7&8)
Multi-Modal Relation Extraction (MMRE) aims at identifying the relation between two entities in texts that contain visual clues. Rich visual content is valuable for the MMRE task, but existing works cannot well model finer associations among different modalities, failing to capture the truly helpful visual information and thus limiting relation extraction performance. In this paper, we propose a novel MMRE framework to better capture the deeper correlations of text, entity pair, and image/objects, so as to mine more helpful information for the task, termed as DGF-PT. We first propose a prompt-based autoregressive encoder, which builds the associations of intra-modal and inter-modal features related to the task, respectively by entity-oriented and object-oriented prefixes. To better integrate helpful visual information, we design a dual-gated fusion module to distinguish the importance of image/objects and further enrich text representations. In addition, a generative decoder is introduced with entity type restriction on relations, better filtering out candidates. Extensive experiments conducted on the benchmark dataset show that our approach achieves excellent performance compared to strong competitors, even in the few-shot situation.

Text Augmented Open Knowledge Graph Completion via Pre-Trained Language Models
Pengcheng Jiang, Shivam Agarwal, Bowen Jin, Xuan Wang, Jimeng Sun and Jiawei Han   11:00-12:30 (Pier 7&8)
The mission of open knowledge graph (KG) completion is to draw new findings from known facts. Existing works that augment KG completion require either (1) factual triples to enlarge the graph reasoning space or (2) manually designed prompts to extract knowledge from a pre-trained language model (PLM), exhibiting limited performance and requiring expensive efforts from experts. To this end, we propose TagReal that automatically generates quality query prompts and retrieves support information from large text corpora to probe knowledge from PLM for KG completion. The results show that TagReal achieves state-of-the-art performance on two benchmark datasets. We find that TagReal has superb performance even with limited training data, outperforming existing embedding-based, graph-based, and PLM-based methods.

An Embarrassingly Easy but Strong Baseline for Nested Named Entity Recognition
Hang Yan, Yu Sun, Xiaonan Li and Xipeng Qiu   11:00-12:30 (Pier 7&8)
Named entity recognition (NER) is the task to detect and classify entity spans in the text. When entity spans overlap between each other, the task is named as nested NER. Span-based methods have been widely used to tackle nested NER. Most of these methods get a score matrix, where each entry corresponds to a span. However, previous work ignores spatial relations in the score matrix. In this paper, we propose using Convolutional Neural Network (CNN) to model these spatial relations. Despite being simple, experiments in three commonly used nested NER datasets show that our model surpasses several recently proposed methods with the same pre-trained encoders. Further analysis shows that using CNN can help the model find more nested entities. Besides, we find that different papers use different sentence tokenizations for the three nested NER datasets, which will influence the comparison. Thus, we release a pre-processing script to facilitate future comparison.

What Is Overlap Knowledge in Event Argument Extraction? APE: A Cross-datasets Transfer Learning Model for EAE
Kaihang Zhang, Kai Shuang, Xinyue Yang, Xuyang Yao and Jinyu Guo   11:00-12:30 (Pier 7&8)
The EAE task extracts a structured event record from an event text. Most existing approaches train the EAE model on each dataset independently and ignore the overlap knowledge across datasets. However, insufficient event records in a single dataset often prevent the existing model from achieving better performance. In this paper, we clearly define the overlap knowledge across datasets and split the knowledge of the EAE task into overlap knowledge across datasets and specific knowledge of the target dataset. We propose APE model to learn the two parts of knowledge in two serial learning phases without causing catastrophic forgetting. In addition, we formulate both learning phases as conditional generation tasks and design Stressing Entity Type Prompt to close the gap between the two phases. The experiments show APE achieves new state-of-the-art with a large margin in the EAE task. When only ten records are available in the target dataset, our model dramatically outperforms the baseline model with average 27.27% F1 gain.

Multi-hop Evidence Retrieval for Cross-document Relation Extraction
Keming Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma and Muhao Chen   11:00-12:30 (Pier 7&8)
Relation Extraction (RE) has been extended to cross-document scenarios because many relations are not simply described in a single document. This inevitably brings the challenge of efficient open-space evidence retrieval to support the inference of cross-document relations, along with the challenge of multi-hop reasoning on top of entities and evidence scattered in an open set of documents. To combat these challenges, we propose Mr.Cod (Multi-hop evidence retrieval for Cross-document relation extraction), which is a multi-hop evidence retrieval method based on evidence path mining and ranking. We explore multiple variants of retrievers to show evidence retrieval is essential in

cross-document RE. We also propose a contextual dense retriever for this setting. Experiments on CodRED show that evidence retrieval with Mr.Cod effectively acquires cross-document evidence and boosts end-to-end RE performance in both closed and open settings.

Graph Propagation based Data Augmentation for Named Entity Recognition
Jiong Cai, Shen Huang, Yong Jiang, Zeqi Tan, Pengjun Xie and Kewei Tu   11:00-12:30 (Pier 7&8)
Data augmentation is an effective solution to improve model performance and robustness for low-resource named entity recognition (NER). However, synthetic data often suffer from poor diversity, which leads to performance limitations. In this paper, we propose a novel Graph Propagated Data Augmentation (GPDA) framework for Named Entity Recognition (NER), leveraging graph propagation to build relationships between labeled data and unlabeled natural texts. By projecting the annotations from the labeled text to the unlabeled text, the unlabeled texts are partially labeled, which has more diversity rather than synthetic annotated data. To strengthen the propagation precision, a simple search engine built on Wikipedia is utilized to fetch related texts of labeled data and to propagate the entity labels to them in the light of the anchor links. Besides, we construct and perform experiments on a real-world low-resource dataset of the E-commerce domain, which will be publicly available to facilitate the low-resource NER research. Experimental results show that GPDA presents substantial improvements over previous data augmentation methods on multiple low-resource NER datasets.

Easy-to-Hard Learning for Information Extraction
Chang Gao, Wenxuan Zhang, Wai Lam and Lidong Bing   11:00-12:30 (Pier 7&8)
Information extraction (IE) systems aim to automatically extract structured information, such as named entities, relations between entities, and events, from unstructured texts. While most existing work addresses a particular IE task, universally modeling various IE tasks with one model has achieved great success recently. Despite their success, they employ a one-stage learning strategy, i.e., directly learning to extract the target structure given the input text, which contradicts the human learning process. In this paper, we propose a unified easy-to-hard learning framework consisting of three stages, i.e., the easy stage, the hard stage, and the main stage, for IE by mimicking the human learning process. By breaking down the learning process into multiple stages, our framework facilitates the model to acquire general IE task knowledge and improve its generalization ability. Extensive experiments across four IE tasks demonstrate the effectiveness of our framework. We achieve new state-of-the-art results on 13 out of 17 datasets.

Type Enhanced BERT for Correcting NER Errors
Kuai Li, Chen Chen, Tao Yang, Tianming Du, Peijie Yu, dong du and Feng Zhang   11:00-12:30 (Pier 7&8)
We introduce the task of correcting named entity recognition (NER) errors without re-training model. After an NER model is trained and deployed in production, it makes prediction errors, which usually need to be fixed quickly. To address this problem, we firstly construct a gazetteer containing named entities and corresponding possible entity types. And then, we propose type enhanced BERT (TyBERT), a method that integrates the named entity's type information into BERT by an adapter layer. When errors are identified, we can repair the model by updating the gazetteer. In other words, the gazetteer becomes a trigger to control NER model's output. The experiment results in multiple corpus show the effectiveness of our method, which outperforms strong baselines.x

Learning Latent Relations for Temporal Knowledge Graph Reasoning
Mengqi Zhang, Yuwei Xia, Qiang Liu, Shu Wu and Liang Wang   11:00-12:30 (Pier 7&8)
Temporal Knowledge Graph (TKG) reasoning aims to predict future facts based on historical data. However, due to the limitations in construction tools and data sources, many important associations between entities may be omitted in TKG. We refer to these missing associations as latent relations. Most existing methods have some drawbacks in explicitly capturing intra-time latent relations between co-occurring entities and inter-time latent relations between entities that appear at different times. To tackle these problems, we propose a new Latent Relations Learning method for TKG reasoning, namely L2TKG. Specifically, we first utilize a Structural Encoder (SE) to obtain representations of entities at each timestamp. We then design a Latent Relations Learning (LRL) module to mine and exploit the intra- and inter-time latent relations. Finally, we extract the temporal representations from the output of SE and LRL for entity prediction. Extensive experiments on four datasets demonstrate the effectiveness of L2TKG.

Silver Syntax Pre-training for Cross-Domain Relation Extraction
Elisa Bassignana, Filip Ginter, Sampo Pyysalo, Rob van der Goot and Barbara Plank   11:00-12:30 (Pier 7&8)
Relation Extraction (RE) remains a challenging task, especially when considering realistic out-of-domain evaluations. One of the main reasons for this is the limited training size of current RE datasets: obtaining high-quality (manually annotated) data is extremely expensive and cannot realistically be repeated for each new domain. An intermediate training step on data from related tasks has shown to be beneficial across many NLP tasks.However, this setup still requires supplementary annotated data, which is often not available. In this paper, we investigate intermediate pre-training specifically for RE. We exploit the affinity between syntactic structure and semantic RE, and identify the syntactic relations which are closely related to RE by being on the shortest dependency path between two entities. We then take advantage of the high accuracy of current syntactic parsers in order to automatically obtain large amounts of low-cost pre-training data. By pre-training our RE model on the relevant syntactic relations, we are able to outperform the baseline in five out of six cross-domain setups, without any additional annotated data.

The Art of Prompting: Event Detection based on Type Specific Prompts
Sijia Wang, Mo Yu and Lifu Huang   11:00-12:30 (Pier 7&8)
We compare various forms of prompts to represent event types and develop a unified framework to incorporate the event type specific prompts for supervised, few-shot, and zero-shot event detection. The experimental results demonstrate that a well-defined and comprehensive event type prompt can significantly improve event detection performance, especially when the annotated data is scarce (few-shot event detection) or not available (zero-shot event detection). By leveraging the semantics of event types, our unified framework shows up to 22.2% F-score gain over the previous state-of-the-art baselines.

Teamwork Is Not Always Good: An Empirical Study of Classifier Drift in Class-incremental Information Extraction
Minqian Liu and Lifu Huang   11:00-12:30 (Pier 7&8)
Class-incremental learning (CIL) aims to develop a learning system that can continually learn new classes from a data stream without forgetting previously learned classes. When learning classes incrementally, the classifier must be constantly updated to incorporate new classes, and the drift in decision boundary may lead to severe forgetting. This fundamental challenge, however, has not yet been studied extensively, especially in the setting where no samples from old classes are stored for rehearsal. In this paper, we take a closer look at how the drift in the classifier leads to forgetting, and accordingly, design four simple yet (super-) effective solutions to alleviate the classifier drift: an Individual Classifiers with Frozen Feature Extractor (ICE) framework where we individually train a classifier for each learning session, and its three variants ICE-PL, ICE-O, and ICE-PL&O which further take the logits of previously learned classes from old sessions or a constant logit of an Other class as constraint to the learning of new classifiers. Extensive experiments and analysis on 6 class-incremental information extraction tasks demonstrate that our solutions, especially ICE-O, consistently show significant improvement over the previous state-of-the-art approaches with up to 44.7% absolute F-score gain, providing a strong baseline and insights for future research on class-incremental learning.

Learning with Partial Annotations for Event Detection

Jian Liu, Dianbo Sui, Kang Liu, Haoyan Liu and Zhe Zhao   11:00-12:30 (Pier 7&8)

Event detection (ED) seeks to discover and classify event instances in plain texts. Previous methods for ED typically adopt supervised learning, requiring fully labeled and high-quality training data. However, in a real-world application, we may not obtain clean training data but only partially labeled one, which could substantially impede the learning process. In this work, we conduct a seminal study for learning with partial annotations for ED. We propose a new trigger localization formulation using contrastive learning to distinguish ground-truth triggers from contexts, showing a decent robustness for addressing partial annotation noise. Impressively, in an extreme scenario where more than 90% of events are unlabeled, our approach achieves an F1 score of over 60%. In addition, we re-annotate and make available two fully annotated subsets of ACE 2005 to serve as an unbiased benchmark for event detection. We hope our approach and data will inspire future studies on this vital yet understudied problem.

DiSCoMaT: Distantly Supervised Composition Extraction from Tables in Materials Science Articles

Tanishq Gupta, Mohd Zaki, Devanshi Khatsuriya, Kausik Hira, N M Anoop Krishnan and Mausam -   11:00-12:30 (Pier 7&8)

A crucial component in the curation of KB for a scientific domain (e.g., materials science, food & nutrition, fuels) is information extraction from tables in the domain's published research articles. To facilitate research in this direction, we define a novel NLP task of extracting compositions of materials (e.g., glasses) from tables in materials science papers. The task involves solving several challenges in concert, such as tables that mention compositions have highly varying structures; text in captions and full paper needs to be incorporated along with data in tables; and regular languages for numbers, chemical compounds, and composition expressions must be integrated into the model. We release a training dataset comprising 4,408 distantly supervised tables, along with 1,475 manually annotated dev and test tables. We also present DiSCoMaT, a strong baseline that combines multiple graph neural networks with several task-specific regular expressions, features, and constraints. We show that DiSCoMaT outperforms recent table processing architectures by significant margins. We release our code and data for further research on this challenging IE task from scientific tables.

CHEER: Centrality-aware High-order Event Reasoning Network for Document-level Event Causality Identification

Meiqi Chen, Yixin Cao, Yan Zhang and Zhiwei Liu   11:00-12:30 (Pier 7&8)

Document-level Event Causality Identification (DECI) aims to recognize causal relations between events within a document. Recent studies focus on building a document-level graph for cross-sentence reasoning, but ignore important causal structures — there are one or two "central" events that prevail throughout the document, with most other events serving as either their cause or consequence. In this paper, we manually annotate central events for a systematical investigation and propose a novel DECI model, CHEER, which performs high-order reasoning while considering event centrality. First, we summarize a general GNN-based DECI model and provide a unified view for better understanding. Second, we design an Event Interaction Graph (EIG) involving the interactions among events (e.g., coreference) and event pairs, e.g., causal transitivity, cause(A, B) AND cause(B, C) → cause(A, C). Finally, we incorporate event centrality information into the EIG reasoning network via well-designed features and multi-task learning. We have conducted extensive experiments on two benchmark datasets. The results present great improvements (5.9% F1 gains on average) and demonstrate the effectiveness of each main component.

OD-RTE: A One-Stage Object Detection Framework for Relational Triple Extraction

Jinzhong Ning, Zhihao Yang, Yuanyuan Sun, Zhizheng Wang and Hongfei LIN   11:00-12:30 (Pier 7&8)

The Relational Triple Extraction (RTE) task is a fundamental and essential information extraction task. Recently, the table-filling RTE methods have received lots of attention. Despite their success, they suffer from some inherent problems such as underutilizing regional information of triple. In this work, we treat the RTE task based on table-filling method as an Object Detection task and propose a one-stage Object Detection framework for Relational Triple Extraction (OD-RTE). In this framework, the vertices-based bounding box detection, coupled with auxiliary global relational triple region detection, ensuring that regional information of triple could be fully utilized. Besides, our proposed decoding scheme could extract all types of triples. In addition, the negative sampling strategy of relations in the training stage improves the training efficiency while alleviating the imbalance of positive and negative relations. The experimental results show that 1) OD-RTE achieves the state-of-the-art performance on two widely used datasets (i.e., NYT and WebNLG). 2) Compared with the best performing table-filling method, OD-RTE achieves faster training and inference speed with lower GPU memory usage. To facilitate future research in this area, the codes are publicly available at https://github.com/NingJinzhong/ODRTE.

Semantic Structure Enhanced Event Causality Identification

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Saiping Guan, Jiafeng Guo and Xueqi Cheng   11:00-12:30 (Pier 7&8)

Event Causality Identification (ECI) aims to identify causal relations between events in unstructured texts. This is a very challenging task, because causal relations are usually expressed by implicit associations between events. Existing methods usually capture such associations by directly modeling the texts with pre-trained language models, which underestimate two kinds of semantic structures vital to the ECI task, namely, event-centric structure and event-associated structure. The former includes important semantic elements related to the events to describe them more precisely, while the latter contains semantic paths between two events to provide possible supports for ECI. In this paper, we study the implicit associations between events by modeling the above explicit semantic structures, and propose a Semantic Structure Integration model (SemSIn). It utilizes a GNN-based event aggregator to integrate the event-centric structure information, and employs an LSTM-based path aggregator to capture the event-associated structure information between two events. Experimental results on three widely used datasets show that SemSIn achieves significant improvements over baseline methods.

Constrained Tuple Extraction with Interaction-Aware Network

Xiaojun Xue, Chunxia Zhang, Tianxiang Xu and Zhendong Niu   11:00-12:30 (Pier 7&8)

Tuples extraction is a fundamental task for information extraction and knowledge graph construction. The extracted tuples are usually represented as knowledge triples consisting of subject, relation, and object. In practice, however, the validity of knowledge triples is associated with and changes with the spatial, temporal, or other kinds of constraints. Motivated by this observation, this paper proposes a constrained tuple extraction (CTE) task to guarantee the validity of knowledge tuples. Formally, the CTE task is to extract constrained tuples from unstructured text, which adds constraints to conventional triples. To this end, we propose an interaction-aware network. Combinatorial interactions among context-specific external features and distinct-granularity internal features are exploited to effectively mine the potential constraints. Moreover, we have built a new dataset containing totally 1,748,826 constrained tuples for training and 3656 ones for evaluation. Experiments on our dataset and the public CaRB dataset demonstrate the superiority of the proposed model. The constructed dataset and the codes are publicly available.

DSP: Discriminative Soft Prompts for Zero-Shot Entity and Relation Extraction

Bo Lv, Xin Liu, Shaojie Dai, Nayu Liu, Fan Yang, Ping Luo and Yue Yu   11:00-12:30 (Pier 7&8)

Prompt-based methods have shown their efficacy in transferring general knowledge within pre-trained language models (PLMs) for low-resource scenarios. Typically, prompt-based methods convert downstream tasks to cloze-style problems and map all labels to verbalizers. However, when applied to zero-shot entity and relation extraction, vanilla prompt-based methods may struggle with the limited coverage of verbalizers to labels and the slow inference speed. In this work, we propose a novel Discriminate Soft Prompts (DSP) approach to take advantage of the prompt-based methods to strengthen the transmission of general knowledge. Specifically, we develop a discriminative

prompt method, which reformulates zero-shot tasks into token discrimination tasks without having to construct verbalizers. Furthermore, to improve the inference speed of the prompt-based methods, we design a soft prompt co-reference strategy, which leverages soft prompts to approximately refer to the vector representation of text tokens. The experimental results show that, our model outperforms baselines on two zero-shot entity recognition datasets with higher inference speed, and obtains a 7.5% average relation F1-score improvement over previous state-of-the-art models on Wiki-ZSL and FewRel.

Adaptive Ordered Information Extraction with Deep Reinforcement Learning
Wenhao Huang, Jiaqing Liang, Zhixu Li, Yanghua Xiao and chuanjun ji 11:00-12:30 (Pier 7&8)
Information extraction (IE) has been studied extensively. The existing methods always follow a fixed extraction order for complex IE tasks with multiple elements to be extracted in one instance such as event extraction. However, we conduct experiments on several complex IE datasets and observe that different extraction orders can significantly affect the extraction results for a great portion of instances, and the ratio of sentences that are sensitive to extraction orders increases dramatically with the complexity of the IE task. Therefore, this paper proposes a novel adaptive ordered IE paradigm to find the optimal element extraction order for different instances, so as to achieve the best extraction results. We also propose an reinforcement learning (RL) based framework to generate optimal extraction order for each instance dynamically. Additionally, we propose a co-training framework adapted to RL to mitigate the exposure bias during the extractor training phase. Extensive experiments conducted on several public datasets demonstrate that our proposed method can beat previous methods and effectively improve the performance of various IE tasks, especially for complex ones.

Learning from a Friend: Improving Event Extraction via Self-Training with Feedback from Abstract Meaning Representation
Zhiyang Xu, Jay Yoon Lee and Lifu Huang 11:00-12:30 (Pier 7&8)
Data scarcity has been the main factor that hinders the progress of event extraction. To overcome this issue, we propose a Self-Training with Feedback (STF) framework that leverages the large-scale unlabeled data and acquires feedback for each new event prediction from the unlabeled data by comparing it to the Abstract Meaning Representation (AMR) graph of the same sentence. Specifically, STF consists of (1) a base event extraction model trained on existing event annotations and then applied to large-scale unlabeled corpora to predict new event mentions as pseudo training samples, and (2) a novel scoring model that takes in each new predicted event trigger, an argument, its argument role, as well as their paths in the AMR graph to estimate a compatibility score indicating the correctness of the pseudo label. The compatibility scores further act as feedback to encourage or discourage the model learning on the pseudo labels during self-training. Experimental results on three benchmark datasets, including ACE05-E, ACE05-E+, and ERE, demonstrate the effectiveness of the STF framework on event extraction, especially event argument extraction, with significant performance gain over the base event extraction models and strong baselines. Our experimental analysis further shows that STF is a generic framework and it can be applied to improve most, if not all, event extraction models by leveraging large-scale unlabeled data, even when high-quality AMR graph annotations are not available.

Retrieve-and-Sample: Document-level Event Argument Extraction via Hybrid Retrieval Augmentation
Yubing Ren, Yanan Cao, Ping Guo, Fang Fang, Wei Ma and Zheng Lin 11:00-12:30 (Pier 7&8)
Recent studies have shown the effectiveness of retrieval augmentation in many generative NLP tasks. These retrieval-augmented methods allow models to explicitly acquire prior external knowledge in a non-parametric manner and regard the retrieved reference instances as cues to augment text generation. These methods use similarity-based retrieval, which is based on a simple hypothesis: the more the retrieved demonstration resembles the original input, the more likely the demonstration label resembles the input label. However, due to the complexity of event labels and sparsity of event arguments, this hypothesis does not always hold in document-level EAE. This raises an interesting question: How do we design the retrieval strategy for document-level EAE? We investigate various retrieval settings from the input and label distribution views in this paper. We further augment document-level EAE with pseudo demonstrations sampled from event semantic regions that can cover adequate alternatives in the same context and event schema. Through extensive experiments on RAMS and WikiEvents, we demonstrate the validity of our newly introduced retrieval-augmented methods and analyze why they work.

A Novel Table-to-Graph Generation Approach for Document-Level Joint Entity and Relation Extraction
Ruoyu Zhang, Yanzeng Li and Lei Zou 11:00-12:30 (Pier 7&8)
Document-level relation extraction (DocRE) aims to extract relations among entities within a document, which is crucial for applications like knowledge graph construction. Existing methods usually assume that entities and their mentions are identified beforehand, which falls short of real-world applications. To overcome this limitation, we propose TaG, a novel table-to-graph generation model for joint extraction of entities and relations at document-level. To enhance the learning of task dependencies, TaG induces a latent graph among mentions, with different types of edges indicating different task information, which is further broadcast with a relational graph convolutional network. To alleviate the error propagation problem, we adapt the hierarchical agglomerative clustering algorithm to back-propagate task information at decoding stage. Experiments on the benchmark dataset, DocRED, demonstrate that TaG surpasses previous methods by a large margin and achieves state-of-the-art results.

Aerial Vision-and-Dialog Navigation
Yue Fan, Winson Chen, Tongzhou Jiang, Chun Zhou, Yi Zhang and Xin Eric Wang 11:00-12:30 (Pier 7&8)
The ability to converse with humans and follow natural language commands is crucial for intelligent unmanned aerial vehicles (a.k.a. drones). It can relieve people's burden of holding a controller all the time, allow multitasking, and make drone control more accessible for people with disabilities or with their hands occupied. To this end, we introduce Aerial Vision-and-Dialog Navigation (AVDN), to navigate a drone via natural language conversation. We build a drone simulator with a continuous photorealistic environment and collect a new AVDN dataset of over 3k recorded navigation trajectories with asynchronous human-human dialogs between commanders and followers. The commander provides initial navigation instruction and further guidance by request, while the follower navigates the drone in the simulator and asks questions when needed. During data collection, followers' attention on the drone's visual observation is also recorded. Based on the AVDN dataset, we study the tasks of aerial navigation from (full) dialog history and propose an effective Human Attention Aided Transformer model (HAA-Transformer), which learns to predict both navigation waypoints and human attention.

MultiCapCLIP: Auto-Encoding Prompts for Zero-Shot Multilingual Visual Captioning
Bang Yang, Fenglin Liu, Xian Wu, Yaowei Wang, Xu Sun and Yuexian Zou 11:00-12:30 (Pier 7&8)
Supervised visual captioning models typically require a large scale of images or videos paired with descriptions in a specific language (i.e., the vision-caption pairs) for training. However, collecting and labeling large-scale datasets is time-consuming and expensive for many scenarios and languages. Therefore, sufficient labeled pairs are usually not available. To deal with the label shortage problem, we present a simple yet effective zero-shot approach MultiCapCLIP that can generate visual captions for different scenarios and languages without any labeled vision-caption pairs of downstream datasets. In the training stage, MultiCapCLIP only requires text data for input. Then it conducts two main steps: 1) retrieving concept prompts that preserve the corresponding domain knowledge of new scenarios; 2) auto-encoding the prompts to learn writing styles to output captions in a desired language. In the testing stage, MultiCapCLIP instead takes visual data as input directly to retrieve the concept prompts to generate the final visual descriptions. The extensive experiments on image and video captioning across four benchmarks and four languages (i.e., English, Chinese, German, and French) confirm the effectiveness of our approach. Compared with state-of-the-art zero-shot and weakly-supervised methods, our method achieves 4.8% and 21.5% absolute improvements in terms of BLEU@4

and CIDEr metrics. Our code is available at https://github.com/yangbang18/MultiCapCLIP.

Generate then Select: Open-ended Visual Question Answering Guided by World Knowledge

Xingyu Fu, Sheng Zhang, Gukyeong Kwon, Pramuditha Perera, Henghui Zhu, Yuhao Zhang, Alexander Hanbo Li, William Yang Wang, Zhiguo Wang, Vittorio Castelli, Patrick Ng, Dan Roth and Bing Xiang
11:00-12:30 (Pier 7&8)

The open-ended Visual Question Answering (VQA) task requires AI models to jointly reason over visual and natural language inputs using world knowledge. Recently, pre-trained Language Models (PLM) such as GPT-3 have been applied to the task and shown to be powerful world knowledge sources. However, these methods suffer from low knowledge coverage caused by PLM bias – the tendency to generate certain tokens over other tokens regardless of prompt changes, and high dependency on the PLM quality – only models using GPT-3 can achieve the best result.

To address the aforementioned challenges, we propose RASO: a new VQA pipeline that deploys a generate-then-select strategy guided by world knowledge for the first time. Rather than following the de facto standard to train a multi-modal model that directly generates the VQA answer, {pasted macro 'MODEL'}name first adopts PLM to generate all the possible answers, and then trains a lightweight answer selection model for the correct answer. As proved in our analysis, RASO expands the knowledge coverage from in-domain training data by a large margin. We provide extensive experimentation and show the effectiveness of our pipeline by advancing the state-of-the-art by 4.1% on OK-VQA, without additional computation cost.

Transforming Visual Scene Graphs to Image Captions

Xu Yang, Jiawei Peng, Zihua Wang, Haiyang Xu, Qinghao Ye, Chenliang Li, Songfang Huang, Fei Huang, Zhangzikang Li and Yu Zhang
11:00-12:30 (Pier 7&8)

We propose to TransForm Scene Graphs into more descriptive Captions (TFSGC). In TFSGC, we apply multi-head attention (MHA) to design the Graph Neural Network (GNN) for embedding scene graphs. After embedding, different graph embeddings contain diverse specific knowledge for generating the words with different part-of-speech, e.g., object/attribute embedding is good for generating nouns/adjectives. Motivated by this, we design a Mixture-of-Expert (MOE)-based decoder, where each expert is built on MHA, for discriminating the graph embeddings to generate different kinds of words. Since both the encoder and decoder are built based on the MHA, as a result, we construct a simple and homogeneous encoder-decoder unlike the previous heterogeneous ones which usually apply Fully-Connected-based GNN and LSTM-based decoder. The homogeneous architecture enables us to unify the training configuration of the whole model instead of specifying different training strategies for diverse sub-networks as in the heterogeneous pipeline, which releases the training difficulty. Extensive experiments on the MS-COCO captioning benchmark validate the effectiveness of our TFSGC. The code is in: https://anonymous.4open.science/r/ACL23_TFSGC.

Revealing Single Frame Bias for Video-and-Language Learning

Jie Lei, Tamara Berg and Mohit Bansal
11:00-12:30 (Pier 7&8)

Training an effective video-and-language model intuitively requires multiple frames as model inputs. However, it is unclear whether using multiple frames is beneficial to downstream tasks, and if yes, whether the performance gain is worth the drastically-increased computation and memory costs resulting from using more frames. In this work, we explore single-frame models for video-and-language learning. On a diverse set of video-and-language tasks (including text-to-video retrieval and video question answering), we show the surprising result that, with large-scale pre-training and a proper frame ensemble strategy at inference time, a single-frame trained model that does not consider temporal information can achieve better performance than existing methods that use multiple frames for training. This result reveals the existence of a strong "static appearance bias" in popular video-and-language datasets. Therefore, to allow for a more comprehensive evaluation of video-and-language models, we propose two new retrieval tasks based on existing fine-grained action recognition datasets that encourage temporal modeling. Our code is available at https://github.com/jayleicn/singularity.

MS-DETR: Natural Language Video Localization with Sampling Moment-Moment Interaction

Wang Jing, Aixin Sun, Hao Zhang and Xiaoli Li
11:00-12:30 (Pier 7&8)

Given a text query, the task of Natural Language Video Localization (NLVL) is to localize a temporal moment in an untrimmed video that semantically matches the query. In this paper, we adopt a proposal-based solution that generates proposals (i.e. candidate moments) and then select the best matching proposal. On top of modeling the cross-modal interaction between candidate moments and the query, our proposed Moment Sampling DETR (MS-DETR) enables efficient moment-moment relation modeling. The core idea is to sample a subset of moments guided by the learnable templates with an adopted DETR framework. To achieve this, we design a multi-scale visual-linguistic encoder, and an anchor-guided moment decoder paired with a set of learnable templates. Experimental results on three public datasets demonstrate the superior performance of MS-DETR.

Tackling Modality Heterogeneity with Multi-View Calibration Network for Multimodal Sentiment Detection

yiwei wei, Shaozu Yuan, Ruosong Yang, Lei Shen, zhangmeizhi li, Longbiao Wang and Meng Chen
11:00-12:30 (Pier 7&8)

With the popularity of social media, detecting sentiment from multimodal posts (e.g. image-text pairs) has attracted substantial attention recently. Existing works focus on fusing different features but ignore the challenge of modality heterogeneity. Specifically, different modalities with inherent disparities may bring three problems: 1) introducing redundant visual features during feature fusion; 2) causing feature shift in the representation space; 3) leading to inconsistent annotations for different modal data. All these issues will increase the difficulty in understanding the sentiment of the multimodal content. In this paper, we propose a novel Multi-View Calibration Network (MVCN) to alleviate the above issues systematically. We first propose a text-guided fusion module with novel Sparse-Attention to reduce the negative impacts of redundant visual elements. We then devise a sentiment-based congruity constraint task to calibrate the feature shift in the representation space. Finally, we introduce an adaptive loss calibration strategy to tackle inconsistent annotated labels. Extensive experiments demonstrate the competitiveness of MVCN over previous approaches and achieve state-of-the-art results on two public benchmark datasets.

LMCap: Few-shot Multilingual Image Captioning by Retrieval Augmented Language Model Prompting

Rita Parada Ramos, Bruno Martins and Desmond Elliott
11:00-12:30 (Pier 7&8)

Multilingual image captioning has recently been tackled by training with large-scale machine translated data, which is an expensive, noisy, and time-consuming process. Without requiring any multilingual caption data, we propose LMCap, an image-blind few-shot multilingual captioning model that works by prompting a language model with retrieved captions. Specifically, instead of following the standard encoder-decoder paradigm, given an image, LMCap first retrieves the captions of similar images using a multilingual CLIP encoder. These captions are then combined into a prompt for an XGLM decoder, in order to generate captions in the desired language. In other words, the generation model does not directly process the image, instead it processes retrieved captions. Experiments on the XM3600 dataset of geographically diverse images show that our model is competitive with fully-supervised multilingual captioning models, without requiring any supervised training on any captioning data.

Learning from Children: Improving Image-Caption Pretraining via Curriculum

Hammad Ayyubi, Rahul Lokesh, Alireza Zareian, Bo Wu and Shih-Fu Chang
11:00-12:30 (Pier 7&8)

Image-caption pretraining has been quite successfully used for downstream vision tasks like zero-shot image classification and object detection. However, image-caption pretraining is still a hard problem – it requires multiple concepts (nouns) from captions to be aligned to several

objects in images. To tackle this problem, we go to the roots – the best learner, children. We take inspiration from cognitive science studies dealing with children's language learning to propose a curriculum learning framework. The learning begins with easy-to-align image caption pairs containing one concept per caption. The difficulty is progressively increased with each new phase by adding one more concept per caption. Correspondingly, the knowledge acquired in each learning phase is utilized in subsequent phases to effectively constrain the learning problem to aligning one new concept-object pair in each phase. We show that this learning strategy improves over vanilla image-caption training in various settings – pretraining from scratch, using a pretrained image or/and pretrained text encoder, low data regime etc.

I Spy a Metaphor: Large Language Models and Diffusion Models Co-Create Visual Metaphors
Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki and Smaranda Muresan 11:00-12:30 (Pier 7&8)
Visual metaphors are powerful rhetorical devices used to persuade or communicate creative ideas through images. Similar to linguistic metaphors, they convey meaning implicitly through symbolism and juxtaposition of the symbols. We propose a new task of generating visual metaphors from linguistic metaphors. This is a challenging task for diffusion-based text-to-image models, such as DALL·E 2, since it requires the ability to model implicit meaning and compositionality. We propose to solve the task through the collaboration between Large Language Models (LLMs) and Diffusion Models: Instruct GPT-3 (davinci-002) with Chain-of-Thought prompting generates text that represents a visual elaboration of the linguistic metaphor containing the implicit meaning and relevant objects, which is then used as input to the diffusion-based text-to-image models. Using a human-AI collaboration framework, where humans interact both with the LLM and the top-performing diffusion model, we create a high-quality dataset containing 6,476 visual metaphors for 1,540 linguistic metaphors and their associated visual elaborations. Evaluation by professional illustrators shows the promise of LLM-Diffusion Model collaboration for this task.To evaluate the utility of our Human-AI collaboration framework and the quality of our dataset, we perform both an intrinsic human-based evaluation and an extrinsic evaluation using visual entailment as a downstream task.

A Language-First Approach for Procedure Planning
Jiateng Liu, Sha Li, Zhenhailong Wang, Manling Li and Heng Ji   11:00-12:30 (Pier 7&8)
Procedure planning, or the ability to predict a series of steps that can achieve a given goal conditioned on the current observation, is critical for building intelligent embodied agents that can assist users in everyday tasks. Encouraged by the recent success of language models (LMs) for zero-shot and few-shot planning, we hypothesize that LMs may be equipped with stronger priors for planning compared to their visual counterparts. To this end, we propose a language-first procedure planning framework with a modularized design: we first align the current and goal observations with corresponding steps and then use a pre-trained LM to predict the intermediate steps. Under this framework, we find that using an image captioning model for alignment can already match state-of-the-art performance and by designing a double retrieval model conditioned over current and goal observations jointly, we can achieve large improvements (19.2%-98.9% relatively higher success rate than state-of-the-art) on both COIN and CrossTask benchmarks. Our work verifies the planning ability of LMs and demonstrates how LMs can serve as a powerful "reasoning engine" even when the input is provided in another modality.

Gradient-Boosted Decision Tree for Listwise Context Model in Multimodal Review Helpfulness Prediction
Thong Thanh Nguyen, Xiaobao Wu, Xinshuai Dong, Cong-Duy T. Nguyen, Zhen Hai, Lidong Bing and Anh Tuan Luu 11:00-12:30 (Pier 7&8)
Multimodal Review Helpfulness Prediction (MRHP) aims to rank product reviews based on predicted helpfulness scores and has been widely applied in e-commerce via presenting customers with useful reviews. Previous studies commonly employ fully-connected neural networks (FCNNs) as the final score predictor and pairwise loss as the training objective. However, FCNNs have been shown to perform inefficient splitting for review features, making the model difficult to clearly differentiate helpful from unhelpful reviews. Furthermore, pairwise objective, which works on review pairs, may not completely capture the MRHP goal to produce the ranking for the entire review list, and possibly induces low generalization during testing. To address these issues, we propose a listwise attention network that clearly captures the MRHP ranking context and a listwise optimization objective that enhances model generalization. We further propose gradient-boosted decision tree as the score predictor to efficaciously partition product reviews' representations. Extensive experiments demonstrate that our method achieves state-of-the-art results and polished generalization performance on two large-scale MRHP benchmark datasets.

Transferring General Multimodal Pretrained Models to Text Recognition
Junyang Lin, Xuancheng Ren, Yichang Zhang, Gao Liu, Peng Wang, An Yang and Chang Zhou   11:00-12:30 (Pier 7&8)
This paper proposes a new method, OFA-OCR, to transfer multimodal pretrained models to text recognition. Specifically, we recast text recognition as image captioning and directly transfer a unified vision-language pretrained model to the end task. Without pretraining on large-scale annotated or synthetic text recognition data, OFA-OCR outperforms the baselines and achieves state-of-the-art performance in the Chinese text recognition benchmark. Additionally, we construct an OCR pipeline with OFA-OCR, and we demonstrate that it can achieve competitive performance with the product-level API.

Prosody-TTS: Improving Prosody with Masked Autoencoder and Conditional Diffusion Model For Expressive Text-to-Speech
Rongjie Huang, Chunlei Zhang, Yi Ren, Zhou Zhao and Dong Yu   11:00-12:30 (Pier 7&8)
Expressive text-to-speech aims to generate high-quality samples with rich and diverse prosody, which is hampered by **dual challenges**: 1) prosodic attributes in highly dynamic voices are difficult to capture and model without intonation; and 2) highly multimodal prosodic representations cannot be well learned by simple regression (e.g., MSE) objectives, which causes blurry and over-smoothing predictions. This paper proposes Prosody-TTS, a two-stage pipeline that enhances **prosody modeling and sampling** by introducing several components: 1) a self-supervised masked autoencoder to model the prosodic representation without relying on text transcriptions or local prosody attributes, which ensures to cover diverse speaking voices with superior generalization; and 2) a diffusion model to sample diverse prosodic patterns within the latent space, which prevents TTS models from generating samples with dull prosodic performance. Experimental results show that Prosody-TTS achieves new state-of-the-art in text-to-speech with natural and expressive synthesis. Both subjective and objective evaluation demonstrate that it exhibits superior audio quality and prosody naturalness with rich and diverse prosodic attributes. Audio samples are available at https://improved_prosody.github.io

Hybrid Transducer and Attention based Encoder-Decoder Modeling for Speech-to-Text Tasks
Yun Tang, Anna Y Sun, Hirofumi Inaguma, Xinyue Chen, Ning Dong, Xutai Ma, Paden D Tomasello and Juan Pino   11:00-12:30 (Pier 7&8)
Transducer and Attention based Encoder-Decoder (AED) are two widely used frameworks for speech-to-text tasks. They are designed for different purposes and each has its own benefits and drawbacks for speech-to-text tasks. In order to leverage strengths of both modeling methods, we propose a solution by combining Transducer and Attention based Encoder-Decoder (TAED) for speech-to-text tasks. The new method leverages AED's strength in non-monotonic sequence to sequence learning while retaining Transducer's streaming property. In the proposed framework, Transducer and AED share the same speech encoder. The predictor in Transducer is replaced by the decoder in the AED model, and the outputs of the decoder are conditioned on the speech inputs instead of outputs from an unconditional language model. The proposed solution ensures that the model is optimized by covering all possible read/write scenarios and creates a matched environment for streaming applications. We evaluate the proposed approach on the MUST-C dataset and the findings demonstrate that TAED performs significantly better than Transducer for offline automatic speech recognition (ASR) and speech-to-text translation (ST) tasks. In the streaming case, TAED outperforms Transducer in the ASR task and one ST direction while comparable results are achieved in another translation direction.

SLUE Phase-2: A Benchmark Suite of Diverse Spoken Language Understanding Tasks
Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan S Sharma, Wei-Lun Wu, Hung-yi Lee, Karen Livescu and Shinji Watanabe
11:00-12:30 (Pier 7&8)
Spoken language understanding (SLU) tasks have been studied for many decades in the speech research community, but have not received as much attention as lower-level tasks like speech and speaker recognition. In this work, we introduce several new annotated SLU benchmark tasks based on freely available speech data, which complement existing benchmarks and address gaps in the SLU evaluation landscape. We contribute four tasks: question answering and summarization involve inference over longer speech sequences; named entity localization addresses the speech-specific task of locating the targeted content in the signal; dialog act classification identifies the function of a given speech utterance. In order to facilitate the development of SLU models that leverage the success of pre-trained speech representations, we will release a new benchmark suite, including for each task (i) curated annotations for a relatively small fine-tuning set, (ii) reproducible pipeline (speech recognizer + text model) and end-to-end baseline models and evaluation metrics, (iii) baseline model performance in various types of systems for easy comparisons. We present the details of data collection and annotation and the performance of the baseline models. We also analyze the sensitivity of pipeline models' performance to the speech recognition accuracy, using more than 20 publicly available speech recognition models.

MOSPC: MOS Prediction Based on Pairwise Comparison
Kexin Wang, Yunlong Zhao, Qianqian Dong, Tom Ko and Mingxuan Wang
11:00-12:30 (Pier 7&8)
As a subjective metric to evaluate the quality of synthesized speech, Mean opinion score(MOS) usually requires multiple annotators to score the same speech. Such an annotation approach requires a lot of manpower and is also time-consuming. MOS prediction model for automatic evaluation can significantly reduce labor cost. In previous works, it is difficult to accurately rank the quality of speech when the MOS scores are close. However, in practical applications, it is more important to correctly rank the quality of synthesis systems or sentences than simply predicting MOS scores. Meanwhile, as each annotator scores multiple audios during annotation, the score is probably a relative value based on the first or the first few speech scores given by the annotator. Motivated by the above two points, we propose a general framework for MOS prediction based on pair comparison (MOSPC), and we utilize C-Mixup algorithm to enhance the generalization performance of MOSPC. The experiments on BVCC and VCC2018 show that our framework outperforms the baselines on most of the correlation coefficient metrics, especially on the metric KTAU related to quality ranking. And our framework also surpasses the strong baseline in ranking accuracy on each fine-grained segment. These results indicate that our framework contributes to improving the ranking accuracy of speech quality.

UMRSpell: Unifying the Detection and Correction Parts of Pre-trained Models towards Chinese Missing, Redundant, and Spelling Correction
Zheyu He, Yujin Zhu, Linlin Wang and Liang Xu
11:00-12:30 (Pier 7&8)
Chinese Spelling Correction (CSC) is the task of detecting and correcting misspelled charac- ters in Chinese texts. As an important step for various downstream tasks, CSC confronts two challenges: 1) Character-level errors consist not only of spelling errors but also of missing and redundant ones that cause variable length between input and output texts, for which most CSC methods could not handle well because of the consistence length of texts required by their inherent detection-correction framework. Con- sequently, the two errors are considered out- side the scope and left to future work, despite the fact that they are widely found and bound to CSC task in Chinese industrial scenario, such as Automatic Speech Recognition (ASR) and Optical Character Recognition (OCR). 2) Most existing CSC methods focus on either detector or corrector and train different mod- els for each one, respectively, leading to in- sufficiency of parameters sharing. To address these issues, we propose a novel model UMR- Spell to learn detection and correction parts together at the same time from a multi-task learning perspective by using a detection trans- mission self-attention matrix, and flexibly deal with both missing, redundant, and spelling er- rors through re-tagging rules. Furthermore, we build a new dataset ECMR-2023 containing five kinds of character-level errors to enrich the CSC task closer to real-world applications. Ex- periments on both SIGHAN benchmarks and ECMR-2023 demonstrate the significant effec- tiveness of UMRSpell over previous represen- tative baselines.

A dynamic programming algorithm for span-based nested named-entity recognition in $O(n^2)$
Caio Corro
11:00-12:30 (Pier 7&8)
Span-based nested named-entity recognition (NER) has a cubic-time complexity using a variant of the CYK algorithm. We show that by adding a supplementary structural constraint on the search space, nested NER has a quadratic-time complexity, that is the same asymptotic complexity than the non-nested case. The proposed algorithm covers a large part of three standard English benchmarks and delivers compa- rable experimental results.

Scaling in Cognitive Modelling: a Multilingual Approach to Human Reading Times
Andrea Gregor de Varda and Marco Marelli
11:00-12:30 (Pier 7&8)
Neural language models are increasingly valued in computational psycholinguistics, due to their ability to provide conditional probability distributions over the lexicon that are predictive of human processing times. Given the vast array of available models, it is of both theoretical and methodological importance to assess what features of a model influence its psychometric quality. In this work we focus on parameter size, showing that larger Transformer-based language models generate probabilistic estimates that are less predictive of early eye-tracking measurements reflecting lexical access and early semantic integration. However, relatively bigger models show an advantage in capturing late eye-tracking measurements that reflect the full semantic and syntactic integration of a word into the current language context. Our results are supported by eye movement data in ten languages and consider four models, spanning from 564M to 4.5B parameters.

Distinguishing Address vs. Reference Mentions of Personal Names in Text
Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, Melissa Ferguson and Stav Atir
11:00-12:30 (Pier 7&8)
Detecting named entities in text has long been a core NLP task. However, not much work has gone into distinguishing whether an entity mention is addressing the entity vs. referring to the entity; e.g., *John, would you turn the light off?* vs. *John turned the light off*. While this distinction is marked by a *vocative case* marker in some languages, many modern Indo-European languages such as English do not use such explicit vocative markers, and the distinction is left to be interpreted in context. In this paper, we present a new annotated dataset that captures the *address* vs. *reference* distinction in English, an automatic tagger that performs at 85% accuracy in making this distinction, and demonstrate how this distinction is important in NLP and computational social science applications in English language.

End-to-End Argument Mining over Varying Rhetorical Structures
Elena Chistova
11:00-12:30 (Pier 7&8)
Rhetorical Structure Theory implies no single discourse interpretation of a text, and the limitations of RST parsers further exacerbate incon- sistent parsing of similar structures. Therefore, it is important to take into account that the same argumentative structure can be found in semantically similar texts with varying rhetorical structures. In this work, the differences between paraphrases within the same argument scheme are evaluated from a rhetorical perspective. The study proposes a deep dependency parsing model to assess the connection between rhetorical and argument structures. The model utilizes rhetorical relations; RST structures of paraphrases serve as training data augmenta- tions. The method allows for end-to-end argumentation analysis using a rhetorical tree instead of a word sequence. It is evaluated on the

bilingual Microtexts corpus, and the first results on fully-fledged argument parsing for the Russian version of the corpus are reported. The results suggest that argument mining can benefit from multiple variants of discourse structure.

Towards Generative Event Factuality Prediction

John Murzaku, Tyler G Osborne, Amittai F Aviram and Owen Rambow   11:00-12:30 (Pier 7&8)
We present a novel end-to-end generative model and system for predicting event factuality holders, targets, and their associated factuality values. We perform the first experiments using all sources and targets of factuality statements from the FactBank corpus. We perform multi-task learning with other tasks and event-factuality corpora to improve on the FactBank source and target task. We argue that careful domain specific target text output format in generative systems is important and verify this with multiple experiments on target text output structure. We redo previous state-of-the-art author-only event factuality experiments and also offer insights towards a generative paradigm for the author-only event factuality prediction task.

Discourse Analysis via Questions and Answers: Parsing Dependency Structures of Questions Under Discussion

Wei-Jen Ko, Yating Wu, Cutter J Dalton, Dananjay T Srinivas, Greg Durrett and Junyi Jessy Li   11:00-12:30 (Pier 7&8)
Automatic discourse processing is bottlenecked by data: current discourse formalisms pose highly demanding annotation tasks involving large taxonomies of discourse relations, making them inaccessible to lay annotators. This work instead adopts the linguistic framework of Questions Under Discussion (QUD) for discourse analysis and seeks to derive QUD structures automatically. QUD views each sentence as an answer to a question triggered in prior context; thus, we characterize relationships between sentences as free-form questions, in contrast to exhaustive fine-grained taxonomies. We develop the first-of-its-kind QUD parser that derives a dependency structure of questions over full documents, trained using a large, crowdsourced question-answering dataset DCQA (Ko et al., 2022). Human evaluation results show that QUD dependency parsing is possible for language models trained with this crowdsourced, generalizable annotation scheme. We illustrate how our QUD structure is distinct from RST trees, and demonstrate the utility of QUD analysis in the context of document simplification. Our findings show that QUD parsing is an appealing alternative for automatic discourse processing.

Connective Prediction for Implicit Discourse Relation Recognition via Knowledge Distillation

Hongyi Wu, Hao Zhou, Man Lan, Yuanbin Wu and Yadong Zhang   11:00-12:30 (Pier 7&8)
Implicit discourse relation recognition (IDRR) remains a challenging task in discourse analysis due to the absence of connectives. Most existing methods utilize one-hot labels as the sole optimization target, ignoring the internal association among connectives. Besides, these approaches spend lots of effort on template construction, negatively affecting the generalization capability. To address these problems,we propose a novel Connective Prediction via Knowledge Distillation (CP-KD) approach to instruct large-scale pre-trained language models (PLMs) mining the latent correlations between connectives and discourse relations, which is meaningful for IDRR. Experimental results on the PDTB 2.0/3.0 and CoNLL2016 datasets show that our method significantly outperforms the state-of-the-art models on coarse-grained and fine-grained discourse relations. Moreover, our approach can be transferred to explicit discourse relation recognition(EDRR) and achieve acceptable performance.

Verifying Annotation Agreement without Multiple Experts: A Case Study with Gujarati SNACS

Maitrey Mehta and Vivek Srikumar   11:00-12:30 (Pier 7&8)
Good datasets are a foundation of NLP research, and form the basis for training and evaluating models of language use. While creating datasets, the standard practice is to verify the annotation consistency using a committee of human annotators. This norm assumes that multiple annotators are available, which is not the case for highly specialized tasks or low-resource languages. In this paper, we ask: Can we evaluate the quality of a dataset constructed by a single human annotator? To address this question, we propose four weak verifiers to help estimate dataset quality, and outline when each may be employed. We instantiate these strategies for the task of semantic analysis of adpositions in Gujarati, a low-resource language, and show that our weak verifiers concur with a double-annotation study. As an added contribution, we also release the first dataset with semantic annotations in Gujarati along with several model baselines.

AxomiyaBERTa: A Phonologically-aware Transformer Model for Assamese

Abhijnan Nath, Sheikh Abdul Mannan and Nikhil Krishnaswamy   11:00-12:30 (Pier 7&8)
Despite their successes in NLP, Transformer-based language models still require extensive computing resources and suffer in low-resource or low-compute settings. In this paper, we present AxomiyaBERTa, a novel BERT model for Assamese, a morphologically-rich low-resource language (LRL) of Eastern India. AxomiyaBERTa is trained only on the masked language modeling (MLM) task, without the typical additional next sentence prediction (NSP) objective, and our results show that in resource-scarce settings for very low-resource languages like Assamese, MLM alone can be successfully leveraged for a range of tasks. AxomiyaBERTa achieves SOTA on token-level tasks like Named Entity Recognition and also performs well on "longer-context" tasks like Cloze-style QA and Wiki Title Prediction, with the assistance of a novel embedding disperser and phonological signals respectively. Moreover, we show that AxomiyaBERTa can leverage phonological signals for even more challenging tasks, such as a novel cross-document coreference task on a translated version of the ECB+ corpus, where we present a new SOTA result for an LRL. Our source code and evaluation scripts may be found at https://github.com/csu-signal/axomiyaberta.

Multilingual Sequence-to-Sequence Models for Hebrew NLP

Matan Eyal, Hila Noga, Roee Aharoni, Idan Szpektor and Reut Tsarfaty   11:00-12:30 (Pier 7&8)
Recent work attributes progress in NLP to large language models (LMs) with increased model size and large quantities of pretraining data. Despite this, current state-of-the-art LMs for Hebrew are both under-parameterized and under-trained compared to LMs in other languages. Additionally, previous work on pretrained Hebrew LMs focused on encoder-only models. While the encoder-only architecture is beneficial for classification tasks, it does not cater well for sub-word prediction tasks, such as Named Entity Recognition, when considering the morphologically rich nature of Hebrew. In this paper we argue that sequence-to-sequence generative architectures are more suitable for large LMs in morphologically rich languages (MRLs) such as Hebrew. We demonstrate this by casting tasks in the Hebrew NLP pipeline as text-to-text tasks, for which we can leverage powerful multilingual, pretrained sequence-to-sequence models as mT5, eliminating the need for a separate, specialized, morpheme-based, decoder. Using this approach, our experiments show substantial improvements over previously published results on all existing Hebrew NLP benchmarks. These results suggest that multilingual sequence-to-sequence models present a promising building block for NLP for MRLs.

Hybrid Knowledge Transfer for Improved Cross-Lingual Event Detection via Hierarchical Sample Selection

Luis Fernando Guzman Nateras, Franck Dernoncourt and Thien Huu Nguyen   11:00-12:30 (Pier 7&8)
In this paper, we address the Event Detection task under a zero-shot cross-lingual setting where a model is trained on a source language but evaluated on a distinct target language for which there is no labeled data available. Most recent efforts in this field follow a direct transfer approach in which the model is trained using language-invariant features and then directly applied to the target language. However, we argue that these methods fail to take advantage of the benefits of the data transfer approach where a cross-lingual model is trained on target-language data and is able to learn task-specific information from syntactical features or word-label relations in the target language. As such, we propose a hybrid knowledge-transfer approach that leverages a teacher-student framework where the teacher and student networks are trained following the direct and data transfer approaches, respectively. Our method is complemented by a hierarchical training-sample selection

scheme designed to address the issue of noisy labels being generated by the teacher model. Our model achieves state-of-the-art results on 9 morphologically-diverse target languages across 3 distinct datasets, highlighting the importance of exploiting the benefits of hybrid transfer.

Mini-Model Adaptation: Efficiently Extending Pretrained Models to New Languages via Aligned Shallow Training
Kelly Marchisio, Patrick Lewis, Yihong Chen and Mikel Artetxe 11:00-12:30 (Pier 7&8)
Prior work shows that it is possible to expand pretrained Masked Language Models (MLMs) to new languages by learning a new set of embeddings, while keeping the transformer body frozen. Despite learning a small subset of parameters, this approach is not compute-efficient, as training the new embeddings requires a full forward and backward pass over the entire model. We propose mini-model adaptation, a compute-efficient alternative that builds a shallow mini-model from a fraction of a large model's parameters. New language-specific embeddings can then be efficiently trained over the mini-model and plugged into the aligned large model for rapid cross-lingual transfer. We explore two approaches to learn mini-models: MINIJOINT, which jointly pretrains the primary model and the mini-model using a single transformer with a secondary MLM head at a middle layer; and MINIPOST, where we start from a regular pretrained model, build a mini-model by extracting and freezing a few layers, and learn a small number of parameters on top. Experiments on XNLI, MLQA and PAWS-X show that mini-model adaptation matches the performance of the standard approach using up to 2.3x less compute on average.

Zero-shot Cross-lingual Transfer With Learned Projections Using Unlabeled Target-Language Data
Ujan Deb, Ridayesh Ramesh Parab and Preethi Jyothi 11:00-12:30 (Pier 7&8)
Adapters have emerged as a parameter-efficient Transformer-based framework for cross-lingual transfer by inserting lightweight language-specific modules (language adapters) and task-specific modules (task adapters) within pretrained multilingual models. Zero-shot transfer is enabled by pairing the language adapter in the target language with an appropriate task adapter in a source language. If our target languages are known apriori, we explore how zero-shot transfer can be further improved within the adapter framework by utilizing unlabeled text during task-specific finetuning. We construct language-specific subspaces using standard linear algebra constructs and selectively project source-language representations into the target language subspace during task-specific finetuning using two schemes. Our experiments on three cross-lingual tasks, Named Entity Recognition (NER), Question Answering (QA) and Natural Language Inference (NLI) yield consistent benefits compared to adapter baselines over a wide variety of target languages with up to 11% relative improvement in NER, 2% relative improvement in QA and 5% relative improvement in NLI.

X-RiSAWOZ: High-Quality End-to-End Multilingual Dialogue Datasets and Few-shot Agents
Mehrad Moradshahi, Tianhao Shen, Kalika Bali, Monojit Choudhury, Gael de Chalendar, Anmol Goel, Sungkyun Kim, Prashant Kodali, Ponnurangam Kumaraguru, Nasredine Semmar, Sina Semnani, Jiwon Seo, Vivek Seshadri, Manish Shrivastava, Michael Sun, Aditya Yadavalli, Chaobin You, Deyi Xiong and Monica S Lam 11:00-12:30 (Pier 7&8)
Task-oriented dialogue research has mainly focused on a few popular languages like English and Chinese, due to the high dataset creation cost for a new language. To reduce the cost, we apply manual editing to automatically translated data. We create a new multilingual benchmark, X-RiSAWOZ, by translating the Chinese RiSAWOZ to 4 languages: English, French, Hindi, Korean; and a code-mixed English-Hindi language. X-RiSAWOZ has more than 18,000 human-verified dialogue utterances for each language, and unlike most multilingual prior work, is an end-to-end dataset for building fully-functioning agents.
The many difficulties we encountered in creating X-RiSAWOZ led us to develop a toolset to accelerate the post-editing of a new language dataset after translation. This toolset improves machine translation with a hybrid entity alignment technique that combines neural with dictionary-based methods, along with many automated and semi-automated validation checks.
We establish strong baselines for X-RiSAWOZ by training dialogue agents in the zero- and few-shot settings where limited gold data is available in the target language. Our results suggest that our translation and post-editing methodology and toolset can be used to create new high-quality multilingual dialogue agents cost-effectively. Our dataset, code, and toolkit are released open-source.

DLAMA: A Framework for Curating Culturally Diverse Facts for Probing the Knowledge of Pretrained Language Models
Amr Keleg and Walid Magdy 11:00-12:30 (Pier 7&8)
A few benchmarking datasets have been released to evaluate the factual knowledge of pretrained language models. These benchmarks (e.g., LAMA, and ParaRel) are mainly developed in English and later are translated to form new multilingual versions (e.g., mLAMA, and mParaRel). Results on these multilingual benchmarks suggest that using English prompts to recall the facts from multilingual models usually yields significantly better and more consistent performance than using non-English prompts. Our analysis shows that mLAMA is biased toward facts from Western countries, which might affect the fairness of probing models. We propose a new framework for curating factual triples from Wikidata that are culturally diverse. A new benchmark DLAMA-v1 is built of factual triples from three pairs of contrasting cultures having a total of 78,259 triples from 20 relation predicates. The three pairs comprise facts representing the (Arab and Western), (Asian and Western), and (South American and Western) countries respectively. Having a more balanced benchmark (DLAMA-v1) supports that mBERT performs better on Western facts than non-Western ones, while monolingual Arabic, English, and Korean models tend to perform better on their culturally proximate facts. Moreover, both monolingual and multilingual models tend to make a prediction that is culturally or geographically relevant to the correct label, even if the prediction is wrong.

Predicting Human Translation Difficulty Using Automatic Word Alignment
Zheng Wei Lim, Trevor Cohn, Charles Kemp and Ekaterina Vylomova 11:00-12:30 (Pier 7&8)
Translation difficulty arises when translators are required to resolve translation ambiguity from multiple possible translations. Translation difficulty can be measured by recording the diversity of responses provided by human translators and the time taken to provide these responses, but these behavioral measures are costly and do not scale. In this work, we use word alignments computed over large scale bilingual corpora to develop predictors of lexical translation difficulty. We evaluate our approach using behavioural data from translations recorded both in and out of context, and report results that improve on a previous embedding-based approach (Thompson et al., 2020). Our work can therefore contribute to a deeper understanding of cross-lingual differences and of causes of translation difficulty.

Unifying Cross-Lingual and Cross-Modal Modeling Towards Weakly Supervised Multilingual Vision-Language Pre-training
Zejun Li, Zhihao Fan, Jingjing Chen, Qi Zhang, Xuanjing Huang and Zhongyu Wei 11:00-12:30 (Pier 7&8)
Multilingual Vision-Language Pre-training (VLP) is a promising but challenging topic due to the lack of large-scale multilingual image-text pairs. Existing works address the problem by translating English data into other languages, which is intuitive and the generated data is usually limited in form and scale. In this paper, we explore a more practical and scalable setting: weakly supervised multilingual VLP with only English image-text pairs and multilingual text corpora. We argue that the universal multilingual representation learned from texts allows the cross-modal interaction learned in English to be transferable to other languages. To this end, we propose a framework to effectively unify cross-lingual and cross-modal pre-training. For unified modeling on different data, we design an architecture with flexible modules to learn different interactions. Moreover, two unified tasks are introduced to efficiently guide the unified cross-lingual cross-modal learning. Extensive experiments demonstrate that our pre-trained model learns universal multilingual multimodal representations, allowing effective cross-lingual transfer on multimodal tasks. Code and models are available at https://github.com/FudanDISC/weakly-supervised-mVLP.

Beyond English-Centric Bitexts for Better Multilingual Language Representation Learning

Barun Patra, Saksham Singhal, Shaohan Huang, Zewen Chi, Li Dong, Furu Wei, Vishrav Chaudhary and Xia Song 11:00-12:30 (Pier 7&8)
In this paper, we elaborate upon recipes for building multilingual representation models that are not only competitive with existing state-of-the-art models but are also more parameter efficient, thereby promoting better adoption in resource-constrained scenarios and practical applications. We show that going beyond English-centric bitexts, coupled with a novel sampling strategy aimed at reducing under-utilization of training data, substantially boosts performance across model sizes for both Electra and MLM pre-training objectives. We introduce XY-LENT: X-Y bitext enhanced Language ENcodings using Transformers which not only achieves state-of-the-art performance over 5 cross-lingual tasks within all model size bands, is also competitive across bands. Our XY-LENT XL variant outperforms XLM-R XXL and exhibits competitive performance with mT5 XXL while being 5x and 6x smaller respectively. We then show that our proposed method helps ameliorate the curse of multilinguality, with the XY-LENT XL achieving 99.3% GLUE performance and 98.5% SQuAD 2.0 performance compared to a SoTA English only model in the same size band. We then analyze our models performance on extremely low resource languages and posit that scaling alone may not be sufficient for improving the performance in this scenario

On-the-fly Cross-lingual Masking for Multilingual Pre-training
Xi Ai and Bin Fang 11:00-12:30 (Pier 7&8)
In multilingual pre-training with the objective of MLM (masked language modeling) on multiple monolingual corpora, multilingual models only learn cross-linguality implicitly from isomorphic spaces formed by overlapping different language spaces due to the lack of explicit cross-lingual forward pass. In this work, we present CLPM (Cross-lingual Prototype Masking), a dynamic and token-wise masking scheme, for multilingual pre-training, using a special token $[\mathcal{C}]_x$ to replace a random token x in the input sentence. $[\mathcal{C}]_x$ is a cross-lingual prototype for x and then forms an explicit cross-lingual forward pass. We instantiate CLPM for the multilingual pre-training phase of UNMT (unsupervised neural machine translation), and experiments show that CLPM can consistently improve the performance of UNMT models on $\{De, Ro, Ne\} \leftrightarrow En$. Beyond UNMT or bilingual tasks, we show that CLPM can consistently improve the performance of multilingual models on cross-lingual classification.

Parameter-Efficient Finetuning for Robust Continual Multilingual Learning
Kartikeya Badola, Shachi Dave and Partha Talukdar 11:00-12:30 (Pier 7&8)
We introduce and study the problem of Continual Multilingual Learning (CML) where a previously trained multilingual model is periodically updated using new data arriving in stages. If the new data is present only in a subset of languages, we find that the resulting model shows improved performance only on the languages included in the latest update (and a few closely related languages) while its performance on all the remaining languages degrade significantly. We address this challenge by proposing LAFT-URIEL, a parameter-efficient finetuning strategy which aims to increase the number of languages on which the model improves after an update, while reducing the magnitude of loss in performance for the remaining languages. LAFT-URIEL uses linguistic knowledge to balance overfitting and knowledge sharing across languages, allowing for an additional 25% of task languages to see an improvement in performance after an update, while also reducing the average magnitude of losses on the remaining languages by 78% relative.

White-Box Multi-Objective Adversarial Attack on Dialogue Generation
Yufei Li, Zexin Li, Yingfan Gao and Cong Liu 11:00-12:30 (Pier 7&8)
Pre-trained transformers are popular in state-of-the-art dialogue generation (DG) systems. Such language models are, however, vulnerable to various adversarial samples as studied in traditional tasks such as text classification, which inspires our curiosity about their robustness in DG systems. One main challenge of attacking DG models is that perturbations on the current sentence can hardly degrade the response accuracy because the unchanged chat histories are also considered for decision-making. Instead of merely pursuing pitfalls of performance metrics such as BLEU, ROUGE, we observe that crafting adversarial samples to force longer generation outputs benefits attack effectiveness—the generated responses are typically irrelevant, lengthy, and repetitive. To this end, we propose a white-box multi-objective attack method called DGSlow. Specifically, DGSlow balances two objectives—generation accuracy and length, via a gradient-based multi-objective optimizer and applies an adaptive searching mechanism to iteratively craft adversarial samples with only a few modifications. Comprehensive experiments on four benchmark datasets demonstrate that DGSlow could significantly degrade state-of-the-art DG models with a higher success rate than traditional accuracy-based methods. Besides, our crafted sentences also exhibit strong transferability in attacking other models.

DSRM: Boost Textual Adversarial Training with Distribution Shift Risk Minimization
SongYang Gao, Shihan Dou, Yan Liu, Xiao Wang, Qi Zhang, Zhongyu Wei, Jin Ma and Ying Shan 11:00-12:30 (Pier 7&8)
Adversarial training is one of the best-performing methods in improving the robustness of deep language models. However, robust models come at the cost of high time consumption, as they require multi-step gradient ascents or word substitutions to obtain adversarial samples. In addition, these generated samples are deficient in grammatical quality and semantic consistency, which impairs the effectiveness of adversarial training. To address these problems, we introduce a novel, effective procedure for instead adversarial training with only clean data. Our procedure, distribution shift risk minimization (DSRM), estimates the adversarial loss by perturbing the input data's probability distribution rather than their embeddings. This formulation results in a robust model that minimizes the expected global loss under adversarial attacks. Our approach requires zero adversarial samples for training and reduces time consumption by up to 70% compared to current best-performing adversarial training methods. Experiments demonstrate that DSRM considerably improves BERT's resistance to textual adversarial attacks and achieves state-of-the-art robust accuracy on various benchmarks.

CASN:Class-Aware Score Network for Textual Adversarial Detection
Rong Bao, Rui Zheng, Liang Ding, Qi Zhang and Dacheng Tao 11:00-12:30 (Pier 7&8)
Adversarial detection aims to detect adversarial samples that threaten the security of deep neural networks, which is an essential step toward building robust AI systems. Density-based estimation is widely considered as an effective technique by explicitly modeling the distribution of normal data and identifying adversarial ones as outliers. However, these methods suffer from significant performance degradation when the adversarial samples lie close to the non-adversarial data manifold. To address this limitation, we propose a score-based generative method to implicitly model the data distribution. Our approach utilizes the gradient of the log-density data distribution and calculates the distribution gap between adversarial and normal samples through multi-step iterations using Langevin dynamics. In addition, we use supervised contrastive learning to guide the gradient estimation using label information, which avoids collapsing to a single data manifold and better preserves the anisotropy of the different labeled data distributions. Experimental results on three text classification tasks upon four advanced attack algorithms show that our approach is a significant improvement (average +15.2 F1 score against previous SOTA) over previous detection methods.

Conformal Nucleus Sampling
Shauli Ravfogel, Yoav Goldberg and Jacob Goldberger 11:00-12:30 (Pier 7&8)
Language models generate text based on successively sampling the next word. A decoding procedure based on nucleus (top-p) sampling chooses from the smallest possible set of words whose cumulative probability exceeds the probability p. In this work, we assess whether a top-p set is indeed aligned with its probabilistic meaning in various linguistic contexts. We employ conformal prediction, a calibration procedure that focuses on the construction of minimal prediction sets according to a desired confidence level, to calibrate the parameter p as a function of the entropy of the next word distribution. We find that OPT models are overconfident, and that calibration shows a moderate inverse scaling with model size.

A Gradient Control Method for Backdoor Attacks on Parameter-Efficient Tuning
Naibin Gu, Peng Fu, Xiyu Liu, Zhengxiao Liu, Zheng Lin and Weiping Wang   11:00-12:30 (Pier 7&8)
Parameter-Efficient Tuning (PET) has shown remarkable performance by fine-tuning only a small number of parameters of the pre-trained language models (PLMs) for the downstream tasks, while it is also possible to construct backdoor attacks due to the vulnerability of pre-trained weights. However, a large reduction in the number of attackable parameters in PET will cause the user's fine-tuning to greatly affect the effectiveness of backdoor attacks, resulting in backdoor forgetting. We find that the backdoor injection process can be regarded as multi-task learning, which has a convergence imbalance problem between the training of clean and poisoned data. And this problem might result in forgetting the backdoor. Based on this finding, we propose a gradient control method to consolidate the attack effect, comprising two strategies. One controls the gradient magnitude distribution cross layers within one task and the other prevents the conflict of gradient directions between tasks. Compared with previous backdoor attack methods in the scenario of PET, our method improve the effect of the attack on sentiment classification and spam detection respectively, which shows that our method is widely applicable to different tasks.

Figurative Language Processing: A Linguistically Informed Feature Analysis of the Behavior of Language Models and Humans
Hyewon Jang, Qi Yu and Diego Frassinelli   11:00-12:30 (Pier 7&8)
Recent years have witnessed a growing interest in investigating what Transformer-based language models (TLMs) actually learn from the training data. This is especially relevant for complex tasks such as the understanding of non-literal meaning. In this work, we probe the performance of three black-box TLMs and two intrinsically transparent white-box models on figurative language classification of sarcasm, similes, idioms, and metaphors. We conduct two studies on the classification results to provide insights into the inner workings of such models. With our first analysis on feature importance, we identify crucial differences in model behavior. With our second analysis using an online experiment with human annotators, we inspect different linguistic characteristics of the four figurative language types.

Hybrid Uncertainty Quantification for Selective Text Classification in Ambiguous Tasks
Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev and Artem Shelmanov   11:00-12:30 (Pier 7&8)
Many text classification tasks are inherently ambiguous, which results in automatic systems having a high risk of making mistakes, in spite of using advanced machine learning models. For example, toxicity detection in user-generated content is a subjective task, and notions of toxicity can be annotated according to a variety of definitions that can be in conflict with one another. Instead of relying solely on automatic solutions, moderation of the most difficult and ambiguous cases can be delegated to human workers. Potential mistakes in automated classification can be identified by using uncertainty estimation (UE) techniques. Although UE is a rapidly growing field within natural language processing, we find that state-of-the-art UE methods estimate only epistemic uncertainty and show poor performance, or under-perform trivial methods for ambiguous tasks such as toxicity detection. We argue that in order to create robust uncertainty estimation methods for ambiguous tasks it is necessary to account also for aleatoric uncertainty. In this paper, we propose a new uncertainty estimation method that combines epistemic and aleatoric UE methods. We show that by using our hybrid method, we can outperform state-of-the-art UE methods for toxicity detection and other ambiguous text classification tasks.

Measuring the Instability of Fine-Tuning
Yupei Du and Dong Nguyen   11:00-12:30 (Pier 7&8)
Fine-tuning pre-trained language models on downstream tasks with varying random seeds has been shown to be unstable, especially on small datasets. Many previous studies have investigated this instability and proposed methods to mitigate it. However, most of these studies only used the standard deviation of performance scores (SD) as their measure, which is a narrow characterization of instability. In this paper, we analyze SD and six other measures quantifying instability of different granularity levels. Moreover, we propose a systematic evaluation framework of these measures' validity. Finally, we analyze the consistency and difference between different measures by reassessing existing instability mitigation methods. We hope our results will inform better measurements of the fine-tuning instability.

Language Model Analysis for Ontology Subsumption Inference
Yuan He, Jiaoyan Chen, Ernesto Jimenez-Ruiz, Hang Dong and Ian Horrocks   11:00-12:30 (Pier 7&8)
Investigating whether pre-trained language models (LMs) can function as knowledge bases (KBs) has raised wide research interests recently. However, existing works focus on simple, triple-based, relational KBs, but omit more sophisticated, logic-based, conceptualised KBs such as OWL ontologies. To investigate an LM's knowledge of ontologies, we propose OntoLAMA, a set of inference-based probing tasks and datasets from ontology subsumption axioms involving both atomic and complex concepts. We conduct extensive experiments on ontologies of different domains and scales, and our results demonstrate that LMs encode relatively less background knowledge of Subsumption Inference (SI) than traditional Natural Language Inference (NLI) but can improve on SI significantly when a small number of samples are given. We will open-source our code and datasets.

ReCode: Robustness Evaluation of Code Generation Models
Shiqi Wang, Zheng Li, Haifeng Qian, Chenghao Yang, Zijian Wang, Mingyue Shang, Varun Kumar, Samson Tan, Baishakhi Ray, Parminder Bhatia, Ramesh Nallapati, Murali Krishna Ramanathan, Dan Roth and Bing Xiang   11:00-12:30 (Pier 7&8)
Code generation models have achieved impressive performance. However, they tend to be brittle as slight edits to a prompt could lead to very different generations; these robustness properties, critical for user experience when deployed in real-life applications, are not well understood. Most existing works on robustness in text or code tasks have focused on classification, while robustness in generation tasks is an uncharted area and to date there is no comprehensive benchmark for robustness in code generation. In this paper, we propose ReCode, a comprehensive robustness evaluation benchmark for code generation models. We customize over 30 transformations specifically for code on docstrings, function and variable names, code syntax, and code format. They are carefully designed to be natural in real-life coding practice, preserve the original semantic meaning, and thus provide multifaceted assessments of a model's robustness performance. With human annotators, we verified that over 90% of the perturbed prompts do not alter the semantic meaning of the original prompt. In addition, we define robustness metrics for code generation models considering the worst-case behavior under each type of perturbation, taking advantage of the fact that executing the generated code can serve as objective evaluation. We demonstrate ReCode on SOTA models using HumanEval, MBPP, as well as function completion tasks derived from them. Interesting observations include: better robustness for CodeGen over InCoder and GPT-J; models are most sensitive to syntax perturbations; more challenging robustness evaluation on MBPP over HumanEval.

Local Interpretation of Transformer Based on Linear Decomposition
Sen Yang, Shujian Huang, wei zou, Jianbing Zhang, Xinyu Dai and Jiajun CHEN   11:00-12:30 (Pier 7&8)
In recent years, deep neural networks (DNNs) have achieved state-of-the-art performance on a wide range of tasks. However, limitations in interpretability have hindered their applications in the real world. This work proposes to interpret neural networks by linear decomposition and finds that the ReLU-activated Transformer can be considered as a linear model on a single input. We further leverage the linearity of the model and propose a linear decomposition of the model output to generate local explanations. Our evaluation of sentiment classification and machine translation shows that our method achieves competitive performance in efficiency and fidelity of explanation. In addition, we demonstrate the potential of our approach in applications with examples of error analysis on multiple tasks.

Defending against Insertion-based Textual Backdoor Attacks via Attribution
Jiazhao Li, Zhuofeng Wu, Wei Ping, Chaowei Xiao and V.G.Vinod Vydiswaran 11:00-12:30 (Pier 7&8)
Textual backdoor attack, as a novel attack model, has been shown to be effective in adding a backdoor to the model during training. Defending against such backdoor attacks has become urgent and important. In this paper, we propose AttDef, an efficient attribution-based pipeline to defend against two insertion-based poisoning attacks, BadNL and InSent. Specifically, we regard the tokens with larger attribution scores as potential triggers since larger attribution words contribute more to the false prediction results and therefore are more likely to be poison triggers. Additionally, we further utilize an external pre-trained language model to distinguish whether input is poisoned or not. We show that our proposed method can generalize sufficiently well in two common attack scenarios (poisoning training data and testing data), which consistently improves previous methods. For instance, AttDef can successfully mitigate both attacks with an average accuracy of 79.97% (56.59% up) and 48.34% (3.99% up) under pre-training and post-training attack defense respectively, achieving the new state-of-the-art performance on prediction recovery over four benchmark datasets.

Feature Interactions Reveal Linguistic Structure in Language Models
Jaap Jumelet and Willem Zuidema 11:00-12:30 (Pier 7&8)
We study feature interactions in the context of feature attribution methods for post-hoc interpretability. In interpretability research, getting to grips with feature interactions is increasingly recognised as an important challenge, because interacting features are key to the success of neural networks. Feature interactions allow a model to build up hierarchical representations for its input, and might provide an ideal starting point for the investigation into linguistic structure in language models. However, uncovering the exact role that these interactions play is also difficult, and a diverse range of interaction attribution methods has been proposed. In this paper, we focus on the question which of these methods most faithfully reflects the inner workings of the target models. We work out a grey box methodology, in which we train models to perfection on a formal language classification task, using PCFGs. We show that under specific configurations, some methods are indeed able to uncover the grammatical rules acquired by a model. Based on these findings we extend our evaluation to a case study on language models, providing novel insights into the linguistic structure that these models have acquired.

HuaSLIM: Human Attention Motivated Shortcut Learning Identification and Mitigation for Large Language models
Yuqi Ren and Deyi Xiong 11:00-12:30 (Pier 7&8)
Large language models have made remarkable progress on a variety of NLP tasks. However, it has been found that they tend to rely on shortcut features that spuriously correlate with labels for prediction, which weakens their generalization on out-of-distribution samples. In this paper, we propose a human attention guided approach to identifying and mitigating shortcut learning, which encourages the LLM-based target model to learn relevant features. We define an attention-based measurement to capture both model and data bias and identify shortcut tokens by exploring both human and neural attention. In a self-distillation framework, we mitigate shortcut learning by dynamically adjusting the distillation temperature according to the detected shortcut tokens and estimated shortcut degree. Additionally, we utilize human attention as a supervisory signal to constrain large language models to pay more attention to relevant tokens. Experimental results on multiple NLP tasks show that our proposed method can effectively identify shortcut tokens, and significantly improve the robustness of large language models on OOD samples, while not undermining the performance on IID data.

Dynamic Transformers Provide a False Sense of Efficiency
Yiming Chen, Simin Chen, Zexin Li, Wei Yang, Cong Liu, Robby Tan and Haizhou Li 11:00-12:30 (Pier 7&8)
Despite much success in natural language processing (NLP), pre-trained language models typically lead to a high computational cost during inference. Multi-exit is a mainstream approach to address this issue by making a trade-off between efficiency and accuracy, where the saving of computation comes from an early exit. However, whether such saving from early-exiting is robust remains unknown. Motivated by this, we first show that directly adapting existing adversarial attack approaches targeting model accuracy cannot significantly reduce inference efficiency. To this end, we propose a simple yet effective attacking framework, SAME, a novel slowdown attack framework on multi-exit models, which is specially tailored to reduce the efficiency of the multi-exit models. By leveraging the multi-exit models' design characteristics, we utilize all internal predictions to guide the adversarial sample generation instead of merely considering the final prediction. Experiments on the GLUE benchmark show that SAME can effectively diminish the efficiency gain of various multi-exit models by 80% on average, convincingly validating its effectiveness and generalization ability.

Prompt Tuning Pushes Farther, Contrastive Learning Pulls Closer: A Two-Stage Approach to Mitigate Social Biases
Yingji Li, Mengnan Du, Xin Wang and Ying Wang 11:00-12:30 (Pier 7&8)
As the representation capability of Pre-trained Language Models (PLMs) improve, there is growing concern that they will inherit social biases from unprocessed corpora. Most previous debiasing techniques used Counterfactual Data Augmentation (CDA) to balance the training corpus. However, CDA slightly modifies the original corpus, limiting the representation distance between different demographic groups to a narrow range. As a result, the debiasing model easily fits the differences between counterfactual pairs, which affects its debiasing performance with limited text resources. In this paper, we propose an adversarial training-inspired two-stage debiasing model using Contrastive learning with Continuous Prompt Augmentation (named CCPA) to mitigate social biases in PLMs' encoding. In the first stage, we propose a data augmentation method based on continuous prompt tuning to push farther the representation distance between sample pairs along different demographic groups. In the second stage, we utilize contrastive learning to pull closer the representation distance between the augmented sample pairs and then fine-tune PLMs' parameters to get debiased encoding. Our approach guides the model to achieve stronger debiasing performance by adding difficulty to the training process. Extensive experiments show that CCPA outperforms baselines in terms of debiasing performance. Meanwhile, experimental results on the GLUE benchmark show that CCPA retains the language modeling capability of PLMs.

Disagreement Matters: Preserving Label Diversity by Jointly Modeling Item and Annotator Label Distributions with DisCo
Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj B Bhensadadia, Ashiqur KhudaBukhsh and Christopher Homan 11:00-12:30 (Pier 7&8)
Annotator disagreement is common whenever human judgment is needed for supervised learning. It is conventional to assume that one label per item represents ground truth. However, this obscures minority opinions, if present. We regard "ground truth" as the distribution of all labels that a population of annotators could produce, if asked (and of which we only have a small sample). We next introduce DisCo (Distribution from Context), a simple neural model that learns to predict this distribution. The model takes annotator-item pairs, rather than items alone, as input, and performs inference by aggregating over all annotators. Despite its simplicity, our experiments show that, on six benchmark datasets, our model is competitive with, and frequently outperforms, other, more complex models that either do not model specific annotators or were not designed for label distribution learning.

An Empirical Analysis of Parameter-Efficient Methods for Debiasing Pre-Trained Language Models
Zhongbin Xie and Thomas Lukasiewicz 11:00-12:30 (Pier 7&8)
The increasingly large size of modern pre-trained language models not only makes them inherit more human-like biases from the training corpora, but also makes it computationally expensive to mitigate such biases. In this paper, we investigate recent parameter-efficient methods in combination with counterfactual data augmentation (CDA) for bias mitigation. We conduct extensive experiments with prefix tuning, prompt

tuning, and adapter tuning on different language models and bias types to evaluate their debiasing performance and abilities to preserve the internal knowledge of a pre-trained model. We find that the parameter-efficient methods (i) are effective in mitigating gender bias, where adapter tuning is consistently the most effective one and prompt tuning is more suitable for GPT-2 than BERT, (ii) are less effective when it comes to racial and religious bias, which may be attributed to the limitations of CDA, and (iii) can perform similarly to or sometimes better than full fine-tuning with improved time and memory efficiency, as well as maintain the internal knowledge in BERT and GPT-2, evaluated via fact retrieval and downstream fine-tuning.

MIL-Decoding: Detoxifying Language Models at Token-Level via Multiple Instance Learning
Xu Zhang and Xiaojun Wan 11:00-12:30 (Pier 7&8)
Despite advances in large pre-trained neural language models, they are prone to generating toxic language, which brings security risks to their applications. We introduce MIL-Decoding, which detoxifies language models at token-level by interpolating it with a trained multiple instance learning (MIL) network. MIL model is trained on a corpus with a toxicity label for each text to predict the overall toxicity and the toxicity of each token in its context. Intuitively, the MIL network computes a toxicity distribution over next tokens according to the generated context which supplements the original language model to avoid toxicity. We evaluate MIL-Decoding with automatic metrics and human evaluation, where MIL-Decoding outperforms other baselines in detoxification while it only hurts generation fluency a little bit.

DITTO: Data-efficient and Fair Targeted Subset Selection for ASR Accent Adaptation
Suraj N Kothawade, Anmol Reddy Mekala, D.Chandra Sekhara SS Hetha Havya, Mayank Kothyari, Rishabh K Iyer, Ganesh Ramakrishnan and Preethi Jyothi 11:00-12:30 (Pier 7&8)
State-of-the-art Automatic Speech Recognition (ASR) systems are known to exhibit disparate performance on varying speech accents. To improve performance on a specific target accent, a commonly adopted solution is to finetune the ASR model using accent-specific labeled speech. However, acquiring large amounts of labeled speech for specific target accents is challenging. Choosing an informative subset of speech samples that are most representative of the target accents becomes important for effective ASR finetuning. To address this problem, we propose DITTO (Data-efficient and faIr Targeted subseT selectiOn that uses Submodular Mutual Information (SMI) functions as acquisition functions to find the most informative set of utterances matching a target accent within a fixed budget. An important feature of DITTO is that it supports fair targeting for multiple accents, i.e. it can automatically select representative data points from multiple accents when the ASR model needs to perform well on more than one accent. We show that compared to other speech selection methods, DITTO is 3-5 times as label-efficient for its improvements on the Indic-TTS and L2 datasets.

It Takes Two to Tango: Navigating Conceptualizations of NLP Tasks and Measurements of Performance
Arjun Subramonian, Xingdi Yuan, Hal Daumé III and Su Lin Blodgett 11:00-12:30 (Pier 7&8)
Progress in NLP is increasingly measured through benchmarks; hence, contextualizing progress requires understanding when and why practitioners may disagree about the validity of benchmarks. We develop a taxonomy of disagreement, drawing on tools from measurement modeling, and distinguish between two types of disagreement: 1) how tasks are conceptualized and 2) how measurements of model performance are operationalized. To provide evidence for our taxonomy, we conduct a meta-analysis of relevant literature to understand how NLP tasks are conceptualized, as well as a survey of practitioners about their impressions of different factors that affect benchmark validity. Our meta-analysis and survey across eight tasks, ranging from coreference resolution to question answering, uncover that tasks are generally not clearly and consistently conceptualized and benchmarks suffer from operationalization disagreements. These findings support our proposed taxonomy of disagreement. Finally, based on our taxonomy, we present a framework for constructing benchmarks and documenting their limitations.

Reimagining Retrieval Augmented Language Models for Answering Queries
Wang-Chiew Tan, Yuliang Li, Pedro Rodriguez, Richard James, Xi Victoria Lin, Alon Halevy and Wen-tau Yih 11:00-12:30 (Pier 7&8)
We present a reality check on large language models and inspect the promise of retrieval-augmented language models in comparison. Such language models are semi-parametric, where models integrate model parameters and knowledge from external data sources to make their predictions, as opposed to the parametric nature of vanilla large language models. We give initial experimental findings that semi-parametric architectures can be enhanced with views, a query analyzer/planner, and provenance to make a significantly more powerful system for question answering in terms of accuracy and efficiency, and potentially for other NLP tasks.

GLUE-X: Evaluating Natural Language Understanding Models from an Out-of-Distribution Generalization Perspective
Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie and Yue Zhang 11:00-12:30 (Pier 7&8)
Pre-trained language models (PLMs) are known to improve the generalization performance of natural language understanding models by leveraging large amounts of data during the pre-training phase. However, the out-of-distribution (OOD) generalization problem remains a challenge in many NLP tasks, limiting the real-world deployment of these methods. This paper presents the first attempt at creating a unified benchmark named GLUE-X for evaluating OOD robustness in NLP models, highlighting the importance of OOD robustness and providing insights on how to measure the robustness of a model and how to improve it. The benchmark includes 13 publicly available datasets for OOD testing, and evaluations are conducted on 8 classic NLP tasks over 21 popularly used PLMs. Our findings confirm the need for improved OOD accuracy in NLP tasks, as significant performance degradation was observed in all settings compared to in-distribution (ID) accuracy.

Revisiting Non-Autoregressive Translation at Scale
Zhihao Wang, Longyue Wang, Jinsong Su, Junfeng Yao and Zhaopeng Tu 11:00-12:30 (Pier 7&8)
In real-world systems, scaling has been critical for improving the translation quality in autoregressive translation (AT), which however has not been well studied for non-autoregressive translation (NAT). In this work, we bridge the gap by systematically studying the impact of scaling on NAT behaviors. Extensive experiments on six WMT benchmarks over two advanced NAT models show that scaling can alleviate the commonly-cited weaknesses of NAT models, resulting in better translation performance. To reduce the side-effect of scaling on decoding speed, we empirically investigate the impact of NAT encoder and decoder on the translation performance. Experimental results on the large-scale WMT20 En-De show that the asymmetric architecture (e.g. bigger encoder and smaller decoder) can achieve comparable performance with the scaling model, while maintaining the superiority of decoding speed with standard NAT models. To this end, we establish a new benchmark by validating scaled NAT models on the scaled dataset, which can be regarded as a strong baseline for future works. We release code and system outputs at https://github.com/DeepLearnXMU/Scaling4NAT.

Reproducibility in NLP: What Have We Learned from the Checklist?
Ian Magnusson, Noah A. Smith and Jesse Dodge 11:00-12:30 (Pier 7&8)
Scientific progress in NLP rests on the reproducibility of researchers' claims. The *CL conferences created the NLP Reproducibility Checklist in 2020 to be completed by authors at submission to remind them of key information to include. We provide the first analysis of the Checklist by examining 10,405 anonymous responses to it. First, we find evidence of an increase in reporting of information on efficiency, validation performance, summary statistics, and hyperparameters after the Checklist's introduction. Further, we show acceptance rate grows for submissions with more Yes responses. We find that the 44% of submissions that gather new data are 5% less likely to be accepted than those that did not; the average reviewer-rated reproducibility of these submissions is also 2% lower relative to the rest. We find that only 46% of

submissions claim to open-source their code, though submissions that do have 8% higher reproducibility score relative to those that do not, the most for any item. We discuss what can be inferred about the state of reproducibility in NLP, and provide a set of recommendations for future conferences, including: a) allowing submitting code and appendices one week after the deadline, and b) measuring dataset reproducibility by a checklist of data collection practices.

Towards Reasoning in Large Language Models: A Survey
Jie Huang and Kevin Chen-Chuan Chang   11:00-12:30 (Pier 7&8)
Reasoning is a fundamental aspect of human intelligence that plays a crucial role in activities such as problem solving, decision making, and critical thinking. In recent years, large language models (LLMs) have made significant progress in natural language processing, and there is observation that these models may exhibit reasoning abilities when they are sufficiently large. However, it is not yet clear to what extent LLMs are capable of reasoning. This paper provides a comprehensive overview of the current state of knowledge on reasoning in LLMs, including techniques for improving and eliciting reasoning in these models, methods and benchmarks for evaluating reasoning abilities, findings and implications of previous research in this field, and suggestions on future directions. Our aim is to provide a detailed and up-to-date review of this topic and stimulate meaningful discussion and future work.

Risks and NLP Design: A Case Study on Procedural Document QA
Nikita Haduong, Alice Gao and Noah A. Smith   11:00-12:30 (Pier 7&8)
As NLP systems are increasingly deployed at scale, concerns about their potential negative impacts have attracted the attention of the research community, yet discussions of risk have mostly been at an abstract level and focused on generic AI or NLP applications. We argue that clearer assessments of risks and harms to users—and concrete strategies to mitigate them—will be possible when we specialize the analysis to more concrete applications and their plausible users. As an illustration, this paper is grounded in cooking recipe procedural document question answering (ProcDocQA), where there are well-defined risks to users such as injuries or allergic reactions. Our case study shows that an existing language model, applied in "zero-shot" mode, quantitatively answers real-world questions about recipes as well or better than the humans who have answered the questions on the web. Using a novel questionnaire informed by theoretical work on AI risk, we conduct a risk-oriented error analysis that could then inform the design of a future system to be deployed with lower risk of harm and better performance.

Session 5 - 16:15-17:45

Interpretability and Analysis of Models for NLP

16:15-17:45 (Metropolitan East)

FLamE: Few-shot Learning from Natural Language Explanations
Yangqiaoyu Zhou, Yiming Zhang and Chenhao Tan   16:15-16:30 (Metropolitan East)
Natural language explanations have the potential to provide rich information that in principle guides model reasoning. Yet, recent work by Lampinen et al. has shown limited utility of natural language explanations in improving classification. To effectively learn from explanations, we present FLamE, a two-stage few-shot learning framework that first generates explanations using GPT-3, and then fine-tunes a smaller model (e.g., RoBERTa) with generated explanations. Our experiments on natural language inference demonstrate effectiveness over strong baselines, increasing accuracy by 17.6% over GPT-3 Babbage and 5.7% over GPT-3 Davinci in e-SNLI. Despite improving classification performance, human evaluation surprisingly reveals that the majority of generated explanations does not adequately justify classification decisions. Additional analyses point to the important role of label-specific cues (e.g., "not know" for the neutral label) in generated explanations.

MGR: Multi-generator Based Rationalization
Wei Liu, Haozhao Wang, Jun Wang, Ruixuan Li, Xinyang Li, YuanKai Zhang and Yang Qiu   16:30-16:45 (Metropolitan East)
Rationalization is to employ a generator and a predictor to construct a self-explaining NLP model in which the generator selects a subset of human-intelligible pieces of the input text to the following predictor. However, rationalization suffers from two key challenges, i.e., spurious correlation and degeneration, where the predictor overfits the spurious or meaningless pieces solely selected by the not-yet well-trained generator and in turn deteriorates the generator. Although many studies have been proposed to address the two challenges, they are usually designed separately and do not take both of them into account. In this paper, we propose a simple yet effective method named MGR to simultaneously solve the two problems. The key idea of MGR is to employ multiple generators such that the occurrence stability of real pieces is improved and more meaningful pieces are delivered to the predictor. Empirically, we show that MGR improves the F1 score by up to 20.9% as compared to state-of-the-art methods.

Efficient Shapley Values Estimation by Amortization for Text Classification
Chenghao Yang, Fan Yin, He He, Kai-Wei Chang, Xiaofei Ma and Bing Xiang   16:45-17:00 (Metropolitan East)
Despite the popularity of Shapley Values in explaining neural text classification models, computing them is prohibitive for large pretrained models due to a large number of model evaluations. In practice, Shapley Values are often estimated with a small number of stochastic model evaluations. However, we show that the estimated Shapley Values are sensitive to random seed choices – the top-ranked features often have little overlap across different seeds, especially on examples with longer input texts. This can only be mitigated by aggregating thousands of model evaluations, which on the other hand, induces substantial computational overheads. To mitigate the trade-off between stability and efficiency, we develop an amortized model that directly predicts each input feature's Shapley Value without additional model evaluations. It is trained on a set of examples whose Shapley Values are estimated from a large number of model evaluations to ensure stability. Experimental results on two text classification datasets demonstrate that our amortized model estimates Shapley Values accurately with up to 60 times speedup compared to traditional methods. Further, our model does not suffer from stability issues as inference is deterministic. We release our code at https://github.com/yangalan123/Amortized-Interpretability.

KNOW How to Make Up Your Mind! Adversarially Detecting and Alleviating Inconsistencies in Natural Language Explanations
Myeongjun Jang, Bodhisattwa Prasad Majumder, Julian McAuley, Thomas Lukasiewicz and Oana-Maria Camburu   17:00-17:15 (Metropolitan East)
While recent works have been considerably improving the quality of the natural language explanations (NLEs) generated by a model to justify its predictions, there is very limited research in detecting and alleviating inconsistencies among generated NLEs. In this work, we leverage external knowledge bases to significantly improve on an existing adversarial attack for detecting inconsistent NLEs. We apply our attack to high-performing NLE models and show that models with higher NLE quality do not necessarily generate fewer inconsistencies. Moreover, we propose an off-the-shelf mitigation method to alleviate inconsistencies by grounding the model into external background knowledge. Our method decreases the inconsistencies of previous high-performing NLE models as detected by our attack.

Information Extraction

16:15-17:45 (Metropolitan Centre)

TAGPRIME: A Unified Framework for Relational Structure Extraction

I-Hung Hsu, Kuan-Hao Huang, Shuning Zhang, Wenxin Cheng, Prem Natarajan, Kai-Wei Chang and Nanyun Peng   16:15-16:30
(Metropolitan Centre)

Many tasks in natural language processing require the extraction of relationship information for a given condition, such as event argument extraction, relation extraction, and task-oriented semantic parsing. Recent works usually propose sophisticated models for each task independently and pay less attention to the commonality of these tasks and to have a unified framework for all the tasks. In this work, we propose to take a unified view of all these tasks and introduce TAGPRIME to address relational structure extraction problems. TAGPRIME is a sequence tagging model that appends priming words about the information of the given condition (such as an event trigger) to the input text. With the self-attention mechanism in pre-trained language models, the priming words make the output contextualized representations contain more information about the given condition, and hence become more suitable for extracting specific relationships for the condition. Extensive experiments and analyses on three different tasks that cover ten datasets across five different languages demonstrate the generality and effectiveness of TAGPRIME.

Linguistic representations for fewer-shot relation extraction across domains

Sireesh Gururaja, Ritam Dutt, Tinglong Liao and Carolyn Rosé   16:30-16:45 (Metropolitan Centre)

Recent work has demonstrated the positive impact of incorporating linguistic representations as additional context and scaffolds on the in-domain performance of several NLP tasks. We extend this work by exploring the impact of linguistic representations on cross-domain performance in a few-shot transfer setting. An important question is whether linguistic representations enhance generalizability by providing features that function as cross-domain pivots. We focus on the task of relation extraction on three datasets of procedural text in two domains, cooking and materials science. Our approach augments a popular transformer-based architecture by alternately incorporating syntactic and semantic graphs constructed by freely available off-the-shelf tools. We examine their utility for enhancing generalization, and investigate whether earlier findings, e.g. that semantic representations can be more helpful than syntactic ones, extend to relation extraction in multiple domains. We find that while the inclusion of these graphs results in significantly higher performance in few-shot transfer, both types of graph exhibit roughly equivalent utility.

Learning Dynamic Contextualised Word Embeddings via Template-based Temporal Adaptation

Xiaohang Tang, Yi Zhou and Danushka Bollegala   16:45-17:00 (Metropolitan Centre)

Dynamic contextualised word embeddings (DCWEs) represent the temporal semantic variations of words. We propose a method for learning DCWEs by time-adapting a pretrained Masked Language Model (MLM) using time-sensitive templates. Given two snapshots C_1 and C_2 of a corpus taken respectively at two distinct timestamps T_1 and T_2, we first propose an unsupervised method to select (a) *pivot* terms related to both C_1 and C_2, and (b) *anchor* terms that are associated with a specific pivot term in each individual snapshot. We then generate prompts by filling manually compiled templates using the extracted pivot and anchor terms. Moreover, we propose an automatic method to learn time-sensitive templates from C_1 and C_2, without requiring any human supervision. Next, we use the generated prompts to adapt a pretrained MLM to T_2 by fine-tuning using those prompts. Multiple experiments show that our proposed method significantly reduces the perplexity of test sentences in C_2, outperforming the current state-of-the-art.

MultiTACRED: A Multilingual Version of the TAC Relation Extraction Dataset

Leonhard Hennig, Philippe Thomas and Sebastian Möller   17:00-17:15 (Metropolitan Centre)

Relation extraction (RE) is a fundamental task in information extraction, whose extension to multilingual settings has been hindered by the lack of supervised resources comparable in size to large English datasets such as TACRED (Zhang et al., 2017). To address this gap, we introduce the MultiTACRED dataset, covering 12 typologically diverse languages from 9 language families, which is created by machine-translating TACRED instances and automatically projecting their entity annotations. We analyze translation and annotation projection quality, identify error categories, and experimentally evaluate fine-tuned pretrained mono- and multilingual language models in common transfer learning scenarios. Our analyses show that machine translation is a viable strategy to transfer RE instances, with native speakers judging more than 83% of the translated instances to be linguistically and semantically acceptable. We find monolingual RE model performance to be comparable to the English original for many of the target languages, and that multilingual models trained on a combination of English and target language data can outperform their monolingual counterparts. However, we also observe a variety of translation and annotation projection errors, both due to the MT systems and linguistic features of the target languages, such as pronoun-dropping, compounding and inflection, that degrade dataset quality and RE model performance.

Can NLI Provide Proper Indirect Supervision for Low-resource Biomedical Relation Extraction?

Jiashu Xu, Mingyu Derek Ma and Muhao Chen   17:15-17:30 (Metropolitan Centre)

Two key obstacles in biomedical relation extraction (RE) are the scarcity of annotations and the prevalence of instances without explicitly pre-defined labels due to low annotation coverage. Existing approaches, which treat biomedical RE as a multi-class classification task, often result in poor generalization in low-resource settings and do not have the ability to make selective prediction on unknown cases but give a guess from seen relations, hindering the applicability of those approaches. We present NBR, which converts biomedical RE as natural language inference formulation through indirect supervision. By converting relations to natural language hypotheses, NBR is capable of exploiting semantic cues to alleviate annotation scarcity. By incorporating a ranking-based loss that implicitly calibrates abstinent instances, NBR learns a clearer decision boundary and is instructed to abstain on uncertain instances. Extensive experiments on three widely-used biomedical RE benchmarks, namely ChemProt, DDI and GAD, verify the effectiveness of NBR in both full-set and low-resource regimes. Our analysis demonstrates that indirect supervision benefits biomedical RE even when a domain gap exists, and combining NLI knowledge with biomedical knowledge leads to the best performance gains.

Open Set Relation Extraction via Unknown-Aware Training

Jun Zhao, Xin Zhao, WenYu Zhan, Qi Zhang, Tao Gui, Zhongyu Wei, Yun Wen Chen, Xiang Gao and Xuanjing Huang   17:30-17:45
(Metropolitan Centre)

The existing supervised relation extraction methods have achieved impressive performance in a closed-set setting, in which the relations remain the same during both training and testing. In a more realistic open-set setting, unknown relations may appear in the test set. Due to the lack of supervision signals from unknown relations, a well-performing closed-set relation extractor can still confidently misclassify them into known relations. In this paper, we propose an unknown-aware training method, regularizing the model by dynamically synthesizing negative instances that can provide the missing supervision signals. Inspired by text adversarial attack, We adaptively apply small but critical

perturbations to original training data, synthesizing **difficult enough** negative instances that are mistaken by the model as known relations, thus facilitating a compact decision boundary. Experimental results show that our method achieves SOTA unknown relation detection without compromising the classification of known relations.

Generation

16:15-17:45 (Metropolitan West)

[CL] Neural Data-to-Text Generation Based on Small Datasets: Comparing the Added Value of Two Semi-Supervised Learning Approaches on Top of a Large Language Model
Chris Lee, Thiago Ferreira, Chris Emmery, Travis Wiltshire and Emiel Krahmer   16:15-16:30 (Metropolitan West)
This study discusses the effect of semi-supervised learning in combination with pretrained language models for data-to-text generation. It is not known whether semi-supervised learning is still helpful when a large-scale language model is also supplemented. This study aims to answer this question by comparing a data-to-text system only supplemented with a language model, to two data-to-text systems that are additionally enriched by a data augmentation or a pseudo-labeling semi-supervised learning approach. Results show that semi-supervised learning results in higher scores on diversity metrics. In terms of output quality, extending the training set of a data-to-text system with a language model using the pseudo-labeling approach did increase text quality scores, but the data augmentation approach yielded similar scores to the system without training set extension. These results indicate that semi-supervised learning approaches can bolster output quality and diversity, even when a language model is also present.

[TACL] Conditional Generation with a Question-Answering Blueprint
Shashi Narayan, Shashi Narayan , Joshua Maynez , Reinald Kim Amplayo , Kuzman Ganchev, Annie Louis, Fantine Huot , Anders Sandholm , Dipanjan Das and Mirella Lapata   16:30-16:45 (Metropolitan West)
The ability to convey relevant and faithful information is critical for many tasks in conditional generation and yet remains elusive for neural seq-to-seq models whose outputs often reveal hallucinations and fail to correctly cover important details. In this work, we advocate planning as a useful intermediate representation for rendering conditional generation less opaque and more grounded. We propose a new conceptualization of text plans as a sequence of question-answer (QA) pairs and enhance existing datasets (e.g., for summarization) with a QA blueprint operating as a proxy for content selection (i.e., what to say) and planning (i.e., in what order). We obtain blueprints automatically by exploiting state-of-the-art question generation technology and convert input-output pairs into input-blueprint-output tuples. We develop Transformer-based models, each varying in how they incorporate the blueprint in the generated output (e.g., as a global plan or iteratively). Evaluation across metrics and datasets demonstrates that blueprint models are more factual than alternatives which do not resort to planning and allow tighter control of the generation output.

HAUSER: Towards Holistic and Automatic Evaluation of Simile Generation
Qianyu He, Yikai Zhang, Jiaqing Liang, Yuncheng Huang, Yanghua Xiao and Yunwen Chen   16:45-17:00 (Metropolitan West)
Similes play an imperative role in creative writing such as story and dialogue generation. Proper evaluation metrics are like a beacon guiding the research of simile generation (SG). However, it remains under-explored as to what criteria should be considered, how to quantify each criterion into metrics, and whether the metrics are effective for comprehensive, efficient, and reliable SG evaluation. To address the issues, we establish HAUSER, a holistic and automatic evaluation system for the SG task, which consists of five criteria from three perspectives and automatic metrics for each criterion. Through extensive experiments, we verify that our metrics are significantly more correlated with human ratings from each perspective compared with prior automatic metrics. Resources of HAUSER are publicly available at https://github.com/Abbey4799/HAUSER.

ByGPT5: End-to-End Style-conditioned Poetry Generation with Token-free Language Models
Jonas Belouadi and Steffen Eger   17:00-17:15 (Metropolitan West)
State-of-the-art poetry generation systems are often complex. They either consist of task-specific model pipelines, incorporate prior knowledge in the form of manually created constraints, or both. In contrast, end-to-end models would not suffer from the overhead of having to model prior knowledge and could learn the nuances of poetry from data alone, reducing the degree of human supervision required. In this work, we investigate end-to-end poetry generation conditioned on styles such as rhyme, meter, and alliteration. We identify and address lack of training data and mismatching tokenization algorithms as possible limitations of past attempts. In particular, we successfully pre-train ByGPT5, a new token-free decoder-only language model, and fine-tune it on a large custom corpus of English and German quatrains annotated with our styles. We show that ByGPT5 outperforms other models such as mT5, ByT5, GPT-2 and ChatGPT, while also being more parameter efficient and performing favorably compared to humans. In addition, we analyze its runtime performance and demonstrate that it is not prone to memorization. We make our code, models, and datasets publicly available.

Are Experts Needed? On Human Evaluation of Counselling Reflection Generation
Zixiu Wu, Simone Balloccu, Ehud Reiter, Rim Helaoui, Diego Reforgiato Recupero and Daniele Riboni   17:15-17:30 (Metropolitan West)
Reflection is a crucial counselling skill where the therapist conveys to the client their interpretation of what the client said. Language models have recently been used to generate reflections automatically, but human evaluation is challenging, particularly due to the cost of hiring experts. Laypeople-based evaluation is less expensive and easier to scale, but its quality is unknown for reflections. Therefore, we explore whether laypeople can be an alternative to experts in evaluating a fundamental quality aspect: coherence and context-consistency. We do so by asking a group of laypeople and a group of experts to annotate both synthetic reflections and human reflections from actual therapists. We find that both laypeople and experts are reliable annotators and that they have moderate-to-strong inter-group correlation, which shows that laypeople can be trusted for such evaluations. We also discover that GPT-3 mostly produces coherent and consistent reflections, and we explore changes in evaluation results when the source of synthetic reflections changes to GPT-3 from the less powerful GPT-2.

Increasing Diversity While Maintaining Accuracy: Text Data Generation with Large Language Models and Human Interventions
John Joon Young Chung, Ece Kamar and Saleema Amershi   17:30-17:45 (Metropolitan West)
Large language models (LLMs) can be used to generate text data for training and evaluating other models. However, creating high-quality datasets with LLMs can be challenging. In this work, we explore human-AI partnerships to facilitate high diversity and accuracy in LLM-based text data generation. We first examine two approaches to diversify text generation: 1) logit suppression, which minimizes the generation of languages that have already been frequently generated, and 2) temperature sampling, which flattens the token sampling probability. We found that diversification approaches can increase data diversity but often at the cost of data accuracy (i.e., text and labels being appropriate for the target domain). To address this issue, we examined two human interventions, 1) label replacement (LR), correcting misaligned labels, and 2) out-of-scope filtering (OOSF), removing instances that are out of the user's domain of interest or to which no considered label applies. With oracle studies, we found that LR increases the absolute accuracy of models trained with diversified datasets by 14.4%. Moreover, we

found that some models trained with data generated with LR interventions outperformed LLM-based few-shot classification. In contrast, OOSF was not effective in increasing model accuracy, implying the need for future work in human-in-the-loop text data generation.

Posters

16:15-17:45 (Frontenac Ballroom and Queen's Quay)

[Industry] Federated Learning of Gboard Language Models with Differential Privacy

Zheng Xu, Yanxiang Zhang, Galen Andrew, Christopher Choquette, Peter Kairouz, Brendan Mcmahan, Jesse Rosenstock and Yuanbo Zhang
16:15-17:45 (Frontenac Ballroom and Queen's Quay)
We train and deploy language models (LMs) with federated learning (FL) and differential privacy (DP) in Google Keyboard (Gboard). The recent DP-Follow the Regularized Leader (DP-FTRL) algorithm is applied to achieve meaningfully formal DP guarantees without requiring uniform sampling of clients. To provide favorable privacy-utility trade-offs, we introduce a new client participation criterion and discuss the implication of its configuration in large scale systems. We show how quantile-based clip estimation can be combined with DP-FTRL to adaptively choose the clip norm during training or reduce the hyperparameter tuning in preparation of training. With the help of pretraining on public data, we trained and deployed more than fifteen Gboard LMs that achieve high utility and ρ-zCDP privacy guarantees with $\rho \in (0.3, 2)$, with one model additionally trained with secure aggregation. We summarize our experience and provide concrete suggestions on DP training for practitioners.

[Industry] KG-FLIP: Knowledge-guided Fashion-domain Language-Image Pre-training for E-commerce

Qinjin Jia, Yang Liu, Daoping Wu, Shaoyuan Xu, Huidong Liu, Jinmiao Fu, Roland Vollgraf and Bryan Wang   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Various Vision-Language Pre-training (VLP) models (e.g., CLIP, BLIP) have sprung up and dramatically advanced the benchmarks for public general-domain datasets (e.g., COCO, Flickr30k). Such models usually learn the cross-modal alignment from large-scale well-aligned image-text datasets without leveraging external knowledge. Adapting these models to downstream applications in specific domains like fashion requires fine-grained in-domain image-text corpus, which are usually less semantically aligned and in small scale that requires efficient pre-training strategies. In this paper, we propose a knowledge-guided fashion-domain language-image pre-training (FLIP) framework that focuses on learning fine-grained representations in e-commerce domain and utilizes external knowledge (i.e., product attribute schema), to improve the pre-training efficiency. Experiments demonstrate that FLIP outperforms previous state-of-the-art VLP models on Amazon data and on the Fashion-Gen dataset by large margins. FLIP has been successfully deployed in the Amazon catalog system to backfill missing attributes and improve the customer shopping experience.

[Industry] AVEN-GR: Attribute Value Extraction and Normalization using product GRaphs

Thomas Ricatte and Donato Crisostomi   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Getting a good understanding of the user intent is vital for e-commerce applications to surface the right product to a given customer query. Query Understanding (QU) systems are essential for this purpose, and many e-commerce providers are working on complex solutions that need to be data efficient and able to capture early emerging market trends. Query Attribute Understanding (QAU) is a sub-component of QU that involves extracting named attributes from user queries and linking them to existing e-commerce entities such as brand, material, color, etc. While extracting named entities from text has been extensively explored in the literature, QAU requires specific attention due to the nature of the queries, which are often short, noisy, ambiguous, and constantly evolving. This paper makes three contributions to QAU. First, we propose a novel end-to-end approach that jointly solves Named Entity Recognition (NER) and Entity Linking (NEL) and enables open-world reasoning for QAU. Second, we introduce a novel method for utilizing product graphs to enhance the representation of query entities. Finally, we present a new dataset constructed from public sources that can be used to evaluate the performance of future QAU systems.

[Industry] PLAtE: A Large-scale Dataset for List Page Web Extraction

Aidan San, Yuan Zhuang, Jan Bakus, Colin Lockard, David Ciemiewicz, Sandeep Atluri, Kevin Small, Yangfeng Ji and Heba Elfardy   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Recently, neural models have been leveraged to significantly improve the performance of information extraction from semi-structured websites. However, a barrier for continued progress is the small number of datasets large enough to train these models. In this work, we introduce the PLAtE (Pages of Lists Attribute Extraction) benchmark dataset as a challenging new web extraction task. PLAtE focuses on shopping data, specifically extractions from product review pages with multiple items encompassing the tasks of: (1) finding product list segmentation boundaries and (2) extracting attributes for each product. PLAtE is composed of 52,898 items collected from 6,694 pages and 156,014 attributes, making it the first large-scale list page web extraction dataset. We use a multi-stage approach to collect and annotate the dataset and adapt three state-of-the-art web extraction models to the two tasks comparing their strengths and weaknesses both quantitatively and qualitatively.

[Industry] Domain-Agnostic Neural Architecture for Class Incremental Continual Learning in Document Processing Platform

Mateusz Wójcik, Witold Kościukiewicz, Mateusz Baran, Tomasz Kajdanowicz and Adam Gonczarek   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Production deployments in complex systems require ML architectures to be highly efficient and usable against multiple tasks. Particularly demanding are classification problems in which data arrives in a streaming fashion and each class is presented separately. Recent methods with stochastic gradient learning have been shown to struggle in such setups or have limitations like memory buffers, and being restricted to specific domains that disable its usage in real-world scenarios. For this reason, we present a fully differentiable architecture based on the Mixture of Experts model, that enables the training of high-performance classifiers when examples from each class are presented separately. We conducted exhaustive experiments that proved its applicability in various domains and ability to learn online in production environments. The proposed technique achieves SOTA results without a memory buffer and clearly outperforms the reference methods.

[Industry] Entity Contrastive Learning in a Large-Scale Virtual Assistant System

Jonathan Rubin, Jason Crowley, George Leung, Morteza Ziyadi and Maria Minakova 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Conversational agents are typically made up of domain (DC) and intent classifiers (IC) that identify the general subject an utterance belongs to and the specific action a user wishes to achieve. In addition, named entity recognition (NER) performs per token labeling to identify specific entities of interest in a spoken utterance. We investigate improving joint IC and NER models using entity contrastive learning that attempts to cluster similar entities together in a learned representation space. We compare a full virtual assistant system trained using entity contrastive learning to a production baseline system that does not use contrastive learning. We present both offline results, using retrospective test sets, as well as live online results from an A/B test that compared the two systems. In both the offline and online settings, entity contrastive training improved overall performance against production baselines. Furthermore, we provide a detailed analysis of learned entity embeddings, in-

cluding both qualitative analysis via dimensionality-reduced visualizations and quantitative analysis by computing alignment and uniformity metrics. We show that entity contrastive learning improves alignment metrics and produces well-formed embedding clusters in representation space.

[Industry] An efficient method for Natural Language Querying on Structured Data
Hanoz Bhathena, Aviral Joshi and Prateek Singh   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
We present an efficient and reliable approach to Natural Language Querying (NLQ) on databases (DB) which is not based on text-to-SQL type semantic parsing. Our approach simplifies the NLQ on structured data problem to the following "bread and butter" NLP tasks: (a) Domain classification, for choosing which DB table to query, whether the question is out-of-scope (b) Multi-head slot/entity extraction (SE) to extract the field criteria and other attributes such as its role (filter, sort etc) from the raw text and (c) Slot value disambiguation (SVD) to resolve/normalize raw spans from SE to format suitable to query a DB. This is a general purpose, DB language agnostic approach and the output can be used to query any DB and return results to the user. Also each of these tasks is extremely well studied, mature, easier to collect data for and enables better error analysis by tracing problems to specific components when something goes wrong.

[Industry] Evaluating Embedding APIs for Information Retrieval
Ehsan Kamalloo, Xinyu Zhang, Odunayo Ogundepo, Nandan Thakur, David Alfonso-hermelo, Mehdi Rezagholizadeh and Jimmy Lin 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
The ever-increasing size of language models curtails their widespread access to the community, thereby galvanizing many companies and startups into offering access to large language models through APIs. One particular API, suitable for dense retrieval, is the semantic embedding API that builds vector representations of a given text. With a growing number of APIs at our disposal, in this paper, our goal is to analyze semantic embedding APIs in realistic retrieval scenarios in order to assist practitioners and researchers in finding suitable services according to their needs. Specifically, we wish to investigate the capabilities of existing APIs on domain generalization and multilingual retrieval. For this purpose, we evaluate the embedding APIs on two standard benchmarks, BEIR, and MIRACL. We find that re-ranking BM25 results using the APIs is a budget-friendly approach and is most effective on English, in contrast to the standard practice, i.e., employing them as first-stage retrievers. For non-English retrieval, re-ranking still improves the results, but a hybrid model with BM25 works best albeit at a higher cost. We hope our work lays the groundwork for thoroughly evaluating APIs that are critical in search and more broadly, in information retrieval.

[Industry] AI Coach Assist: An Automated Approach for Call Recommendation in Contact Centers for Agent Coaching
Md Tahmid Rahman Laskar, Cheng Chen, Xue-yong Fu, Mahsa Azizi, Shashi Bhushan and Simon Corston-oliver   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
In recent years, the utilization of Artificial Intelligence (AI) in the contact center industry is on the rise. One area where AI can have a significant impact is in the coaching of contact center agents. By analyzing call transcripts, AI can quickly determine which calls are most relevant for coaching purposes, and provide relevant feedback and insights to the contact center manager or supervisor. In this paper, we present "AI Coach Assis", which leverages the pre-trained transformer-based language models to determine whether a given call is coachable or not based on the quality assurance (QA) queries/questions asked by the contact center managers or supervisors. The system was trained and evaluated on a large dataset collected from real-world contact centers and provides an efficient and effective way to determine which calls are most relevant for coaching purposes. Extensive experimental evaluation demonstrates the potential of AI Coach Assist to improve the coaching process, resulting in enhancing the performance of contact center agents.

[Industry] A Static Evaluation of Code Completion by Large Language Models
Hantian Ding, Varun Kumar, Yuchen Tian, Zijian Wang, Rob Kwiatkowski, Xiaopeng Li, Murali Krishna Ramanathan, Baishakhi Ray, Parminder Bhatia and Sudipta Sengupta   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Large language models trained on code have shown great potential to increase productivity of software developers. Several execution-based benchmarks have been proposed to evaluate functional correctness of model-generated code on simple programming problems. Nevertheless, it is expensive to perform the same evaluation on complex real-world projects considering the execution cost. On the other hand, static analysis tools such as linters, which can detect errors without running the program, haven't been well explored for evaluating code generation models. In this work, we propose a static evaluation framework to quantify static errors in Python code completions, by leveraging Abstract Syntax Trees. Compared with execution-based evaluation, our method is not only more efficient, but also applicable to code in the wild. For experiments, we collect code context from open source repos to generate one million function bodies using public models. Our static analysis reveals that Undefined Name and Unused Variable are the most common errors among others made by language models. Through extensive studies, we also show the impact of sampling temperature, model size, and context on static errors in code completions.

[Industry] Predicting Customer Satisfaction with Soft Labels for Ordinal Classification
Etienne Manderscheid and Matthias Lee   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
In a typical call center, only up to 8% of callers leave a Customer Satisfaction (CSAT) survey response at the end of the call, and these tend to be customers with strongly positive or negative experiences. To manage this data sparsity and response bias, we outline a predictive CSAT deep learning algorithm that infers CSAT on the 1-5 scale on inbound calls to the call center with minimal latency. The key metric to maximize is the precision for CSAT = 1 (lowest CSAT). We maximize this metric in two ways. First, reframing the problem as a binary class, rather than five-class problem during model fine-tuning, and then mapping binary outcomes back to five classes using temperature-scaled model probabilities. Second, using soft labels to represent the classes. The result is a production model able to support key customer workflows with high accuracy over millions of calls a month.

[Industry] Application-Agnostic Language Modeling for On-Device ASR
Markus Nussbaum-thom, Lyan Verwimp and Youssef Oualil   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
On-device automatic speech recognition systems face several challenges compared to server-based systems. They have to meet stricter constraints in terms of speed, disk size and memory while maintaining the same accuracy. Often they have to serve several ap- plications with different distributions at once, such as communicating with a virtual assistant and speech-to-text. The simplest solution to serve multiple applications is to build application-specific (language) models, but this leads to an increase in memory. Therefore, we explore different data- and architecture-driven language modeling approaches to build a single application-agnostic model. We propose two novel feed-forward architectures that find an optimal trade off between different on-device constraints. In comparison to the application-specific solution, one of our novel approaches reduces the disk size by half, while maintaining speed and accuracy of the original model.

[Industry] Semantic Ambiguity Detection in Sentence Classification using Task-Specific Embeddings
Jong Myoung Kim, Young-jun Lee, Sangkeun Jung and Ho-jin Choi   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Ambiguity is a major obstacle to providing services based on sentence classification. However, because of the structural limitations of the service, there may not be sufficient contextual information to resolve the ambiguity. In this situation, we focus on ambiguity detection so that service design considering ambiguity is possible. We utilize similarity in a semantic space to detect ambiguity in service scenarios and training data. In addition, we apply task-specific embedding to improve performance. Our results demonstrate that ambiguities and resulting labeling errors in training data or scenarios can be detected. Additionally, we confirm that it can be used to debug services

[Industry] What, When, and How to Ground: Designing User Persona-Aware Conversational Agents for Engaging Dialogue
Deuksin Kwon, Sunwoo Lee, Ki Hyun Kim, Seojin Lee, Taeyoon Kim and Eric Davis   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
This paper presents a method for building a personalized open-domain dialogue system to address the WWH (WHAT, WHEN, and HOW) problem for natural response generation in a commercial setting, where personalized dialogue responses are heavily interleaved with casual response turns. The proposed approach involves weighted dataset blending, negative persona information augmentation methods, and the design of personalized conversation datasets to address the challenges of WWH in personalized, open-domain dialogue systems. Our work effectively balances dialogue fluency and tendency to ground, while also introducing a response-type label to improve the controllability and explainability of the grounded response. The combination of these methods leads to more fluent conversations, as evidenced by subjective human evaluations as well as objective evaluations.

[Industry] Reducing cohort bias in natural language understanding systems with targeted self-training scheme
Dieu-thu Le, Gabriela Hernandez, Bei Chen and Melanie Bradford   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Bias in machine learning models can be an issue when the models are trained on particular types of data that do not generalize well, causing under performance in certain groups of users. In this work, we focus on reducing the bias related to new customers in a digital voice assistant system. It is observed that natural language understanding models often have lower performance when dealing with requests coming from new users rather than experienced users. To mitigate this problem, we propose a framework that consists of two phases (1) a fixing phase with four active learning strategies used to identify important samples coming from new users, and (2) a self training phase where a teacher model trained from the first phase is used to annotate semi-supervised samples to expand the training data with relevant cohort utterances. We explain practical strategies that involve an identification of representative cohort-based samples through density clustering as well as employing implicit customer feedbacks to improve new customers' experience. We demonstrate the effectiveness of our approach in a real world large scale voice assistant system for two languages, German and French through both offline experiments as well as A/B testings.

[Industry] KoSBI: A Dataset for Mitigating Social Bias Risks Towards Safer Large Language Model Applications
Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim and Jung-woo Ha   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Large language models (LLMs) not only learn natural text generation abilities but also social biases against different demographic groups from real-world data. This poses a critical risk when deploying LLM-based applications. Existing research and resources are not readily applicable in South Korea due to the differences in language and culture, both of which significantly affect the biases and targeted demographic groups. This limitation requires localized social bias datasets to ensure the safe and effective deployment of LLMs. To this end, we present KosBi, a new social bias dataset of 34k pairs of contexts and sentences in Korean covering 72 demographic groups in 15 categories. We find that through filtering-based moderation, social biases in generated content can be reduced by 16.47%p on average for HyperClova (30B and 82B), and GPT-3.

[Industry] CWSeg: An Efficient and General Approach to Chinese Word Segmentation
Dedong Li, Rui Zhao and Fei Tan   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
In this work, we report our efforts in advancing Chinese Word Segmentation for the purpose of rapid deployment in different applications. The pre-trained language model (PLM) based segmentation methods have achieved state-of-the-art (SOTA) performance, whereas this paradigm also poses challenges in the deployment. It includes the balance between performance and cost, segmentation ambiguity due to domain diversity and vague words boundary, and multi-grained segmentation. In this context, we propose a simple yet effective approach, namely CWSeg, to augment PLM-based schemes by developing cohort training and versatile decoding strategies. Extensive experiments on benchmark datasets demonstrate the efficiency and generalization of our approach. The corresponding segmentation system is also implemented for practical usage and the demo is recorded.

[Industry] Extracting Text Representations for Terms and Phrases in Technical Domains
Francesco Fusco and Diego Antognini   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Extracting dense representations for terms and phrases is a task of great importance for knowledge discovery platforms targeting highly-technical fields. Dense representations are used as features for downstream components and have multiple applications ranging from ranking results in search to summarization. Common approaches to create dense representations include training domain-specific embeddings with self-supervised setups or using sentence encoder models trained over similarity tasks. In contrast to static embeddings, sentence encoders do not suffer from the out-of-vocabulary (OOV) problem, but impose significant computational costs. In this paper, we propose a fully unsupervised approach to text encoding that consists of training small character-based models with the objective of reconstructing large pre-trained embedding matrices. Models trained with this approach can not only match the quality of sentence encoders in technical domains, but are 5 times smaller and up to 10 times faster, even on high-end GPUs.

[Industry] Weighted Contrastive Learning With False Negative Control to Help Long-tailed Product Classification
Tianqi Wang, Lei Chen, Xiaodan Zhu, Younghun Lee and Jing Gao   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Item categorization (IC) aims to classify product descriptions into leaf nodes in a categorical taxonomy, which is a key technology used in a wide range of applications. Along with the fact that most datasets often has a long-tailed distribution, classification performances on tail labels tend to be poor due to scarce supervision, causing many issues in real-life applications. To address IC task's long-tail issue, K-positive contrastive loss (KCL) is proposed on image classification task and can be applied on the IC task when using text-based contrastive learning, e.g., SimCSE. However, one shortcoming of using KCL has been neglected in previous research: false negative (FN) instances may harm the KCL's representation learning. To address the FN issue in the KCL, we proposed to re-weight the positive pairs in the KCL loss with a regularization that the sum of weights should be constrained to K+1 as close as possible. After controlling FN instances with the proposed method, IC performance has been further improved and is superior to other LT-addressing methods.

[Industry] RadLing: Towards Efficient Radiology Report Understanding
Rikhiya Ghosh, Oladimeji Farri, Sanjeev Kumar Karn, Manuela Danu, Ramya Vunikili and Larisa Micu   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Most natural language tasks in the radiology domain use language models pre-trained on biomedical corpus. There are few pretrained language models trained specifically for radiology, and fewer still that have been trained in a low data setting and gone on to produce comparable results in fine-tuning tasks. We present RadLing, a continuously pretrained language model using ELECTRA-small architecture, trained using over 500K radiology reports that can compete with state-of-the-art results for fine tuning tasks in radiology domain. Our main contribution in this paper is knowledge-aware masking which is an taxonomic knowledge-assisted pre-training task that dynamically masks tokens to inject knowledge during pretraining. In addition, we also introduce an knowledge base-aided vocabulary extension to adapt the general tokenization vocabulary to radiology domain.

[Industry] NAG-NER: a Unified Non-Autoregressive Generation Framework for Various NER Tasks
Xinpeng Zhang, Ming Tan, Jingfan Zhang and Wei Zhu   16:15-17:45 (Frontenac Ballroom and Queen's Quay)

Recently, the recognition of flat, nested, and discontinuous entities by a unified generative model framework has received increasing attention both in the research field and industry. However, the current generative NER methods force the entities to be generated in a predefined order, suffering from error propagation and inefficient decoding. In this work, we propose a unified non-autoregressive generation (NAG) framework for general NER tasks, referred to as NAG-NER. First, we propose to generate entities as a set instead of a sequence, avoiding error propagation. Second, we propose incorporating NAG in NER tasks for efficient decoding by treating each entity as a target sequence. Third, to enhance the generation performances of the NAG decoder, we employ the NAG encoder to detect potential entity mentions. Extensive experiments show that our NAG-NER model outperforms the state-of-the-art generative NER models on three benchmark NER datasets of different types and two of our proprietary NER tasks.\footnote{Code will be publicly available to the research community upon acceptance.}

[Industry] CUPID: Curriculum Learning Based Real-Time Prediction using Distillation

Arindam Bhattacharya, Ankith Ms, Ankit Gandhi, Vijay Huddar, Atul Saroop and Rahul Bhagat   16:15-17:45 (Frontenac Ballroom and Queen's Quay)

Relevance in E-commerce Product Search is crucial for providing customers with accurate results that match their query intent. With recent advancements in NLP and Deep Learning, Transformers have become the default choice for relevance classification tasks. In such a setting, the relevance model uses query text and product title as input features, and estimates if the product is relevant for the customer query. While cross-attention in Transformers enables a more accurate relevance prediction in such a setting, its high evaluation latency makes it unsuitable for real-time predictions in which thousands of products must be evaluated against a user query within few milliseconds. To address this issue, we propose CUPID: a Curriculum learning based real-time Prediction using Distillation that utilizes knowledge distillation within a curriculum learning setting to learn a simpler architecture that can be evaluated within low latency budgets. In a bi-lingual relevance prediction task, our approach shows an 302 bps improvement on English and 676 bps improvement for low-resource Arabic, while maintaining the low evaluation latency on CPUs.

[Industry] KAFA: Rethinking Image Ad Understanding with Knowledge-Augmented Feature Adaptation of Vision-Language Models

Zhiwei Jia, Pradyumna Narayana, Arjun Akula, Garima Pruthi, Hao Su, Sugato Basu and Varun Jampani 16:15-17:45 (Frontenac Ballroom and Queen's Quay)

Image ad understanding is a crucial task with wide real-world applications. Although highly challenging with the involvement of diverse atypical scenes, real-world entities, and reasoning over scene-texts, how to interpret image ads is relatively under-explored, especially in the era of foundational vision-language models (VLMs) featuring impressive generalizability and adaptability. In this paper, we perform the first empirical study of image ad understanding through the lens of pre-trained VLMs. We benchmark and reveal practical challenges in adapting these VLMs to image ad understanding. We propose a simple feature adaptation strategy to effectively fuse multimodal information for image ads and further empower it with knowledge of real-world entities. We hope our study draws more attention to image ad understanding which is broadly relevant to the advertising industry.

[Industry] MathPrompter: Mathematical Reasoning using Large Language Models

Shima Imani, Liang Du and Harsh Shrivastava   16:15-17:45 (Frontenac Ballroom and Queen's Quay)

Large Language Models (LLMs) have limited performance when solving arithmetic reasoning tasks and often provide incorrect answers. Unlike natural language understanding, math problems typically have a single correct answer, making the task of generating accurate solutions more challenging for LLMs. To the best of our knowledge, we are not aware of any LLMs that indicate their level of confidence in their responses which fuels a trust deficit in these models impeding their adoption. To address this deficiency, we propose 'MathPrompter', a technique that improves performance of LLMs on arithmetic problems along with increased reliance in the predictions. MathPrompter uses the Zero-shot chain-of-thought prompting technique to generate multiple algebraic expressions or python functions to solve the same math problem in different ways and thereby raise the confidence level in the output results. This is in contrast to other prompt based CoT methods, where there is no check on the validity of the intermediate steps followed. Our technique improves over state-of-the-art on the 'MultiArith' dataset (78.7% -> 92.5%) evaluated using 175B parameter GPT-based LLM.

[Industry] Distilled Language Models are economically efficient for the enterprise. ...mostly.

Kristen Howell, Gwen Christian, Pavel Fomitchov, Gitit Kehat, Julianne Marzulla, Leanne Rolston, Jadin Tredup, Ilana Zimmerman, Ethan Selfridge and Joseph Bradley   16:15-17:45 (Frontenac Ballroom and Queen's Quay)

Contacting customer service via chat is a common practice. Because employing customer service agents is expensive, many companies are turning to NLP that assists human agents by auto-generating responses that can be used directly or with modifications. With their ability to handle large context windows, Large Language Models (LLMs) are a natural fit for this use case. However, their efficacy must be balanced with the cost of training and serving them. This paper assesses the practical cost and impact of LLMs for the enterprise as a function of the usefulness of the responses that they generate. We present a cost framework for evaluating an NLP model's utility for this use case and apply it to a single brand as a case study in the context of an existing agent assistance product. We compare three strategies for specializing an LLM — prompt engineering, fine-tuning, and knowledge distillation — using feedback from the brand's customer service agents. We find that the usability of a model's responses can make up for a large difference in inference cost for our case study brand, and we extrapolate our findings to the broader enterprise space.

[Industry] DISCOSQA: A Knowledge Base Question Answering System for Space Debris based on Program Induction

Paul Darm, Antonio Valerio Miceli Barone, Shay B. Cohen and Annalisa Riccardi   16:15-17:45 (Frontenac Ballroom and Queen's Quay)

Space program agencies execute complex satellite operations that need to be supported by the technical knowledge contained in their extensive information systems. Knowledge Base (KB) databases are an effective way of storing and accessing such information to scale. In this work we present a system, developed for the European Space Agency, that can answer complex natural language queries, to support engineers in accessing the information contained in a KB that models the orbital space debris environment. Our system is based on a pipeline which first generates a program sketch from a natural language question, then specializes the sketch into a concrete query program with mentions of entities, attributes and relations, and finally executes the program against the database. This pipeline decomposition approach enables us to train the system by leveraging out-of-domain data and semi-synthetic data generated by GPT-3, thus reducing overfitting and shortcut learning even with limited amount of in-domain training data.

[Industry] Unified Contextual Query Rewriting

Yingxue Zhou, Jie Hao, Mukund Rungta, Yang Liu, Eunah Cho, Xing Fan, Yanbin Lu, Vishal Vasudevan, Kellen Gillespie and Zeynab Raeesy 16:15-17:45 (Frontenac Ballroom and Queen's Quay)

Query rewriting (QR) is an important technique for user friction (i.e. recovering ASR error or system error) reduction and contextual carry-over (i.e. ellipsis and co-reference) in conversational AI systems. Recently, generation-based QR models have achieved promising results on these two tasks separately. Although these two tasks have many similarities such as they both use the previous dialogue along with the current request as model input, there is no unified model to solve them jointly. To this end, we propose a unified contextual query rewriting model that unifies QR for both reducing friction and contextual carryover purpose. Moreover, we involve multiple auxiliary tasks such as trigger prediction and NLU interpretation tasks to boost the performance of the rewrite. We leverage the text-to-text unified framework which uses independent tasks with weighted loss to account for task importance. Then we propose new unified multitask learning strategies including

a sequential model which outputs one sentence for multi-tasks, and a hybrid model where some tasks are independent and some tasks are sequentially generated. Our experimental results demonstrate the effectiveness of the proposed unified learning methods.

[Industry] Exploring Zero and Few-shot Techniques for Intent Classification
Soham Parikh, Mitul Tiwari, Prashil Tumbade and Quaizar Vohra   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Conversational NLU providers often need to scale to thousands of intent-classification models where new customers often face the cold-start problem. Scaling to so many customers puts a constraint on storage space as well. In this paper, we explore four different zero and few-shot intent classification approaches with this low-resource constraint: 1) domain adaptation, 2) data augmentation, 3) zero-shot intent classification using descriptions large language models (LLMs), and 4) parameter-efficient fine-tuning of instruction-finetuned language models. Our results show that all these approaches are effective to different degrees in low-resource settings. Parameter-efficient fine-tuning using T-few recipe on Flan-T5 yields the best performance even with just one sample per intent. We also show that the zero-shot method of prompting LLMs using intent descriptions is also very competitive.

[Industry] Generate-then-Retrieve: Intent-Aware FAQ Retrieval in Product Search
Zhiyu Chen, Jason Choi, Besnik Fetahu, Oleg Rokhlenko and Shervin Malmasi   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Frequently Asked Question (FAQ) retrieval aims at retrieving question-answer pairs for a given a user query. Integrating FAQ retrieval with product search can not only empower users to make more informed purchase decisions, but also enhance user retention through efficient post-purchase support. Providing FAQ content without disrupting user's shopping experience poses challenges on deciding when and how to show FAQ results.

Our proposed intent-aware FAQ retrieval consists of (1) an intent classifier that predicts whether the query is looking for an FAQ; (2) a reformulation model that rewrites query into a natural question. Offline evaluation demonstrates that our approach improves 12% in Hit@1 on retrieving ground-truth FAQs, while reducing latency by 95% compared to baseline systems. These improvements are further validated by real user feedback, where more than 99% of users consider FAQs displayed on top of product search results is helpful. Overall, our findings show promising directions for integrating FAQ retrieval into product search at scale.

[Industry] EvolveMT: an Ensemble MT Engine Improving Itself with Usage Only
Kamer Yüksel, Ahmet Gunduz, Mohamed Al-badrashiny and Hassan Sawaf   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
This work proposes a method named EvolveMT for the efficient combination of multiple machine translation (MT) engines. The method selects the output from one engine for each segment, using online learning techniques to predict the most appropriate system for each translation request. A neural quality estimation metric supervises the method without requiring reference translations. The method's online learning capability enables it to adapt to changes in the domain or MT engines dynamically, eliminating the requirement for retraining. The method selects a subset of translation engines to be called based on the source sentence features. The degree of exploration is configurable according to the desired quality-cost trade-off. Results from custom datasets demonstrate that EvolveMT achieves similar translation accuracy at a lower cost than selecting the best translation of each segment from all translations using an MT quality estimator. To the best of our knowledge, EvolveMT is the first MT system that adapts itself after deployment to incoming translation requests from the production environment without needing costly retraining on human feedback.

[Industry] MobileNMT: Enabling Translation in 15MB and 30ms
Ye Lin, Xiaohui Wang, Zhexi Zhang, Mingxuan Wang, Tong Xiao and Jingbo Zhu   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Deploying NMT models on mobile devices is essential for privacy, low latency, and offline scenarios. For high model capacity, NMT models are rather large. Running these models on devices is challenging with limited storage, memory, computation, and power consumption. Existing work either only focuses on a single metric such as FLOPs or general engine which is not good at auto-regressive decoding. In this paper, we present MobileNMT, a system that can translate in 15MB and 30ms on devices. We propose a series of principles for model compression when combined with quantization. Further, we implement an engine that is friendly to INT8 and decoding. With the co-design of model and engine, compared with the existing system, we speed up 47.0x and save 99.5% of memory with only 11.6% loss of BLEU. Our code will be publicly available after the anonymity period.

[Industry] Multi-doc Hybrid Summarization via Salient Representation Learning
Min Xiao   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Multi-document summarization is gaining more and more attention recently and serves as an invaluable tool to obtain key facts among a large information pool. In this paper, we proposed a multi-document hybrid summarization approach, which simultaneously generates a human-readable summary and extracts corresponding key evidences based on multi-doc inputs. To fulfill that purpose, we crafted a salient representation learning method to induce latent salient features, which are effective for joint evidence extraction and summary generation. In order to train this model, we conducted multi-task learning to optimize a composited loss, constructed over extractive and abstractive sub-components in a hierarchical way. We implemented the system based on a ubiquotously adopted transformer architecture and conducted experimental studies on multiple datasets across two domains, achieving superior performance over the baselines.

[Industry] Toward More Accurate and Generalizable Evaluation Metrics for Task-Oriented Dialogs
Abishek Komma, Nagesh Panyam Chandrasekarasastry, Timothy Leffel, Anuj Goyal, Angeliki Metallinou, Spyros Matsoukas and Aram Galstyan   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Measurement of interaction quality is a critical task for the improvement of large-scale spoken dialog systems. Existing approaches to dialog quality estimation either focus on evaluating the quality of individual turns, or collect dialog-level quality measurements from end users immediately following an interaction. In contrast to these approaches, we introduce a new dialog-level annotation workflow called Dialog Quality Annotation (DQA). DQA expert annotators evaluate the quality of dialogs as a whole, and also label dialogs for attributes such as goal completion and user sentiment. In this contribution, we show that: (i) while dialog quality cannot be completely decomposed into dialog-level attributes, there is a strong relationship between some objective dialog attributes and judgments of dialog quality; (ii) for the task of dialog-level quality estimation, a supervised model trained on dialog-level annotations outperforms methods based purely on aggregating turn-level features; and (iii) the proposed evaluation model shows better domain generalization ability compared to the baselines. On the basis of these results, we argue that having high-quality human-annotated data is an important component of evaluating interaction quality for large industrial-scale voice assistant platforms.

[Industry] Mitigating the Burden of Redundant Datasets via Batch-Wise Unique Samples and Frequency-Aware Losses
Donato Crisostomi, Andrea Caciolai, Alessandro Pedrani, Kay Rottmann, Alessandro Manzotti, Enrico Palumbo and Davide Bernardi 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Datasets used to train deep learning models in industrial settings often exhibit skewed distributions with some samples repeated a large number of times. This paper presents a simple yet effective solution to reduce the increased burden of repeated computation on redundant datasets. Our approach eliminates duplicates at the batch level, without altering the data distribution observed by the model, making it model-agnostic and easy to implement as a plug-and-play module. We also provide a mathematical expression to estimate the reduction in training time that our approach provides. Through empirical evidence, we show that our approach significantly reduces training times on various models

across datasets with varying redundancy factors, without impacting their performance on the Named Entity Recognition task, both on publicly available datasets and in real industrial settings. In the latter, the approach speeds training by up to 87%, and by 46% on average, with a drop in model performance of 0.2% relative at worst. We finally release a modular and reusable codebase to further advance research in this area.

[Industry] Search Query Spell Correction with Weak Supervision in E-commerce
Vishal Kakkar, Chinmay Sharma, Madhura Pande and Surender Kumar   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Misspelled search queries in e-commerce can lead to empty or irrelevant products. Besides inadvertent typing mistakes, most spell mistakes occur because the user does not know the correct spelling, hence typing it as it is pronounced colloquially. This colloquial typing creates countless misspelling patterns for a single correct query. In this paper, we first systematically analyze and group different spell errors into error classes and then leverage the state-of-the-art Transformer model for contextual spell correction. We overcome the constraint of limited human labelled data by proposing novel synthetic data generation techniques for voluminous generation of training pairs needed by data hungry Transformers, without any human intervention. We further utilize weakly supervised data coupled with curriculum learning strategies to improve on tough spell mistakes without regressing on the easier ones. We show significant improvements from our model on human labeled data and online A/B experiments against multiple state-of-art models.

[Industry] Referring to Screen Texts with Voice Assistants
Shruti Bhargava, Anand Dhoot, Ing-marie Jonsson, Hoang Long Nguyen, Alkesh Patel, Hong Yu and Vincent Renkens   16:15-17:45
(Frontenac Ballroom and Queen's Quay)
Voice assistants help users make phone calls, send messages, create events, navigate and do a lot more. However assistants have limited capacity to understand their users' context. In this work, we aim to take a step in this direction. Our work dives into a new experience for users to refer to phone numbers, addresses, email addresses, urls, and dates on their phone screens. We focus on reference understanding, which is particularly interesting when, similar to visual grounding, there are multiple similar texts on screen. We collect a dataset and propose a lightweight general purpose model for this novel experience. Since consuming pixels directly is expensive, our system is designed to rely only on text extracted from the UI. Our model is modular, offering flexibility, better interpretability and efficient run time memory.

[Industry] Weakly supervised hierarchical multi-task classification of customer questions
Jitenkumar Rana, Promod Yenigalla, Chetan Aggarwal, Sandeep Sricharan Mukku, Manan Soni and Rashmi Patange   16:15-17:45
(Frontenac Ballroom and Queen's Quay)
Identifying granular and actionable topics from customer questions (CQ) posted on e-commerce websites helps surface the missing information expected by customers on the product detail page (DP), provide insights to brands and sellers on what critical product information that the customers are looking before making a purchase decision and helps enrich the catalog quality to improve the overall customer experience (CX). We propose a weakly supervised Hierarchical Multi-task Classification Framework (HMCF) to identify topics from customer questions at various granularities. Complexity lies in creating a list of granular topics (taxonomy) for 1000s of product categories and building a scalable classification system. To this end, we introduce a clustering based Taxonomy Creation and Data Labeling (TCDL) module for creating taxonomy and labelled data with minimal supervision. Using TCDL module, taxonomy and labelled data creation task reduces to 2 hours as compared to 2 weeks of manual efforts by a subject matter expert. For classification, we propose a two level HMCF that performs multi-class classification to identify coarse level-1 topic and leverages NLI based label-aware approach to identify granular level-2 topic. We showcase that HMCF (based on BERT and NLI) a) achieves absolute improvement of 13% in Top-1 accuracy over single-task non-hierarchical baselines b) learns a generic domain invariant function that can adapt to constantly evolving taxonomy (open label set) without need of re-training. c) reduces model deployment efforts significantly since it needs only one model that caters to 1000s of product categories.

[Industry] xPQA: Cross-Lingual Product Question Answering in 12 Languages
Xiaoyu Shen, Akari Asai, Bill Byrne and Adria De Gispert   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Product Question Answering (PQA) systems are key in e-commerce applications as they provide responses to customers' questions as they shop for products. While existing work on PQA focuses mainly on English, in practice there is need to support multiple customer languages while leveraging product information available in English. To study this practical industrial task, we present xPQA, a large-scale annotated cross-lingual PQA dataset in 12 languages, and report results in (1) candidate ranking, to select the best English candidate containing the information to answer a non-English question; and (2) answer generation, to generate a natural-sounding non-English answer based on the selected English candidate. We evaluate various approaches involving machine translation at runtime or offline, leveraging multilingual pre-trained LMs, and including or excluding xPQA training data. We find that in-domain data is essential as cross-lingual rankers trained on other domains perform poorly on the PQA task, and that translation-based approaches are most effective for candidate ranking while multilingual finetuning works best for answer generation. Still, there remains a significant performance gap between the English and the cross-lingual test sets.

[TACL] The Parallelism Tradeoff: Limitations of Log-Precision Transformers
William Merrill and Ashish Sabharwal   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Despite their omnipresence in modern NLP, characterizing the computational power of transformer neural nets remains an interesting open question. We prove that transformers whose arithmetic precision is logarithmic in the number of input tokens (and whose feedforward nets are computable using space linear in their input) can be simulated by constant-depth logspace-uniform threshold circuits. This provides insight on the power of transformers using known results in complexity theory. For example, if $L \neq P$ (i.e., not all poly-time problems can be solved using logarithmic space), then transformers cannot even accurately solve linear equalities or check membership in an arbitrary context-free grammar with empty productions. Our result intuitively emerges from the transformer architecture's high parallelizability. We thus speculatively introduce the idea of a fundamental parallelism tradeoff: any model architecture as parallelizable as the transformer will obey limitations similar to it. Since parallelism is key to training models at massive scale, this suggests a potential inherent weakness of the scaling paradigm.

[SRW] Multi-Dialectal Representation Learning of Sinitic Phonology
Zhibai Jia   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Representation learning of Sinitic Phonology using a knowledge graph based method

[SRW] Classical Out-of-Distribution Detection Methods Benchmark in Text Classification Tasks
Mateusz Baran, Joanna Baran, Mateusz Wójcik, Maciej Zięba and Adam Gonczarek   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
The paper evaluates existing OOD detection methods for NLP systems and emphasizes the need for further research to develop more effective approaches to ensure safety and trustworthiness of NLP systems.

[SRW] A State-Vector Framework For Dataset Effects
Esmat Sahak, Zining Zhu and Frank Rudzicz   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
We propose a state-vector framework to analyze the effects of datasets on deep learning language models using probing tasks.

[SRW] Probing for Hyperbole in Pre-Trained Language Models
Nina Schneidermann, Daniel Hershcovich and Bolette Pedersen   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
This paper contributes to hyperbole identification research in NLP with two probing tasks (edge and MDL probing) for 3 pre-trained language models, as well as an attempt to shed light on problems annotating hyperbole.

[SRW] Acquiring Frame Element Knowledge with Deep Metric Learning for Semantic Frame Induction
Kosuke Yamada, Ryohei Sasano and Koichi Takeda   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
In the argument clustering for semantic frame induction, we propose a method that applies deep metric learning.

[Demo] NeuroX Library for Neuron Analysis of Deep NLP Models
Nadir Durrani, Hassan Sajjad and Fahim Dalvi   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Neuron analysis provides insights into how knowledge is structured in representations and discovers the role of neurons in the network. In addition to developing an understanding of our models, neuron analysis enables various applications such as debiasing, domain adaptation and architectural search. We present NeuroX, a comprehensive open-source toolkit to conduct neuron analysis of natural language processing models. It implements various interpretation methods under a unified API, and provides a framework for data processing and evaluation, thus making it easier for researchers and practitioners to perform neuron analysis. The Python toolkit is available at https://www.github.com/fdalvi/NeuroX.

Demo Video available at: https://youtu.be/mLhs2YMx4u8

[Demo] Hierarchy Builder: Organizing Textual Spans into a Hierarchy to Facilitate Navigation
Yoav Goldberg, Hillel Taub-Tabib and Itay Yair   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Information extraction systems often produce hundreds to thousands of strings on a specific topic. We present a method that facilitates better consumption of these strings, in an ex- ploratory setting in which a user wants to both get a broad overview of what's available, and a chance to dive deeper on some aspects. The sys- tem works by grouping similar items together, and arranging the remaining items into a hierarchical navigable DAG structure. We apply the method to medical information extraction.

[Demo] The ROOTS Search Tool: Data Transparency for LLMs
Anna Rogers, Yacine Jernite, Sasha Luccioni, Gérard Dupont, Hugo Laurençon, Paulo Villegas, Christopher Akiki and Aleksandra Piktus
16:15-17:45 (Frontenac Ballroom and Queen's Quay)
ROOTS is a 1.6TB multilingual text corpus developed for the training of BLOOM, currently the largest language model explicitly accompanied by commensurate data governance efforts. In continuation of these efforts, we present the ROOTS Search Tool: a search engine over the entire ROOTS corpus offering both fuzzy and exact search capabilities. ROOTS is the largest corpus to date that can be investigated this way. The ROOTS Search Tool is open-sourced and available on Hugging Face Spaces: https://huggingface.co/spaces/bigscience-data/roots-search. We describe our implementation and the possible use cases of our tool.

[Demo] UINAUIL: A Unified Benchmark for Italian Natural Language Understanding
Viviana Patti, Cristina Bosco, Alessio Bosca, Livio Bioglio and Valerio Basile   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
This paper introduces the Unified Interactive Natural Understanding of the Italian Language (UINAUIL), a benchmark of six tasks for Italian Natural Language Understanding. We present a description of the tasks and software library that collects the data from the European Language Grid, harmonizes the data format, and exposes functionalities to facilitates data manipulation and the evaluation of custom models. We also present the results of tests conducted with available Italian and multilingual language models on UINAUIL, providing an updated picture of the current state of the art in Italian NLU.

[Demo] Inseq: An Interpretability Toolkit for Sequence Generation Models
Oskar van der Wal, Ludwig Sickert, Nils Feldhus and Gabriele Sarti   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Past work in natural language processing interpretability focused mainly on popular classification tasks while largely overlooking generation settings, partly due to a lack of dedicated tools. In this work, we introduce Inseq, a Python library to democratize access to interpretability analyses of sequence generation models. Inseq enables intuitive and optimized extraction of models' internal information and feature importance scores for popular decoder-only and encoder-decoder Transformers architectures. We showcase its potential by adopting it to highlight gender biases in machine translation models and locate factual knowledge inside GPT-2. Thanks to its extensible interface supporting cutting-edge techniques such as contrastive feature attribution, Inseq can drive future advances in explainable natural language generation, centralizing good practices and enabling fair and reproducible model evaluations.

[Demo] Which Spurious Correlations Impact Reasoning in NLI Models? A Visual Interactive Diagnosis through Data-Constrained Counterfactuals
Mennatallah El-Assady, Afra Amini and Robin Chan   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
We present a human-in-the-loop dashboard tailored to diagnosing potential spurious features that NLI models rely on for predictions. The dashboard enables users to generate diverse and challenging examples by drawing inspiration from GPT-3 suggestions. Additionally, users can receive feedback from a trained NLI model on how challenging the newly created example is and make refinements based on the feedback. Through our investigation, we discover several categories of spurious correlations that impact the reasoning of NLI models, which we group into three categories: Semantic Relevance, Logical Fallacies, and Bias. Based on our findings, we identify and describe various research opportunities, including diversifying training data and assessing NLI models' robustness by creating adversarial test suites.

[Demo] VisKoP: Visual Knowledge oriented Programming for Interactive Knowledge Base Question Answering
Juanzi Li, Lei Hou, Peng Zhang, jianjun xu, Hailong Jin, Jifan Yu, Amy Xin, Shulin Cao, Xin Lv, YUANYONG CHEN and Zijun Yao   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
We present Visual Knowledge oriented Programming platform (VisKoP), a knowledge base question answering (KBQA) system that integrates human into the loop to edit and debug the knowledge base (KB) queries. VisKoP not only provides a neural program induction module, which converts natural language questions into knowledge oriented program language (KoPL), but also maps KoPL programs into graphical elements. KoPL programs can be edited with simple graphical operators, such as <i>"dragging"</i> to add knowledge operators and <i>"slot filling"</i> to designate operator arguments. Moreover, VisKoP provides auto-completion for its knowledge base schema and users can easily debug the KoPL program by checking its intermediate results. To facilitate the practical KBQA on a million-entity-level KB, we design a highly efficient KoPL execution engine for the back-end. Experiment results show that VisKoP is highly efficient and user interaction can fix a large portion of wrong KoPL programs to acquire the correct answer. The VisKoP online demo, highly efficient KoPL engine, and screencast video are now publicly available.

Human-in-the-loop Evaluation for Early Misinformation Detection: A Case Study of COVID-19 Treatments
Ethan Adrian Mendes, Yang Chen, Wei Xu and Alan Ritter   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
We present a human-in-the-loop evaluation framework for fact-checking novel misinformation claims and identifying social media messages

that support them. Our approach extracts check-worthy claims, which are aggregated and ranked for review. Stance classifiers are then used to identify tweets supporting novel misinformation claims, which are further reviewed to determine whether they violate relevant policies. To demonstrate the feasibility of our approach, we develop a baseline system based on modern NLP methods for human-in-the-loop fact-checking in the domain of COVID-19 misinformation. We make our data and annotation guidelines available to support the evaluation of human-in-the-loop systems that identify novel misinformation directly from raw user-generated content.

SWiPE: A Dataset for Document-Level Simplification of Wikipedia Pages
Philippe Laban, Jesse Vig, Wojciech Kryscinski, Shafiq Joty, Caiming Xiong and Chien-Sheng Jason Wu 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Text simplification research has mostly focused on sentence-level simplification, even though many desirable edits - such as adding relevant background information or reordering content - may require document-level context. Prior work has also predominantly framed simplification as a single-step, input-to-output task, only implicitly modeling the fine-grained, span-level edits that elucidate the simplification process. To address both gaps, we introduce the SWiPE dataset, which reconstructs the document-level editing process from English Wikipedia (EW) articles to paired Simple Wikipedia (SEW) articles. In contrast to prior work, SWiPE leverages the entire revision history when pairing pages in order to better identify simplification edits. We work with Wikipedia editors to annotate 5,000 EW-SEW document pairs, labeling more than 40,000 edits with proposed 19 categories. To scale our efforts, we propose several models to automatically label edits, achieving an F-1 score of up to 70.9, indicating that this is a tractable but challenging NLU task. Finally, we categorize the edits produced by several simplification models and find that SWiPE-trained models generate more complex edits while reducing unwanted edits.

A New Aligned Simple German Corpus
Vanessa Toborek, Moritz Busch, Malte Boßert, Christian Bauckhage and Pascal Welke 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
"Leichte Sprache", the German counterpart to Simple English, is a regulated language aiming to facilitate complex written language that would otherwise stay inaccessible to different groups of people. We present a new sentence-aligned monolingual corpus for Simple German – German. It contains multiple document-aligned sources which we have aligned using automatic sentence-alignment methods. We evaluate our alignments based on a manually labelled subset of aligned documents. The quality of our sentence alignments, as measured by the F1-score, surpasses previous work. We publish the dataset under CC BY-SA and the accompanying code under MIT license.

Target-Based Offensive Language Identification
Marcos Zampieri, Skye Morgan, Kai North, Tharindu Ranasinghe, Austin Simmmons, Paridhi Khandelwal, Sara Rosenthal and Preslav Nakov 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
We present TBO, a new dataset for Target-based Offensive language identification. TBO contains post-level annotations regarding the harmfulness of an offensive post and token-level annotations comprising of the target and the offensive argument expression. Popular offensive language identification datasets for social media focus on annotation taxonomies only at the post level and more recently, some datasets have been released that feature only token-level annotations. TBO is an important resource that bridges the gap between post-level and token-level annotation datasets by introducing a single comprehensive unified annotation taxonomy. We use the TBO taxonomy to annotate post-level and token-level offensive language on English Twitter posts. We release an initial dataset of over 4,500 instances collected from Twitter and we carry out multiple experiments to compare the performance of different models trained and tested on TBO.

Multilingual Knowledge Graph Completion with Language-Sensitive Multi-Graph Attention
Rongchuan Tang, Yang Zhao, Chengqing Zong and Yu Zhou 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Multilingual Knowledge Graph Completion (KGC) aims to predict missing links with multilingual knowledge graphs. However, existing approaches suffer from two main drawbacks: (a) alignment dependency: the multilingual KGC is always realized with joint entity or relation alignment, which introduces additional alignment models and increases the complexity of the whole framework; (b) training inefficiency: the trained model will only be used for the completion of one target KG, although the data from all KGs are simultaneously used. To address these drawbacks, we propose a novel multilingual KGC framework with language-sensitive multi-graph attention such that the missing links on all given KGs can be inferred by a universal knowledge completion model. Specifically, we first build a relational graph neural network by sharing the embeddings of aligned nodes to transfer language-independent knowledge. Meanwhile, a language-sensitive multi-graph attention (LSMGA) is proposed to deal with the information inconsistency among different KGs. Experimental results show that our model achieves significant improvements on the DBP-5L and E-PKG datasets.

HAHE: Hierarchical Attention for Hyper-Relational Knowledge Graphs in Global and Local Level
Haoran Luo, Haihong E, Yuhao Yang, Yikai Guo, Mingzhi Sun, Tianyu Yao, Zichen Tang, Kaiyang Wan, Meina Song and Wei Lin 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Link Prediction on Hyper-relational Knowledge Graphs (HKG) is a worthwhile endeavor. HKG consists of hyper-relational facts (H-Facts), composed of a main triple and several auxiliary attribute-value qualifiers, which can effectively represent factually comprehensive information. The internal structure of HKG can be represented as a hypergraph-based representation globally and a semantic sequence-based representation locally. However, existing research seldom simultaneously models the graphical and sequential structure of HKGs, limiting HKGs' representation. To overcome this limitation, we propose a novel Hierarchical Attention model for HKG Embedding (HAHE), including global-level and local-level attention. The global-level attention can model the graphical structure of HKG using hypergraph dual-attention layers, while the local-level attention can learn the sequential structure inside H-Facts via heterogeneous self-attention layers. Experiment results indicate that HAHE achieves state-of-the-art performance in link prediction tasks on HKG standard datasets. In addition, HAHE addresses the issue of HKG multi-position prediction for the first time, increasing the applicability of the HKG link prediction task. Our code is publicly available.

Text-to-SQL Error Correction with Language Models of Code
Ziru Chen, Shijie Chen, Michael White, Raymond Mooney, Ali Payani, Jayanth Srinivasa, Yu Su and Huan Sun 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Despite recent progress in text-to-SQL parsing, current semantic parsers are still not accurate enough for practical use. In this paper, we investigate how to build automatic text-to-SQL error correction models. Noticing that token-level edits are out of context and sometimes ambiguous, we propose building clause-level edit models instead. Besides, while most language models of code are not specifically pre-trained for SQL, they know common data structures and their operations in programming languages such as Python. Thus, we propose a novel representation for SQL queries and their edits that adheres more closely to the pre-training corpora of language models of code. Our error correction model improves the exact set match accuracy of different parsers by 2.4-6.5 and obtains up to 4.3 point absolute improvement over two strong baselines.

Characterization of Stigmatizing Language in Medical Records
Keith Harrigan, Ayah Zirikly, Brant Chee, Alya Ahmad, Anne R Links, Somnath Saha, Mary Catherine Beach and Mark Dredze 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Widespread disparities in clinical outcomes exist between different demographic groups in the United States. A new line of work in medical sociology has demonstrated physicians often use stigmatizing language in electronic medical records within certain groups, such as black

patients, which may exacerbate disparities. In this study, we characterize these instances at scale using a series of domain-informed NLP techniques. We highlight important differences between this task and analogous bias-related tasks studied within the NLP community (e.g., classifying microaggressions). Our study establishes a foundation for NLP researchers to contribute timely insights to a problem domain brought to the forefront by recent legislation regarding clinical documentation transparency. We release data, code, and models.

GreenKGC: A Lightweight Knowledge Graph Completion Method

Yun Cheng Wang, Xiou Ge, Bin Wang and C.-C. Jay Kuo   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Knowledge graph completion (KGC) aims to discover missing relationships between entities in knowledge graphs (KGs). Most prior KGC work focuses on learning embeddings for entities and relations through a simple score function. Yet, a higher-dimensional embedding space is usually required for a better reasoning capability, which leads to larger model size and hinders applicability to real-world problems (e.g., large-scale KGs or mobile/edge computing). A lightweight modularized KGC solution, called GreenKGC, is proposed in this work to address this issue. GreenKGC consists of three modules: representation learning, feature pruning, and decision learning, to extract discriminant KG features and make accurate predictions on missing relationships using classifiers and negative sampling. Experimental results demonstrate that, in low dimensions, GreenKGC can outperform SOTA methods in most datasets. In addition, low-dimensional GreenKGC can achieve competitive or even better performance against high-dimensional models with a much smaller model size.

Neural Machine Translation for Mathematical Formulae

Felix Petersen, Moritz Schubotz, Andre Greiner-Petter and Bela Gipp   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
We tackle the problem of neural machine translation of mathematical formulae between ambiguous presentation languages and unambiguous content languages. Compared to neural machine translation on natural language, mathematical formulae have a much smaller vocabulary and much longer sequences of symbols, while their translation requires extreme precision to satisfy mathematical information needs. In this work, we perform the tasks of translating from LaTeX to Mathematica as well as from LaTeX to semantic LaTeX. While recurrent, recursive, and transformer networks struggle with preserving all contained information, we find that convolutional sequence-to-sequence networks achieve 95.1% and 90.7% exact matches, respectively.

FutureTOD: Teaching Future Knowledge to Pre-trained Language Model for Task-Oriented Dialogue

*Weihao Zeng, Keqing He, Yejie Wang, Chen Zeng, Jingang Wang, Yunsen Xian and Weiran Xu*16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Pre-trained language models based on general text enable huge success in the NLP scenario. But the intrinsical difference of linguistic patterns between general text and task-oriented dialogues makes existing pre-trained language models less useful in practice. Current dialogue pre-training methods rely on a contrastive framework and face the challenges of both selecting true positives and hard negatives. In this paper, we propose a novel dialogue pre-training model, FutureTOD, which distills future knowledge to the representation of the previous dialogue context using a self-training framework. Our intuition is that a good dialogue representation both learns local context information and predicts future information. Extensive experiments on diverse downstream dialogue tasks demonstrate the effectiveness of our model, especially the generalization, robustness, and learning discriminative dialogue representations capabilities.

History Semantic Graph Enhanced Conversational KBQA with Temporal Information Modeling

Hao Sun, Yang Li, Liwei Deng, Bowen Li, Binyuan Hui, Binhua Li, Yunshi Lan, Yan Zhang and Yongbin Li 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Context information modeling is an important task in conversational KBQA. However, existing methods usually assume the independence of utterances and model them in isolation. In this paper, we propose a History Semantic Graph Enhanced Conversational KBQA model (HSGE) that is able to effectively model long-range semantic dependencies in conversation history while maintaining low computational cost. The framework incorporates a context-aware encoder, which employs a dynamic memory decay mechanism and models context at different levels of granularity. We evaluate HSGE on a widely used benchmark dataset for complex sequential question answering. Experimental results demonstrate that it outperforms existing baselines averaged on all question types.

I Cast Detect Thoughts: Learning to Converse and Guide with Intents and Theory-of-Mind in Dungeons and Dragons

Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi and Prithviraj Ammanabrolu   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
We propose a novel task, G4C, to study teacher-student natural language interactions in a goal-driven and grounded environment. Dungeons and Dragons (D&D), a role-playing game, provides an ideal setting to investigate such interactions. Here, the Dungeon Master (DM), i.e., the teacher, guides the actions of several players—students, each with their own personas and abilities—to achieve shared goals grounded in a fantasy world. Our approach is to decompose and model these interactions into (1) the DM's intent to guide players toward a given goal; (2) the DM's guidance utterance to the players expressing this intent; and (3) a theory-of-mind (ToM) model that anticipates the players' reaction to the guidance one turn into the future. We develop a novel reinforcement learning (RL) method for training a DM that generates guidance for players by rewarding utterances where the intent matches the ToM-anticipated player actions. Human and automated evaluations show that a DM trained to explicitly model intents and incorporate ToM of the players using RL generates better-quality guidance that is 3x more likely to fulfill the DM's intent than a vanilla natural language generation (NLG) approach.

RobuT: A Systematic Study of Table QA Robustness Against Human-Annotated Adversarial Perturbations

Yilun Zhao, Chen Zhao, Linyong Nan, Zhenting Qi, Wenlin Zhang, Xiangru Tang, Boyu Mi and Dragomir Radev   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Despite significant progress having been made in question answering on tabular data (Table QA), it's unclear whether, and to what extent existing Table QA models are robust to task-specific perturbations, e.g., replacing key question entities or shuffling table columns. To systematically study the robustness of Table QA models, we propose a benchmark called RobuT, which builds upon existing Table QA datasets (WTQ, WikiSQL-Weak, and SQA) and includes human-annotated adversarial perturbations in terms of table header, table content, and question. Our results indicate that both state-of-the-art Table QA models and large language models (e.g., GPT-3) with few-shot learning falter in these adversarial sets. We propose to address this problem by using large language models to generate adversarial examples to enhance training, which significantly improves the robustness of Table QA models.

Ambiguous Learning from Retrieval: Towards Zero-shot Semantic Parsing

Shan Wu, Chunlei Xin, Hongyu Lin, Xianpei Han, Cao Liu, Jiansong Chen, Fan Yang, Guanglu Wan and Le Sun   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Current neural semantic parsers take a supervised approach requiring a considerable amount of training data which is expensive and difficult to obtain. Thus, minimizing the supervision effort is one of the key challenges in semantic parsing. In this paper, we propose the Retrieval as Ambiguous Supervision framework, in which we construct a retrieval system based on pretrained language models to collect high-coverage candidates. Assuming candidates always contain the correct ones, we convert zero-shot task into ambiguously supervised task. To improve the precision and coverage of such ambiguous supervision, we propose a confidence-driven self-training algorithm, in which a semantic parser is learned and exploited to disambiguate the candidates iteratively. Experimental results show that our approach significantly outperforms the

state-of-the-art zero-shot semantic parsing methods.

LI-RAGE: Late Interaction Retrieval Augmented Generation with Explicit Signals for Open-Domain Table Question Answering
Weizhe Lin, Rexhina Blloshmi, Bill Byrne, Adria de Gispert and Gonzalo Iglesias 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Recent open-domain TableQA models are typically implemented as retriever-reader pipelines. The retriever component is usually a variant of the Dense Passage Retriever, which computes the similarities between questions and tables based on a single representation of each. These fixed vectors can be insufficient to capture fine-grained features of potentially very big tables with heterogeneous row/column information. We address this limitation by 1) applying late interaction models which enforce a finer-grained interaction between question and table embeddings at retrieval time. In addition, we 2) incorporate a joint training scheme of the retriever and reader with explicit table-level signals, and 3) embed a binary relevance token as a prefix to the answer generated by the reader, so we can determine at inference time whether the table used to answer the question is reliable and filter accordingly. The combined strategies set a new state-to-the-art performance on two public open-domain TableQA datasets.

Reasoning Implicit Sentiment with Chain-of-Thought Prompting
Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li and Tat-Seng Chua 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
While sentiment analysis systems try to determine the sentiment polarities of given targets based on the key opinion expressions in input texts, in implicit sentiment analysis (ISA) the opinion cues come in an implicit and obscure manner. Thus detecting implicit sentiment requires the common-sense and multi-hop reasoning ability to infer the latent intent of opinion. Inspired by the recent chain-of-thought (CoT) idea, in this work we introduce a Three-hop Reasoning (THOR) CoT framework to mimic the human-like reasoning process for ISA. We design a three-step prompting principle for THOR to step-by-step induce the implicit aspect, opinion, and finally the sentiment polarity. Our THOR+Flan-T5 (11B) pushes the state-of-the-art (SoTA) by over 6% F1 on supervised setup. More strikingly, THOR+GPT3 (175B) boosts the SoTA by over 50% F1 on zero-shot setting.

Similarity-weighted Construction of Contextualized Commonsense Knowledge Graphs for Knowledge-intense Argumentation Tasks
Moritz Plenz, Juri Opitz, Philipp Heinisch, Philipp Cimiano and Anette Frank 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Arguments often do not make explicit how a conclusion follows from its premises. To compensate for this lack, we enrich arguments with structured background knowledge to support knowledge-intense argumentation tasks. We present a new unsupervised method for constructing Contextualized Commonsense Knowledge Graphs (CCKGs) that selects contextually relevant knowledge from large knowledge graphs (KGs) efficiently and at high quality. Our work goes beyond context-insensitive knowledge extraction heuristics by computing semantic similarity between KG triplets and textual arguments. Using these triplet similarities as weights, we extract contextualized knowledge paths that connect a conclusion to its premise, while maximizing similarity to the argument. We combine multiple paths into a CCKG that we optionally prune to reduce noise and raise precision. Intrinsic evaluation of the quality of our graphs shows that our method is effective for (re)constructing human explanation graphs. Manual evaluations in a large-scale knowledge selection setup verify high recall and precision of implicit CSK in the CCKGs. Finally, we demonstrate the effectiveness of CCKGs in a knowledge-insensitive argument quality rating task, outperforming strong baselines and rivaling a GPT-3 based system.

ArgAnalysis35K : A large-scale dataset for Argument Quality Analysis
Omkar Jayant Joshi, Priya N Pitre and Yashodhara Haribhakta 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Argument Quality Detection is an emerging field in NLP which has seen significant recent development. However, existing datasets in this field suffer from a lack of quality, quantity and diversity of topics and arguments, specifically the presence of vague arguments that are not persuasive in nature. In this paper, we leverage a combined experience of 10+ years of Parliamentary Debating to create a dataset that covers significantly more topics and has a wide range of sources to capture more diversity of opinion. With 34,890 high-quality argument-analysis pairs (a term we introduce in this paper), this is also the largest dataset of its kind to our knowledge. In addition to this contribution, we introduce an innovative argument scoring system based on instance-level annotator reliability and propose a quantitative model of scoring the relevance of arguments to a range of topics.

Trading Syntax Trees for Wordpieces: Target-oriented Opinion Words Extraction with Wordpieces and Aspect Enhancement
Samuel Mensah, Kai Sun and Nikolaos Aletras 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
State-of-the-art target-oriented opinion word extraction (TOWE) models typically use BERT-based text encoders that operate on the word level, along with graph convolutional networks (GCNs) that incorporate syntactic information extracted from syntax trees. These methods achieve limited gains with GCNs and have difficulty using BERT wordpieces. Meanwhile, BERT wordpieces are known to be effective at representing rare words or words with insufficient context information. To address this issue, this work trades syntax trees for BERT wordpieces by entirely removing the GCN component from the methods' architectures. To enhance TOWE performance, we tackle the issue of aspect representation loss during encoding. Instead of solely utilizing a sentence as the input, we use a sentence-aspect pair. Our relatively simple approach achieves state-of-the-art results on benchmark datasets and should serve as a strong baseline for further research.

Tracing Linguistic Markers of Influence in a Large Online Organisation
Prashant Khare, Ravi Shekhar, Mladen Karan, Stephen McQuistin, Colin Perkins, Ignacio Castro, Gareth Tyson, Patrick G.T. Healey and Matthew Purver 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Social science and psycholinguistic research have shown that power and status affect how people use language in a range of domains. Here, we investigate a similar question in a large, distributed, consensus-driven community with little traditional power hierarchy – the Internet Engineering Task Force (IETF), a collaborative organisation that designs internet standards. Our analysis based on lexical categories (LIWC) and BERT, shows that participants' levels of influence can be predicted from their email text, and identify key linguistic differences (e.g., certain LIWC categories, such as "WE" are positively correlated with high-influence). We also identify the differences in language use for the same person before and after becoming influential.

Discourse-Level Representations can Improve Prediction of Degree of Anxiety
Swanie Juhng, Matthew Matero, Vasudha Varadarajan, Johannes C Eichstaedt, Adithya V Ganesan and H. Andrew Schwartz 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Anxiety disorders are the most common of mental illnesses, but relatively little is known about how to detect them from language. The primary clinical manifestation of anxiety is worry associated cognitive distortions, which are likely expressed at the discourse-level of semantics. Here, we investigate the development of a modern linguistic assessment for degree of anxiety, specifically evaluating the utility of discourse-level information in addition to lexical-level large language model embeddings. We find that a combined lexico-discourse model outperforms models based solely on state-of-the-art contextual embeddings (RoBERTa), with discourse-level representations derived from Sentence-BERT and DiscRE both providing additional predictive power not captured by lexical-level representations. Interpreting the model, we find that discourse patterns of causal explanations, among others, were used significantly more by those scoring high in anxiety, dovetailing with psychological literature.

From Dogwhistles to Bullhorns: Unveiling Coded Rhetoric with Language Models

Julia Mendelsohn, Ronan Le Bras, Yejin Choi and Maarten Sap   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Dogwhistles are coded expressions that simultaneously convey one meaning to a broad audience and a second, often hateful or provocative, meaning to a narrow in-group; they are deployed to evade both political repercussions and algorithmic content moderation. For example, the word "cosmopolitan" in a sentence such as "we need to end the cosmopolitan experiment" can mean "worldly" to many but also secretly mean "Jewish" to a select few. We present the first large-scale computational investigation of dogwhistles. We develop a typology of dogwhistles, curate the largest-to-date glossary of over 300 dogwhistles with rich contextual information and examples, and analyze their usage in historical U.S. politicians' speeches. We then assess whether a large language model (GPT-3) can identify dogwhistles and their meanings, and find that GPT-3's performance varies widely across types of dogwhistles and targeted groups. Finally, we show that harmful content containing dogwhistles avoids toxicity detection, highlighting online risks presented by such coded language. This work sheds light on the theoretical and applied importance of dogwhistles in both NLP and computational social science, and provides resources to facilitate future research in modeling dogwhistles and mitigating their online harms.

HiPool: Modeling Long Documents Using Graph Neural Networks
Irene Li, Aosong Feng, Dragomir Radev and Rex Ying   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Encoding long sequences in Natural Language Processing (NLP) is a challenging problem. Though recent pretraining language models achieve satisfying performances in many NLP tasks, they are still restricted by a pre-defined maximum length, making them challenging to be extended to longer sequences. So some recent works utilize hierarchies to model long sequences. However, most of them apply sequential models for upper hierarchies, suffering from long dependency issues. In this paper, we alleviate these issues through a graph-based method. We first chunk the sequence with a fixed length to model the sentence-level information. We then leverage graphs to model intra- and cross-sentence correlations with a new attention mechanism. Additionally, due to limited standard benchmarks for long document classification (LDC), we propose a new challenging benchmark, totaling six datasets with up to 53k samples and 4034 average tokens' length. Evaluation shows our model surpasses competitive baselines by 2.6% in F1 score, and 4.8% on the longest sequence dataset. Our method is shown to outperform hierarchical sequential models with better performance and scalability, especially for longer sequences.

DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains
Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille and Pierre-Antoine Gourraud 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
In recent years, pre-trained language models (PLMs) achieve the best performance on a wide range of natural language processing (NLP) tasks. While the first models were trained on general domain data, specialized ones have emerged to more effectively treat specific domains. In this paper, we propose an original study of PLMs in the medical domain on French language. We compare, for the first time, the performance of PLMs trained on both public data from the web and private data from healthcare establishments. We also evaluate different learning strategies on a set of biomedical tasks. In particular, we show that we can take advantage of already existing biomedical PLMs in a foreign language by further pre-train it on our targeted data. Finally, we release the first specialized PLMs for the biomedical field in French, called DrBERT, as well as the largest corpus of medical data under free license on which these models are trained.

KALM: Knowledge-Aware Integration of Local, Document, and Global Contexts for Long Document Understanding
Shangbin Feng, Zhaoxuan Tan, Wenqian Zhang, Zhenyu Lei and Yulia Tsvetkov   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
With the advent of pre-trained language models (LMs), increasing research efforts have been focusing on infusing commonsense and domain-specific knowledge to prepare LMs for downstream tasks. These works attempt to leverage knowledge graphs, the de facto standard of symbolic knowledge representation, along with pre-trained LMs. While existing approaches leverage external knowledge, it remains an open question how to jointly incorporate knowledge graphs represented in varying contexts — from local (e.g., sentence), document-level, to global knowledge, to enable knowledge-rich and interpretable exchange across contexts. In addition, incorporating varying contexts can especially benefit long document understanding tasks that leverage pre-trained LMs, typically bounded by the input sequence length. In light of these challenges, we propose KALM, a language model that jointly leverages knowledge in local, document-level, and global contexts for long document understanding. KALM firstly encodes long documents and knowledge graphs into the three knowledge-aware context representations. KALM then processes each context with context-specific layers. These context-specific layers are followed by a ContextFusion layer that facilitates knowledge exchange to derive an overarching document representation. Extensive experiments demonstrate that KALM achieves state-of-the-art performance on three long document understanding tasks across 6 datasets/settings. Further analyses reveal that the three knowledge-aware contexts are complementary and they all contribute to model performance, while the importance and information exchange patterns of different contexts vary on different tasks and datasets.

Plug-and-Play Knowledge Injection for Pre-trained Language Models
Zhengyan Zhang, Zhiyuan Zeng, Yankai Lin, Huadong Wang, Deming Ye, Chaojun Xiao, Xu Han, Zhiyuan Liu, Peng Li, Maosong Sun and Jie Zhou   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Injecting external knowledge can improve the performance of pre-trained language models (PLMs) on various downstream NLP tasks. However, massive retraining is required to deploy new knowledge injection methods or knowledge bases for downstream tasks. In this work, we are the first to study how to improve the flexibility and efficiency of knowledge injection by reusing existing downstream models. To this end, we explore a new paradigm *plug-and-play knowledge injection*, where knowledge bases are injected into frozen existing downstream models by a *knowledge plugin*. Correspondingly, we propose a plug-and-play injection method *map-tuning*, which trains a mapping of knowledge embeddings to enrich model inputs with mapped embeddings while keeping model parameters frozen. Experimental results on three knowledge-driven NLP tasks show that existing injection methods are not suitable for the new paradigm, while map-tuning effectively improves the performance of downstream models. Moreover, we show that a frozen downstream model can be well adapted to different domains with different mapping networks of domain knowledge. Our code and models are available at https://github.com/THUNLP/Knowledge-Plugin.

LAMBADA: Backward Chaining for Automated Reasoning in Natural Language
Mehran Kazemi, Najoung Kim, Deepti Bhatia, Xin Xu and Deepak Ramachandran   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Remarkable progress has been made on automated reasoning with natural text, by using Large Language Models (LLMs) and methods such as Chain-of-Thought prompting and Selection-Inference. These techniques search for proofs in the forward direction from axioms to the conclusion, which suffers from a combinatorial explosion of the search space, and thus high failure rates for problems requiring longer chains of reasoning. The classical automated reasoning literature has shown that reasoning in the backward direction (i.e. from intended conclusion to supporting axioms) is significantly more efficient at proof-finding. Importing this intuition into the LM setting, we develop a Backward Chaining algorithm, called LAMBADA, that decomposes reasoning into four sub-modules, that are simply implemented by few-shot prompted LLM inference. We show that LAMBADA achieves sizable accuracy boosts over state-of-the-art forward reasoning methods on two challenging logical reasoning datasets, particularly when deep and accurate proof chains are required.

Can LMs Learn New Entities from Descriptions? Challenges in Propagating Injected Knowledge
Yasumasa Onoe, Michael J.Q. Zhang, Shankar Padmanabhan, Greg Durrett and Eunsol Choi 16:15-17:45 (Frontenac Ballroom and Queen's Quay)

Pre-trained language models (LMs) are used for knowledge intensive tasks like question answering, but their knowledge gets continuously outdated as the world changes. Prior work has studied targeted updates to LMs, injecting individual facts and evaluating whether the model learns these facts while not changing predictions on other contexts. We take a step forward and study LMs' abilities to make inferences based on injected facts (or propagate those facts): for example, after learning that something is a TV show, does an LM predict that you can watch it? We study this with two cloze-style tasks: an existing dataset of real-world sentences about novel entities (ECBD) as well as a new controlled benchmark with manually designed templates requiring varying levels of inference about injected knowledge. Surprisingly, we find that existing methods for updating knowledge (gradient-based fine-tuning and modifications of this approach) show little propagation of injected knowledge. These methods improve performance on cloze instances only when there is lexical overlap between injected facts and target inferences. Yet, prepending entity definitions in an LM's context improves performance across all settings, suggesting that there is substantial headroom for parameter-updating approaches for knowledge injection.

On the Efficacy of Sampling Adapters
Clara Meister, Tiago Pimentel, Luca Malagutti, Ethan Wilcox and Ryan Cotterell   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Sampling-based decoding strategies are widely employed for generating text from probabilistic models, yet standard ancestral sampling often results in text that is degenerate or incoherent. To alleviate this issue, various modifications to a model's sampling distribution, such as top-p or top-k sampling, have been introduced and are now ubiquitously used in language generation systems. We propose a unified framework for understanding these techniques, which we term sampling adapters. Sampling adapters often lead to qualitatively better text, which raises the question: From a formal perspective, how are they changing the token-level distributions of language generation models? And why do these local changes lead to higher-quality text? We argue that the shift they enforce can be viewed as a trade-off between precision and recall: while the model loses its ability to produce certain strings, its precision rate on desirable text increases. While this trade-off is not reflected in standard metrics of distribution quality (such as perplexity), we find that several precision-emphasizing measures indeed indicate that sampling adapters can lead to probability distributions more aligned with the true distribution. Further, these measures correlate with higher sequence-level quality scores, specifically, Mauve.

Compositional Data Augmentation for Abstractive Conversation Summarization
Siru Ouyang, Jiaao Chen, Jiawei Han and Diyi Yang   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Recent abstractive conversation summarization systems generally rely on large-scale datasets with annotated summaries. However, collecting and annotating these conversations can be a time-consuming and labor-intensive task. To address this issue, in this work, we present a sub-structure level compositional data augmentation method, COMPO, for generating diverse and high-quality pairs of conversations and summaries. Specifically, COMPO first extracts conversation structures like topic splits and action triples as basic units. Then we organize these semantically meaningful conversation snippets compositionally to create new training instances. Additionally, we explore noise-tolerant settings in both self-training and joint-training paradigms to make the most of these augmented samples. Our experiments on benchmark datasets, SAMSum and DialogSum, show that COMPO substantially outperforms prior baseline methods by achieving a nearly 10% increase of ROUGE scores with limited data. Code is available at https://github.com/ozyysHr/Compo.

Exploring the Impact of Layer Normalization for Zero-shot Neural Machine Translation
Zhuoyuan Mao, Raj Dabre, Qianying Liu, Haiyue Song, Chenhui Chu and Sadao Kurohashi   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
This paper studies the impact of layer normalization (LayerNorm) on zero-shot translation (ZST). Recent efforts for ZST often utilize the Transformer architecture as the backbone, with LayerNorm at the input of layers (PreNorm) set as the default. However, Xu et al. (2019) has revealed that PreNorm carries the risk of overfitting the training data. Based on this, we hypothesize that PreNorm may overfit supervised directions and thus have low generalizability for ZST. Through experiments on OPUS, IWSLT, and Europarl datasets for 54 ZST directions, we demonstrate that the original Transformer setting of LayerNorm after residual connections (PostNorm) consistently outperforms PreNorm by up to 12.3 BLEU points. We then study the performance disparities by analyzing the differences in off-target rates and structural variations between PreNorm and PostNorm. This study highlights the need for careful consideration of the LayerNorm setting for ZST.

Subset Retrieval Nearest Neighbor Machine Translation
Hiroyuki Deguchi, Taro Watanabe, Yusuke Matsui, Masao Utiyama, Hideki Tanaka and Eiichiro Sumita   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
k-nearest-neighbor machine translation (kNN-MT) (Khandelwal et al., 2021) boosts the translation performance of trained neural machine translation (NMT) models by incorporating example-search into the decoding algorithm. However, decoding is seriously time-consuming, i.e., roughly 100 to 1,000 times slower than standard NMT, because neighbor tokens are retrieved from all target tokens of parallel data in each timestep. In this paper, we propose "Subset kNN-MT", which improves the decoding speed of kNN-MT by two methods: (1) retrieving neighbor target tokens from a subset that is the set of neighbor sentences of the input sentence, not from all sentences, and (2) efficient distance computation technique that is suitable for subset neighbor search using a look-up table. Our proposed method achieved a speed-up of up to 132.2 times and an improvement in BLEU score of up to 1.6 compared with kNN-MT in the WMT'19 De-En translation task and the domain adaptation tasks in De-En and En-Ja.

PromptNER: Prompt Locating and Typing for Named Entity Recognition
Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu and Yueting Zhuang 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Prompt learning is a new paradigm for utilizing pre-trained language models and has achieved great success in many tasks. To adopt prompt learning in the NER task, two kinds of methods have been explored from a pair of symmetric perspectives, populating the template by enumerating spans to predict their entity types or constructing type-specific prompts to locate entities. However, these methods not only require a multi-round prompting manner with a high time overhead and computational cost, but also require elaborate prompt templates, that are difficult to apply in practical scenarios. In this paper, we unify entity locating and entity typing into prompt learning, and design a dual-slot multi-prompt template with the position slot and type slot to prompt locating and typing respectively. Multiple prompts can be input to the model simultaneously, and then the model extracts all entities by parallel predictions on the slots. To assign labels for the slots during training, we design a dynamic template filling mechanism that uses the extended bipartite graph matching between prompts and the ground-truth entities. We conduct experiments in various settings, including resource-rich flat and nested NER datasets and low-resource in-domain and cross-domain datasets. Experimental results show that the proposed model achieves a significant performance improvement, especially in the cross-domain few-shot setting, which outperforms the state-of-the-art model by +7.7% on average.

ACLM: A Selective-Denoising based Generative Data Augmentation Approach for Low-Resource Complex NER
Sreyan Ghosh, Utkarsh Tyagi, Manan Suri, Sonal Kumar, Ramaneswaran S and Dinesh Manocha   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Complex Named Entity Recognition (NER) is the task of detecting linguistically complex named entities in low-context text. In this paper, we present ACLM Attention-map aware keyword selection for Conditional Language Model fine-tuning), a novel data augmentation approach based on conditional generation, to address the data scarcity problem in low-resource complex NER. ACLM alleviates the context-entity

mismatch issue, a problem existing NER data augmentation techniques suffer from and often generates incoherent augmentations by placing complex named entities in the wrong context. ACLM builds on BART and is optimized on a novel text reconstruction or denoising task - we use selective masking (aided by attention maps) to retain the named entities and certain keywords in the input sentence that provide contextually relevant additional knowledge or hints about the named entities. Compared with other data augmentation strategies, ACLM can generate more diverse and coherent augmentations preserving the true word sense of complex entities in the sentence. We demonstrate the effectiveness of ACLM both qualitatively and quantitatively on monolingual, cross-lingual, and multilingual complex NER across various low-resource settings. ACLM outperforms all our neural baselines by a significant margin (1%-36%). In addition, we demonstrate the application of ACLM to other domains that suffer from data scarcity (e.g., biomedical). In practice, ACLM generates more effective and factual augmentations for these domains than prior methods.

The Role of Global and Local Context in Named Entity Recognition
Arthur Amalvy, Vincent Labatut and Richard Dufour   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Pre-trained transformer-based models have recently shown great performance when applied to Named Entity Recognition (NER). As the complexity of their self-attention mechanism prevents them from processing long documents at once, these models are usually applied in a sequential fashion. Such an approach unfortunately only incorporates local context and prevents leveraging global document context in long documents such as novels, which might hinder performance. In this article, we explore the impact of global document context, and its relationships with local context. We find that correctly retrieving global document context has a greater impact on performance than only leveraging local context, prompting for further research on how to better retrieve that context.

Event Extraction as Question Generation and Answering
Di Lu, Shihao Ran, Joel Tetreault and Alejandro Jaimes   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Recent work on Event Extraction has reframed the task as Question Answering (QA), with promising results. The advantage of this approach is that it addresses the error propagation issue found in traditional token-based classification approaches by directly predicting event arguments without extracting candidates first. However, the questions are typically based on fixed templates and they rarely leverage contextual information such as relevant arguments. In addition, prior QA-based approaches have difficulty handling cases where there are multiple arguments for the same role. In this paper, we propose QGA-EE, which enables a Question Generation (QG) model to generate questions that incorporate rich contextual information instead of using fixed templates. We also propose dynamic templates to assist the training of QG model. Experiments show that QGA-EE outperforms all prior single-task-based models on the ACE05 English dataset.

RED^{FM}: a Filtered and Multilingual Relation Extraction Dataset
Pere Lluís Huguet Cabot, Simone Tedeschi, Axel-Cyrille Ngonga Ngomo and Roberto Navigli 16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Relation Extraction (RE) is a task that identifies relationships between entities in a text, enabling the acquisition of relational facts and bridging the gap between natural language and structured knowledge. However, current RE models often rely on small datasets with low coverage of relation types, particularly when working with languages other than English.

In this paper, we address the above issue and provide two new resources that enable the training and evaluation of multilingual RE systems. First, we present SRED^{FM}, an automatically annotated dataset covering 18 languages, 400 relation types, 13 entity types, totaling more than 40 million triplet instances. Second, we propose RED^{FM}, a smaller, human-revised dataset for seven languages that allows for the evaluation of multilingual RE systems. To demonstrate the utility of these novel datasets, we experiment with the first end-to-end multilingual RE model, mREBEL, that extracts triplets, including entity types, in multiple languages. We release our resources and model checkpoints at https://www.github.com/babelscape/rebel.

From the One, Judge of the Whole: Typed Entailment Graph Construction with Predicate Generation
Zhibin Chen, Yansong Feng and Dongyan Zhao   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Entailment Graphs (EGs) have been constructed based on extracted corpora as a strong and explainable form to indicate context-independent entailment relation in natural languages. However, EGs built by previous methods often suffer from the severe sparsity issues, due to limited corpora available and the long-tail phenomenon of predicate distributions. In this paper, we propose a multi-stage method, Typed Predicate-Entailment Graph Generator (TP-EGG), to tackle this problem. Given several seed predicates, TP-EGG builds the graphs by generating new predicates and detecting entailment relations among them. The generative nature of TP-EGG helps us leverage the recent advances from large pretrained language models (PLMs), while avoiding the reliance on carefully prepared corpora. Experiments on benchmark datasets show that TP-EGG can generate high-quality and scale-controllable entailment graphs, achieving significant in-domain improvement over state-of-the-art EGs and boosting the performance of down-stream inference tasks.

Reanalyzing L2 Preposition Learning with Bayesian Mixed Effects and a Pretrained Language Model
Jakob Prange and Man Ho Ivy Wong   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
We use both Bayesian and neural models to dissect a data set of Chinese learners' pre- and post-interventional responses to two tests measuring their understanding of English prepositions. The results mostly replicate previous findings from frequentist analyses and newly reveal crucial interactions between student ability, task type, and stimulus sentence. Given the sparsity of the data as well as high diversity among learners, the Bayesian method proves most useful; but we also see potential in using language model probabilities as predictors of grammaticality and learnability.

Unsupervised Selective Rationalization with Noise Injection
Adam Storek, Melanie Subbiah and Kathleen McKeown   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
A major issue with using deep learning models in sensitive applications is that they provide no explanation for their output. To address this problem, unsupervised selective rationalization produces rationales alongside predictions by chaining two jointly-trained components, a rationale generator and a predictor. Although this architecture guarantees that the prediction relies solely on the rationale, it does not ensure that the rationale contains a plausible explanation for the prediction. We introduce a novel training technique that effectively limits generation of implausible rationales by injecting noise between the generator and the predictor. Furthermore, we propose a new benchmark for evaluating unsupervised selective rationalization models using movie reviews from existing datasets. We achieve sizeable improvements in rationale plausibility and task accuracy over the state-of-the-art across a variety of tasks, including our new benchmark, while maintaining or improving model faithfulness.

Entity Tracking in Language Models
Najoung Kim and Sebastian Schuster   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
Keeping track of how states of entities change as a text or dialog unfolds is a key prerequisite to discourse understanding. Yet, there have been few systematic investigations into the ability of large language models (LLMs) to track discourse entities. In this work, we present a task probing to what extent a language model can infer the final state of an entity given an English description of the initial state and a series of state-changing operations. We use this task to first investigate whether Flan-T5, GPT-3 and GPT-3.5 can track the state of entities, and find that only GPT-3.5 models, which have been pretrained on large amounts of code, exhibit this ability. We then investigate whether smaller

models pretrained primarily on text can learn to track entities, through finetuning T5 on several training/evaluation splits. While performance degrades for more complex splits, we find that even when evaluated on a different set of entities from training or longer operation sequences, a finetuned model can perform non-trivial entity tracking. Taken together, these results suggest that language models can learn to track entities but pretraining on text corpora alone does not make this capacity surface.

Towards a Common Understanding of Contributing Factors for Cross-Lingual Transfer in Multilingual Language Models: A Review
Fred Philippy, Siwen Guo and Shohreh Haddadan   16:15-17:45 (Frontenac Ballroom and Queen's Quay)
In recent years, pre-trained Multilingual Language Models (MLLMs) have shown a strong ability to transfer knowledge across different languages. However, given that the aspiration for such an ability has not been explicitly incorporated in the design of the majority of MLLMs, it is challenging to obtain a unique and straightforward explanation for its emergence. In this review paper, we survey literature that investigates different factors contributing to the capacity of MLLMs to perform zero-shot cross-lingual transfer and subsequently outline and discuss these factors in detail. To enhance the structure of this review and to facilitate consolidation with future studies, we identify five categories of such factors. In addition to providing a summary of empirical evidence from past studies, we identify consensuses among studies with consistent findings and resolve conflicts among contradictory ones. Our work contextualizes and unifies existing research streams which aim at explaining the cross-lingual potential of MLLMs. This review provides, first, an aligned reference point for future research and, second, guidance for a better-informed and more efficient way of leveraging the cross-lingual capacity of MLLMs.

Semantics: Lexical

16:15-17:45 (Pier 2&3)

DimonGen: Diversified Generative Commonsense Reasoning for Explaining Concept Relationships
Chenzhengyi Liu, Jie Huang, Kerui Zhu and Kevin Chen-Chuan Chang   16:15-16:30 (Pier 2&3)
In this paper, we propose DimonGen, which aims to generate diverse sentences describing concept relationships in various everyday scenarios. To support this, we first create a benchmark dataset for this task by adapting the existing CommonGen dataset. We then propose a two-stage model called MoREE to generate the target sentences. MoREE consists of a mixture of retrievers model that retrieves diverse context sentences related to the given concepts, and a mixture of generators model that generates diverse sentences based on the retrieved contexts. We conduct experiments on the DimonGen task and show that MoREE outperforms strong baselines in terms of both the quality and diversity of the generated sentences. Our results demonstrate that MoREE is able to generate diverse sentences that reflect different relationships between concepts, leading to a comprehensive understanding of concept relationships.

Does GPT-3 Grasp Metaphors? Identifying Metaphor Mappings with Generative Language Models
Lennart Wachowiak and Dagmar Gromann   16:30-16:45 (Pier 2&3)
Conceptual metaphors present a powerful cognitive vehicle to transfer knowledge structures from a source to a target domain. Prior neural approaches focus on detecting whether natural language sequences are metaphoric or literal. We believe that to truly probe metaphoric knowledge in pre-trained language models, their capability to detect this transfer should be investigated. To this end, this paper proposes to probe the ability of GPT-3 to detect metaphoric language and predict the metaphor's source domain without any pre-set domains. We experiment with different training sample configurations for fine-tuning and few-shot prompting on two distinct datasets. When provided 12 few-shot samples in the prompt, GPT-3 generates the correct source domain for a new sample with an accuracy of 65.15% in English and 34.65% in Spanish. GPT's most common error is a hallucinated source domain for which no indicator is present in the sentence. Other common errors include identifying a sequence as literal even though a metaphor is present and predicting the wrong source domain based on specific words in the sequence that are not metaphorically related to the target domain.

Learning to Substitute Spans towards Improving Compositional Generalization
Zhaoyi Li, Ying Wei and Defu Lian   16:45-17:00 (Pier 2&3)
Despite the rising prevalence of neural sequence models, recent empirical evidences suggest their deficiency in compositional generalization. One of the current de-facto solutions to this problem is compositional data augmentation, aiming to incur additional compositional inductive bias. Nonetheless, the improvement offered by existing handcrafted augmentation strategies is limited when successful systematic generalization of neural sequence models requires multi-grained compositional bias (i.e., not limited to either lexical or structural biases only) or differentiation of training sequences in an imbalanced difficulty distribution. To address the two challenges, we first propose a novel compositional augmentation strategy dubbed Span Substitution (SpanSub) that enables multi-grained composition of substantial substructures in the whole training set. Over and above that, we introduce the Learning to Substitute Span (L2S2) framework which empowers the learning of span substitution probabilities in SpanSub in an end-to-end manner by maximizing the loss of neural sequence models, so as to outweigh those challenging compositions with elusive concepts and novel surroundings. Our empirical results on three standard compositional generalization benchmarks, including SCAN, COGS and GeoQuery (with an improvement of at most 66.5%, 10.3%, 1.2%, respectively), demonstrate the superiority of SpanSub, L2S2 and their combination.

LexSym: Compositionality as Lexical Symmetry
Ekin Akyurek and Jacob Andreas   17:00-17:15 (Pier 2&3)
In tasks like semantic parsing, instruction following, and question answering, standard deep networks fail to generalize compositionally from small datasets. Many existing approaches overcome this limitation with model architectures that enforce a compositional process of sentence interpretation. In this paper, we present a domain-general and model-agnostic formulation of compositionality as a constraint on symmetries of data distributions rather than models. Informally, we prove that whenever a task can be solved by a compositional model, there is a corresponding data augmentation scheme — a procedure for transforming examples into other well-formed examples — that imparts compositional inductive bias on any model trained to solve the same task. We describe a procedure called LexSym that discovers these transformations automatically, then applies them to training data for ordinary neural sequence models. Unlike existing compositional data augmentation procedures, LexSym can be deployed agnostically across text, structured data, and even images. It matches or surpasses state-of-the-art, task-specific models on COGS semantic parsing, SCAN and Alchemy instruction following, and CLEVR-CoGenT visual question answering datasets.

Interpretable Word Sense Representations via Definition Generation: The Case of Semantic Change Analysis
Mario Giulianelli, Iris Luden, Raquel Fernandez and Andrey Kutuzov   17:15-17:30 (Pier 2&3)
We propose using automatically generated natural language definitions of contextualised word usages as interpretable word and word sense representations. Given a collection of usage examples for a target word, and the corresponding data-driven usage clusters (i.e., word senses), a definition is generated for each usage with a specialised Flan-T5 language model, and the most prototypical definition in a usage cluster is

chosen as the sense label.

We demonstrate how the resulting sense labels can make existing approaches to semantic change analysis more interpretable, and how they can allow users — historical linguists, lexicographers, or social scientists — to explore and intuitively explain diachronic trajectories of word meaning. Semantic change analysis is only one of many possible applications of the 'definitions as representations' paradigm. Beyond being human-readable, contextualised definitions also outperform token or usage sentence embeddings in word-in-context semantic similarity judgements, making them a new promising type of lexical representation for NLP.

CLCL: Non-compositional Expression Detection with Contrastive Learning and Curriculum Learning
Jianing Zhou, Ziheng Zeng and Suma Bhat   17:30-17:45 (Pier 2&3)

Non-compositional expressions present a substantial challenge for natural language processing (NLP) systems, necessitating more intricate processing compared to general language tasks, even with large pre-trained language models. Their non-compositional nature and limited availability of data resources further compound the difficulties in accurately learning their representations. This paper addresses both of these challenges. By leveraging contrastive learning techniques to build improved representations it tackles the non-compositionality challenge. Additionally, we propose a dynamic curriculum learning framework specifically designed to take advantage of the scarce available data for modeling non-compositionality. Our framework employs an easy-to-hard learning strategy, progressively optimizing the model's performance by effectively utilizing available training data. Moreover, we integrate contrastive learning into the curriculum learning approach to maximize its benefits. Experimental results demonstrate the gradual improvement in the model's performance on idiom usage recognition and metaphor detection tasks. Our evaluation encompasses six datasets, consistently affirming the effectiveness of the proposed framework. Our models available at https://github.com/zhjjn/CLCL.git.

Linguistic Theories, Cognitive Modeling, and Psycholinguistics

16:15-17:45 (Pier 7&8)

[TACL] Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times?
Byung-Doh Oh and William Schuler   16:15-16:30 (Pier 7&8)

This work presents results using multiple large pretrained language models showing that models with more parameters and lower perplexity nonetheless yield surprisal estimates that are less predictive of human reading times, replicating and expanding upon earlier results limited to just GPT-2 (Oh et al., 2022). First, regression analyses show a strictly monotonic, positive log-linear relationship between perplexity and fit to reading times for the more recently released five GPT-Neo variants and eight OPT variants on two separate datasets, providing strong empirical support for this trend. Subsequently, analysis of residual errors reveals a systematic deviation of the larger variants, such as under-predicting reading times of named entities and making compensatory overpredictions for reading times of function words such as modals and conjunctions. These results suggest that the propensity of larger Transformer-based models to 'memorize' sequences during training makes their surprisal estimates diverge from humanlike expectations, which warrants caution in using pretrained language models to study human language processing.

Dependency resolution at the syntax-semantics interface: psycholinguistic and computational insights on control dependencies
Iria de-Dios-Flores, Juan Pablo Garcia Amboage and Marcos Garcia   16:30-16:45 (Pier 7&8)

Using psycholinguistic and computational experiments we compare the ability of humans and several pre-trained language models to correctly identify control dependencies in Spanish sentences such as 'José le prometió/ordenó a María ser ordenado/a' ('Joseph promised/ordered Mary to be tidy'). These structures underlie complex anaphoric and agreement relations at the interface of syntax and semantics, allowing us to study lexically-guided antecedent retrieval processes. Our results show that while humans correctly identify the (un)acceptability of the strings, language models often fail to identify the correct antecedent in non-adjacent dependencies, showing their reliance on linearity. Additional experiments on Galician reinforce these conclusions. Our findings are equally valuable for the evaluation of language models' ability to capture linguistic generalizations, as well as for psycholinguistic theories of anaphor resolution.

Cognitive Reframing of Negative Thoughts through Human-Language Model Interaction
Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, David Wadden, Khendra G Lucas, Adam Miner, Theresa Nguyen and Tim Althoff   16:45-17:00 (Pier 7&8)

A proven therapeutic technique to overcome negative thoughts is to replace them with a more hopeful "reframed thought." Although therapy can help people practice and learn this Cognitive Reframing of Negative Thoughts, clinician shortages and mental health stigma commonly limit people's access to therapy. In this paper, we conduct a human-centered study of how language models may assist people in reframing negative thoughts. Based on psychology literature, we define a framework of seven linguistic attributes that can be used to reframe a thought. We develop automated metrics to measure these attributes and validate them with expert judgements from mental health practitioners. We collect a dataset of 600 situations, thoughts and reframes from practitioners and use it to train a retrieval-enhanced in-context learning model that effectively generates reframed thoughts and controls their linguistic attributes. To investigate what constitutes a "high-quality" reframe, we conduct an IRB-approved randomized field study on a large mental health website with over 2,000 participants. Amongst other findings, we show that people prefer highly empathic or specific reframes, as opposed to reframes that are overly positive. Our findings provide key implications for the use of LMs to assist people in overcoming negative thoughts.

Exploring How Generative Adversarial Networks Learn Phonological Representations
Jingyi Chen and Micha Elsner   17:00-17:15 (Pier 7&8)

This paper explores how Generative Adversarial Networks (GANs) learn representations of phonological phenomena. We analyze how GANs encode contrastive and non-contrastive nasality in French and English vowels by applying the ciwGAN architecture (Begus, 2021). Begus claims that ciwGAN encodes linguistically meaningful representations with categorical variables in its latent space and manipulating the latent variables shows an almost one to one corresponding control of the phonological features in ciwGAN's generated outputs. However, our results show an interactive effect of latent variables on the features in the generated outputs, which suggests the learned representations in neural networks are different from the phonological representations proposed by linguists. On the other hand, ciwGAN is able to distinguish contrastive and noncontrastive features in English and French by encoding them differently. Comparing the performance of GANs learning from different languages results in a better understanding of what language specific features contribute to developing language specific phonological representations. We also discuss the role of training data frequencies in phonological feature learning.

A Method for Studying Semantic Construal in Grammatical Constructions with Interpretable Contextual Embedding Spaces
Gabriella Chronis, Kyle Mahowald and Katrin Erk   17:15-17:30 (Pier 7&8)

We study semantic construal in grammatical constructions using large language models. First, we project contextual word embeddings into three interpretable semantic spaces, each defined by a different set of psycholinguistic feature norms. We validate these interpretable spaces

and then use them to automatically derive semantic characterizations of lexical items in two grammatical constructions: nouns in subject or object position within the same sentence, and the AANN construction (e.g., 'a beautiful three days'). We show that a word in subject position is interpreted as more agentive than the very same word in object position, and that the nouns in the AANN construction are interpreted as more measurement-like than when in the canonical alternation. Our method can probe the distributional meaning of syntactic constructions at a templatic level, abstracted away from specific lexemes.

Main Conference: Wednesday, July 12, 2023

Session 6 - 09:00-10:30

NLP Applications

09:00-10:30 (Metropolitan East)

Enhancing Grammatical Error Correction Systems with Explanations
Yuejiao Fei, Leyang Cui, Sen Yang, Wai Lam, Zhenzhong Lan and Shuming Shi 09:00-09:15 (Metropolitan East)
Grammatical error correction systems improve written communication by detecting and correcting language mistakes. To help language learners better understand why the GEC system makes a certain correction, the causes of errors (evidence words) and the corresponding error types are two key factors. To enhance GEC systems with explanations, we introduce EXPECT, a large dataset annotated with evidence words and grammatical error types. We propose several baselines and anlysis to understand this task. Furthermore, human evaluation verifies our explainable GEC system's explanations can assist second-language learners in determining whether to accept a correction suggestion and in understanding the associated grammar rule.

PMAES: Prompt-mapping Contrastive Learning for Cross-prompt Automated Essay Scoring
Yuan Chen and Xia Li 09:15-09:30 (Metropolitan East)
Current cross-prompt automated essay scoring (AES) is a challenging task due to the large discrepancies between different prompts, such as different genres and expressions. The main goal of current cross-prompt AES systems is to learn enough shared features between the source and target prompts to grade well on the target prompt. However, because the features are captured based on the original prompt representation, they may be limited being directly extracted directly between essays. In fact, when the representations of two prompts are more similar, we can gain more shared features between them. Based on this motivation, in this paper, we propose a learning strategy called "prompt-mapping" to learn about more consistent representations of source and target prompts. In this way, we can obtain more shared features between the two prompts and use them to better represent the essays for the target prompt. Experimental results on the ASAP++ dataset demonstrate the effectiveness of our method. We also design experiments in different settings to show that our method can be applied in different scenarios. Our code is available at https://github.com/gdufsnlp/PMAES.

DARE: Towards Robust Text Explanations in Biomedical and Healthcare Applications
Adam Daniel Ivankay, Mattia Rigotti and Pascal Frossard 09:30-09:45 (Metropolitan East)
Along with the successful deployment of deep neural networks in several application domains, the need to unravel the black-box nature of these networks has seen a significant increase recently. Several methods have been introduced to provide insight into the inference process of deep neural networks. However, most of these explainability methods have been shown to be brittle in the face of adversarial perturbations of their inputs in the image and generic textual domain. In this work we show that this phenomenon extends to specific and important high stakes domains like biomedical datasets. In particular, we observe that the robustness of explanations should be characterized in terms of the accuracy of the explanation in linking a model's inputs and its decisions - faithfulness - and its relevance from the perspective of domain experts - plausibility. This is crucial to prevent explanations that are inaccurate but still look convincing in the context of the domain at hand. To this end, we show how to adapt current attribution robustness estimation methods to a given domain, so as to take into account domain-specific plausibility. This results in our DomainAdaptiveAREstimator (DARE) attribution robustness estimator, allowing us to properly characterize the domain-specific robustness of faithful explanations. Next, we provide two methods, adversarial training and FAR training, to mitigate the brittleness characterized by DARE, allowing us to train networks that display robust attributions. Finally, we empirically validate our methods with extensive experiments on three established biomedical benchmarks.

Injecting knowledge into language generation: a case study in auto-charting after-visit care instructions from medical dialogue
Maksim Eremeev, Ilya Valmianski, Xavier Amatriain and Anitha Kannan 09:45-10:00 (Metropolitan East)
Factual correctness is often the limiting factor in practical applications of natural language generation in high-stakes domains such as health-care. An essential requirement for maintaining factuality is the ability to deal with rare tokens. This paper focuses on rare tokens that appear in both the source and the reference sequences, and which, when missed during generation, decrease the factual correctness of the output text. For high-stake domains that are also knowledge-rich, we show how to use knowledge to (a) identify which rare tokens that appear in both source and reference are important and (b) uplift their conditional probability. We introduce the "utilization rate" that encodes knowledge and serves as a regularizer by maximizing the marginal probability of selected tokens. We present a study in a knowledge-rich domain of healthcare, where we tackle the problem of generating after-visit care instructions based on patient-doctor dialogues. We verify that, in our dataset, specific medical concepts with high utilization rates are underestimated by conventionally trained sequence-to-sequence models. We observe that correcting this with our approach to knowledge injection reduces the uncertainty of the model as well as improves factuality and coherence without negatively impacting fluency.

Byte-Level Grammatical Error Correction Using Synthetic and Curated Corpora
Svanhvít Lilja Ingólfsdóttir, Petur Orri Ragnarsson, Haukur Páll Jónsson, Haukur Barri Simonarson, Vilhjalmur Thorsteinsson and Vésteinn Snæbjarnarson 10:00-10:15 (Metropolitan East)
Grammatical error correction (GEC) is the task of correcting typos, spelling, punctuation and grammatical issues in text. Approaching the problem as a sequence-to-sequence task, we compare the use of a common subword unit vocabulary and byte-level encoding. Initial synthetic training data is created using an error-generating pipeline, and used for finetuning two subword-level models and one byte-level model. Models are then finetuned further on hand-corrected error corpora, including texts written by children, university students, dyslexic and second-language writers, and evaluated over different error types and error origins. We show that a byte-level model enables higher correction quality than a subword approach, not only for simple spelling errors, but also for more complex semantic, stylistic and grammatical issues. In particular, initial training on synthetic corpora followed by finetuning on a relatively small parallel corpus of real-world errors helps the byte-level model correct a wide range of commonly occurring errors. Our experiments are run for the Icelandic language but should hold for other similar languages, and in particular to morphologically rich ones.

Adaptive and Personalized Exercise Generation for Online Language Learning
Peng Cui and Mrinmaya Sachan 10:15-10:30 (Metropolitan East)
Adaptive learning aims to provide customized educational activities (e.g., exercises) to address individual learning needs. However, manual construction and delivery of such activities is a laborious process. Thus, in this paper, we study a novel task of adaptive and personalized

exercise generation for online language learning. To this end, we combine a knowledge tracing model that estimates each student's evolving knowledge states from their learning history and a controlled text generation model that generates exercise sentences based on the student's current estimated knowledge state and instructor requirements of desired properties (e.g., domain knowledge and difficulty). We train and evaluate our model on real-world learner interaction data from Duolingo and demonstrate that LMs guided by student states can generate superior exercises. Then, we discuss the potential use of our model in educational applications using various simulations. These simulations show that our model can adapt to students' individual abilities and can facilitate their learning efficiency by personalizing learning sequences.

Machine Learning for NLP

09:00-10:30 (Metropolitan Centre)

RankCSE: Unsupervised Sentence Representations Learning via Learning to Rank
Jiduan Liu, Jiahao Liu, Qifan Wang, Jingang Wang, Wei Wu, Yunsen Xian, Dongyan Zhao, Kai Chen and Rui Yan 09:00-09:15 (Metropolitan Centre)
Unsupervised sentence representation learning is one of the fundamental problems in natural language processing with various downstream applications. Recently, contrastive learning has been widely adopted which derives high-quality sentence representations by pulling similar semantics closer and pushing dissimilar ones away. However, these methods fail to capture the fine-grained ranking information among the sentences, where each sentence is only treated as either positive or negative. In many real-world scenarios, one needs to distinguish and rank the sentences based on their similarities to a query sentence, e.g., very relevant, moderate relevant, less relevant, irrelevant, etc. In this paper, we propose a novel approach, RankCSE, for unsupervised sentence representation learning, which incorporates ranking consistency and ranking distillation with contrastive learning into a unified framework. In particular, we learn semantically discriminative sentence representations by simultaneously ensuring ranking consistency between two representations with different dropout masks, and distilling listwise ranking knowledge from the teacher. An extensive set of experiments are conducted on both semantic textual similarity (STS) and transfer (TR) tasks. Experimental results demonstrate the superior performance of our approach over several state-of-the-art baselines.

Lifting the Curse of Capacity Gap in Distilling Language Models
Chen Zhang, Yang Yang, Jiahao Liu, Jingang Wang, Yunsen Xian, Benyou Wang and Dawei Song 09:15-09:30 (Metropolitan Centre)
Pretrained language models (LMs) have shown compelling performance on various downstream tasks, but unfortunately they require a tremendous amount of inference compute. Knowledge distillation finds a path to compress LMs to small ones with a teacher-student paradigm. However, when the capacity gap between the teacher and the student is large, a curse of capacity gap appears, invoking a deficiency in distilling LMs. While a few studies have been carried out to fill the gap, the curse is not yet well tackled. In this paper, we aim at lifting the curse of capacity gap via enlarging the capacity of the student without notably increasing the inference compute. Largely motivated by sparse activation regime of mixture of experts (MoE), we propose a mixture of minimal experts (MiniMoE), which imposes extra parameters to the student but introduces almost no additional inference compute. Experimental results on GLUE and CoNLL demonstrate the curse of capacity gap is lifted by the magic of MiniMoE to a large extent. MiniMoE also achieves the state-of-the-art performance at small FLOPs compared with a range of competitive baselines. With a compression rate as much as $\sim 50 \times$, MiniMoE preserves $\sim 95\%$ GLUE score of the teacher.

Consistency Regularization Training for Compositional Generalization
Yongjing Yin, Jiali Zeng, Yafu Li, Fandong Meng, Jie Zhou and Yue Zhang 09:30-09:45 (Metropolitan Centre)
Existing neural models have difficulty generalizing to unseen combinations of seen components. To achieve compositional generalization, models are required to consistently interpret (sub)expressions across contexts. Without modifying model architectures, we improve the capability of Transformer on compositional generalization through consistency regularization training, which promotes representation consistency across samples and prediction consistency for a single sample. Experimental results on semantic parsing and machine translation benchmarks empirically demonstrate the effectiveness and generality of our method. In addition, we find that the prediction consistency scores on in-distribution validation sets can be an alternative for evaluating models during training, when commonly-used metrics are not informative.

Graph-based Relation Mining for Context-free Out-of-vocabulary Word Embedding Learning
Ziran Liang, Yuyin Lu, HeGang Chen and Yanghui Rao 09:45-10:00 (Metropolitan Centre)
The out-of-vocabulary (OOV) words are difficult to represent while critical to the performance of embedding-based downstream models. Prior OOV word embedding learning methods failed to model complex word formation well. In this paper, we propose a novel graph-based relation mining method, namely GRM, for OOV word embedding learning. We first build a Word Relationship Graph (WRG) based on word formation and associate OOV words with their semantically relevant words, which can mine the relational information inside word structures. Subsequently, our GRM can infer high-quality embeddings for OOV words through passing and aggregating semantic attributes and relational information in the WRG, regardless of contextual richness. Extensive experiments demonstrate that our model significantly outperforms state-of-the-art baselines on both intrinsic and downstream tasks when faced with OOV words.

WhitenedCSE: Whitening-based Contrastive Learning of Sentence Embeddings
Wenjie Zhuo, Yifan Sun, Xiaohan Wang, Linchao Zhu and Yi Yang 10:15-10:30 (Metropolitan Centre)
This paper presents a whitening-based contrastive learning method for sentence embedding learning (WhitenedCSE), which combines contrastive learning with a novel shuffled group whitening. Generally, contrastive learning pulls distortions of a single sample (i.e., positive samples) close and push negative samples far away, correspondingly facilitating the alignment and uniformity in the feature space. A popular alternative to the "pushing" operation is whitening the feature space, which scatters all the samples for uniformity. Since the whitening and the contrastive learning have large redundancy w.r.t. the uniformity, they are usually used separately and do not easily work together. For the first time, this paper integrates whitening into the contrastive learning scheme and facilitates two benefits. 1) Better uniformity. We find that these two approaches are not totally redundant but actually have some complementarity due to different uniformity mechanism. 2) Better alignment. We randomly divide the feature into multiple groups along the channel axis and perform whitening independently within each group. By shuffling the group division, we derive multiple distortions of a single sample and thus increase the positive sample diversity. Consequently, using multiple positive samples with enhanced diversity further improves contrastive learning due to better alignment. Extensive experiments on seven semantic textual similarity tasks show our method achieves consistent improvement over the contrastive learning baseline and sets new states of the art, e.g., 78.78% (+2.53% based on BERT{pasted macro 'BA'}) Spearman correlation on STS tasks.

Machine Translation

09:00-10:30 (Metropolitan West)

Improving Translation Quality Estimation with Bias Mitigation

Hui Huang, Shuangzhi Wu, Kehai Chen, Hui Di, Muyun Yang and Tiejun Zhao   09:15-09:30 (Metropolitan West)

State-of-the-art translation Quality Estimation (QE) models are proven to be biased. More specifically, they over-rely on monolingual features while ignoring the bilingual semantic alignment. In this work, we propose a novel method to mitigate the bias of the QE model and improve estimation performance. Our method is based on the contrastive learning between clean and noisy sentence pairs. We first introduce noise to the target side of the parallel sentence pair, forming the negative samples. With the original parallel pairs as the positive sample, the QE model is contrastively trained to distinguish the positive samples from the negative ones. This objective is jointly trained with the regression-style quality estimation, so as to prevent the QE model from overfitting to monolingual features. Experiments on WMT QE evaluation datasets demonstrate that our method improves the estimation performance by a large margin while mitigating the bias.

Test-time Adaptation for Machine Translation Evaluation by Uncertainty Minimization

Runzhe Zhan, Xuebo Liu, Derek F. Wong, Cuilian Zhang, Lidia S. Chao and Min Zhang   09:30-09:45 (Metropolitan West)

The neural metrics recently received considerable attention from the research community in the automatic evaluation of machine translation. Unlike text-based metrics that have interpretable and consistent evaluation mechanisms for various data sources, the reliability of neural metrics in assessing out-of-distribution data remains a concern due to the disparity between training data and real-world data. This paper aims to address the inference bias of neural metrics through uncertainty minimization during test time, without requiring additional data. Our proposed method comprises three steps: uncertainty estimation, test-time adaptation, and inference. Specifically, the model employs the prediction uncertainty of the current data as a signal to update a small fraction of parameters during test time and subsequently refine the prediction through optimization. To validate our approach, we apply the proposed method to three representative models and conduct experiments on the WMT21 benchmarks. The results obtained from both in-domain and out-of-distribution evaluations consistently demonstrate improvements in correlation performance across different models. Furthermore, we provide evidence that the proposed method effectively reduces model uncertainty. The code is publicly available at https://github.com/NLP2CT/TaU.

Towards Higher Pareto Frontier in Multilingual Machine Translation

yichong huang, Xiaocheng Feng, Xinwei Geng, Baohang Li and Bing Qin   09:45-10:00 (Metropolitan West)

Multilingual neural machine translation has witnessed remarkable progress in recent years. However, the long-tailed distribution of multilingual corpora poses a challenge of Pareto optimization, i.e., optimizing for some languages may come at the cost of degrading the performance of others. Existing balancing training strategies are equivalent to a series of Pareto optimal solutions, which trade off on a Pareto frontierIn Pareto optimization, Pareto optimal solutions refer to solutions in which none of the objectives can be improved without sacrificing at least one of the other objectives. The set of all Pareto optimal solutions forms a Pareto frontier.. In this work, we propose a new training framework, Pareto Mutual Distillation (Pareto-MD), towards pushing the Pareto frontier outwards rather than making trade-offs. Specifically, Pareto-MD collaboratively trains two Pareto optimal solutions that favor different languages and allows them to learn from the strengths of each other via knowledge distillation. Furthermore, we introduce a novel strategy to enable stronger communication between Pareto optimal solutions and broaden the applicability of our approach. Experimental results on the widely-used WMT and TED datasets show that our method significantly pushes the Pareto frontier and outperforms baselines by up to +2.46 BLEUOur code will be released upon acceptance..

Causes and Cures for Interference in Multilingual Translation

Uri Shaham, Maha Elbayad, Vedanuj Goswami, Omer Levy and Shruti Bhosale   10:00-10:15 (Metropolitan West)

Multilingual machine translation models can benefit from synergy between different language pairs, but also suffer from interference. While there is a growing number of sophisticated methods that aim to eliminate interference, our understanding of interference as a phenomenon is still limited. This work identifies the main factors that contribute to interference in multilingual machine translation. Through systematic experimentation, we find that interference (or synergy) are primarily determined by model size, data size, and the proportion of each language pair within the total dataset. We observe that substantial interference occurs mainly when the model is very small with respect to the available training data, and that using standard transformer configurations with less than one billion parameters largely alleviates interference and promotes synergy. Moreover, we show that tuning the sampling temperature to control the proportion of each language pair in the data is key to balancing the amount of interference between low and high resource language pairs effectively, and can lead to superior performance overall.

Breeding Machine Translations: Evolutionary approach to survive and thrive in the world of automated evaluation

Josef Jon and Ondřej Bojar   10:15-10:30 (Metropolitan West)

We propose a genetic algorithm (GA) based method for modifying n-best lists produced by a machine translation (MT) system. Our method offers an innovative approach to improving MT quality and identifying weaknesses in evaluation metrics. Using common GA operations (mutation and crossover) on a list of hypotheses in combination with a fitness function (an arbitrary MT metric), we obtain novel and diverse outputs with high metric scores. With a combination of multiple MT metrics as the fitness function, the proposed method leads to an increase in translation quality as measured by other held-out automatic metrics. With a single metric (including popular ones such as COMET) as the fitness function, we find blind spots and flaws in the metric. This allows for an automated search for adversarial examples in an arbitrary metric, without prior assumptions on the form of such example. As a demonstration of the method, we create datasets of adversarial examples and use them to show that reference-free COMET is substantially less robust than the reference-based version.

Posters

09:00-10:30 (Frontenac Ballroom and Queen's Quay)

[CL] Onception: Active Learning with Expert Advice for Real World Machine Translation

Vânia Mendonça, Ricardo Rei, Luísa Coheur and Alberto Sardinha   09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Active learning can play an important role in low-resource settings (i.e., where annotated data is scarce), by selecting which instances may be more worthy to annotate. Most active learning approaches for Machine Translation assume the existence of a pool of sentences in a source language, and rely on human annotators to provide translations or post-edits, which can still be costly. In this article, we apply active learning to a real-world human-in-the-loop scenario in which we assume that: (1) the source sentences may not be readily available, but instead arrive in a stream; (2) the automatic translations receive feedback in the form of a rating, instead of a correct/edited translation, since the human-in-the-loop might be a user looking for a translation, but not be able to provide one. To tackle the challenge of deciding whether each incoming pair source–translations is worthy to query for human feedback, we resort to a number of stream-based active learning query strategies. Moreover, because we do not know in advance which query strategy will be the most adequate for a certain language pair and set of Machine Translation models, we propose to dynamically combine multiple strategies using prediction with expert advice. Our experiments

on different language pairs and feedback settings show that using active learning allows us to converge on the best Machine Translation systems with fewer human interactions. Furthermore, combining multiple strategies using prediction with expert advice outperforms several individual active learning strategies with even fewer interactions, particularly in partial feedback settings.

[TACL] Questions Are All You Need to Train a Dense Passage Retriever
Devendra Sachan, Mike Lewis, Dani Yogatama, Luke Zettlemoyer, Joelle Pineau and Manzil Zaheer   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We introduce ART, a new corpus-level autoencoding approach for training dense retrieval models that does not require any labeled training data. Dense retrieval is a central challenge for open-domain tasks, such as Open QA, where state-of-the-art methods typically require large supervised datasets with custom hard-negative mining and denoising of positive examples. ART, in contrast, only requires access to unpaired inputs and outputs (e.g. questions and potential answer passages). It uses a new passage retrieval autoencoding scheme, where (1) an input question is used to retrieve a set of evidence passages, and (2) the passages are then used to compute the probability of reconstructing the original question. Training for retrieval based on question reconstruction enables effective unsupervised learning of both passage and question encoders, which can be later incorporated into complete Open QA systems without any further finetuning. Extensive experiments demonstrate that ART obtains state-of-the-art results on multiple QA retrieval benchmarks with only generic initialization from a pretrained language model, removing the need for labeled data and task-specific losses.

[TACL] Robust Dialogue State Tracking with Weak Supervision and Sparse Data
Michael Heck, Nurul Lubis, Carel van Niekerk, Shutong Feng, Christian Geishauser, Hsien-Chin Lin and Milica Gašić   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Generalising dialogue state tracking (DST) to new data is especially challenging due to the strong reliance on abundant and fine-grained supervision during training. Sample sparsity, distributional shift and the occurrence of new concepts and topics frequently lead to severe performance degradation during inference. In this paper we propose a training strategy to build extractive DST models without the need for fine-grained manual span labels. Two novel input-level dropout methods mitigate the negative impact of sample sparsity. We propose a new model architecture with a unified encoder that supports value as well as slot independence by leveraging the attention mechanism. We combine the strengths of triple copy strategy DST and value matching to benefit from complementary predictions without violating the principle of ontology independence. Our experiments demonstrate that an extractive DST model can be trained without manual span labels. Our architecture and training strategies improve robustness towards sample sparsity, new concepts and topics, leading to state-of-the-art performance on a range of benchmarks. We further highlight our model's ability to effectively learn from non-dialogue data.

[TACL] Sub-Character Tokenization for Chinese Pretrained Language Models
*Chenglei Si, Zhengyan Zhang, Yingfa Chen, Fanchao Qi, Xiaozhi Wang, Zhiyuan Liu, Yasheng Wang, Qun Liu and Maosong Sun*09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Tokenization is fundamental to pretrained language models (PLMs). Existing tokenization methods for Chinese PLMs typically treat each character as an indivisible token. However, they ignore the unique feature of the Chinese writing system where additional linguistic information exists below the character level, i.e., at the sub-character level. To utilize such information, we propose sub-character (SubChar for short) tokenization. Specifically, we first encode the input text by converting each Chinese character into a short sequence based on its glyph or pronunciation, and then construct the vocabulary based on the encoded text with sub-word segmentation. Experimental results show that Sub-Char tokenizers have two main advantages over existing tokenizers: 1) They can tokenize inputs into much shorter sequences, thus improving the computational efficiency. 2) Pronunciation-based SubChar tokenizers can encode Chinese homophones into the same transliteration sequences and produce the same tokenization output, hence being robust to homophone typos. At the same time, models trained with SubChar tokenizers perform competitively on downstream tasks. We release our code and models at https://github.com/thunlp/SubCharTokenization to facilitate future work.

[TACL] Evaluating Transformer Models and Human Behaviors on Chinese Character Naming
Xiaomeng Ma and Lingyu Gao   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Neural network models have been proposedto explain the grapheme-phoneme mappingprocess in humans for many alphabet languages.These models not only successfullylearned the correspondence of the letterstrings and their pronunciation, but alsocaptured human behavior in nonce wordnaming tasks. How would the neural modelsperform for a non-alphabet language (Chinese)unknown character task? How wellwould the model capture human behavior?In this study, we evaluate a set of transformermodels and compare their performanceswith human behaviors on unknownChinese character naming task. We foundthat the models and humans behaved verysimilarly that they have similar accuracy distributionfor each character and have a substantialoverlap in answers. In addition, themodels' answers are highly correlated withhumans' answers. These results suggestedthat the transformer models can well capturehuman's character naming behavior.

[TACL] Domain-Specific Word Embeddings with Structure Prediction
David Lassner, Anne Baillot, Shinichi Nakajima and Stephanie Brandl   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Complementary to finding good general word embeddings, an important question for representation learning is to find dynamic word embeddings, e.g., across time or domain. Current methods do not offer a way to use or predict information on structure between sub-corpora, time or domain and dynamic embeddings can only be compared after post-alignment. We propose novel word embedding methods that provide general word representations for the whole corpus, domain-specific representations for each sub-corpus, sub-corpus structure, and embedding alignment simultaneously. We present an empirical evaluation on New York Times articles and two English Wikipedia datasets with articles on science and philosophy. Our method, called Word2Vec with Structure Prediction (W2VPred), provides better performance than baselines in terms of the general analogy tests, domain-specific analogy tests, and multiple specific word embedding evaluations as well as structure prediction performance when no structure is given a priori. As a use case in the field of Digital Humanities we demonstrate how to raise novel research questions for high literature from the German Text Archive.

[TACL] Transparency Helps Reveal When Language Models Learn Meaning
Zhaofeng Wu, William Merrill, Hao Peng, Iz Beltagy and Noah Smith   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Many current NLP systems are built from language models trained to optimize unsupervised objectives on large amounts of raw text. Under what conditions might such a procedure acquire meaning? Our systematic experiments with synthetic data reveal that, with languages where all expressions have context-independent denotations (i.e., languages with strong transparency), both autoregressive and masked language models successfully learn to emulate semantic relations between expressions.However, when denotations are changed to be context-dependent with the language otherwise unmodified, this ability degrades. Turning to natural language, our experiments with a specific phenomenon – referential opacity – add to the growing body of evidence that current language models do not represent natural language semantics well. We show this failure relates to the context-dependent nature of natural language form-meaning mappings.

[SRW] Gender Stereotyping in Popular Children's Videos
Tiasa Singha Roy, Mallikarjuna Tupakula, Sumeet Kumar and Ashiqur Khudabukhsh   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Using the top 100 YouTube Kids Channel

[SRW] The Turing Quest: Can Transformers Make Good NPCs?
Qi Chen Gao and Ali Emami   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We explored the generation of NPC dialogue using a zero-shot prompting method as well as the ability of LMs to self-evaluate and score dialogue with few-shot learning.

[SRW] Making the Most Out of the Limited Context Length: Predictive Power Varies with Clinical Note Type and Note Section
Hongyi Zheng, Yixin Zhu, Lavender Jiang, Kyunghyun Cho and Eric Oermann   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We propose a data-driven framework to select clinical note sections with high predictive power.

[SRW] Intriguing Effect of the Correlation Prior on ICD-9 Code Assignment
Zihao Yang, Chenkang Zhang, Muru Wu, Xujin Liu, Lavender Jiang, Kyunghyun Cho and Eric Oermann 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
This paper investigates the usefulness of correlation bias in improving language models' performance on predicting imbalanced clinical codes from discharge summaries.

[SRW] Can LMs Store and Retrieve 1-to-N Relational Knowledge?
Haruki Nagasawa, Benjamin Heinzerling, Kazuma Kokuta and Kentaro Inui   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Our study aimed to explore the feasibility of using LMs as KBs, and we focused specifically on 1-to-N relational knowledge, an area that has not been extensively researched, and proposed a comprehensive approach that involved identifying the unique characteristics of this type of knowledge, designing appropriate training methods, and developing evaluation perspectives.

[SRW] MedTem2.0: Prompt-based Temporal Classification of Treatment Events from Discharge Summaries
Yang Cui, Lifeng Han and Goran Nenadic   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We use Prompt-based learning on LLMs for Temporal Classification of Treatment Events from Discharge Summaries of clinical data.

[SRW] Building a Buzzer-quiz Answering System
*Naoya Sugiura, Kosuke Yamada, Ryohei Sasano, Koichi Takeda and Katsuhiko Toyama*09:00-10:30 (Frontenac Ballroom and Queen's Quay)
This paper presents two types of buzzer-quiz answering systems that can predict the answer from only part of a question and then proposes a method to estimate the accuracy of the answers for each system by using the internal scores of each model.

[SRW] Second Language Acquisition of Neural Language Models
Miyu Oba, Tatsuki Kuribayashi, Hiroki Ouchi and Taro Watanabe   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We trained bilingual LMs with a scenario similar to human L2 acquisition and analyzed their cross-lingual transfer from linguistic perspectives.

[SRW] Semantic Accuracy in Natural Language Generation: A Thesis Proposal
Patricia Schmidtova   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We propose a thesis in which we explore how evaluation and interpretability techniques could lead to better natural language generation systems.

[SRW] Towards Reliable Misinformation Mitigation: Generalization, Uncertainty, and GPT-4
Kellin Pelrine, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph and Reihaneh Rabbany   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Framework towards better real-world misinformation detection through investigation of generalization, soft classification, and GPT-4.

[Demo] PrimeQA: The Prime Repository for State-of-the-Art Multilingual Question Answering Research and Development
Salim Roukos, Radu Florian, Juergen Bross, Riyaz Bhat, Md Arafat Sultan, Yulong Li, Vishwajeet Kumar, Rong Zhang, Scott McCarley, Sara Rosenthal, Mihaela Bornea, Kshitij Fadnis, Martin Franz, Bhavani Iyer, Jaydeep Sen and Avi Sil   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
The field of Question Answering (QA) has made remarkable progress in recent years, thanks to the advent of large pre-trained language models, newer realistic benchmark datasets with leaderboards, and novel algorithms for key components such as retrievers and readers. In this paper, we introduce PrimeQA: a one-stop and open-source QA repository with an aim to democratize QA research and facilitate easy replication of state-of-the-art (SOTA) QA methods. PrimeQA supports core QA functionalities like retrieval and reading comprehension as well as auxiliary capabilities such as question generation. It has been designed as an end-to-end toolkit for various use cases: building front-end applications, replicating SOTA methods on public benchmarks, and expanding pre-existing methods. PrimeQA is available at: https://github.com/primeqa.

[Demo] A Practical Toolkit for Multilingual Question and Answer Generation
Jose Camacho-Collados, Fernando Alva-Manchego and Asahi Ushio   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Generating questions along with associated answers from a text has applications in several domains, such as creating reading comprehension tests for students, or improving document search by providing auxiliary questions and answers based on the query. Training models for question and answer generation (QAG) is not straightforward due to the expected structured output (i.e. a list of question and answer pairs), as it requires more than generating a single sentence. This results in a small number of publicly accessible QAG models. In this paper, we introduce AutoQG, an online service for multilingual QAG along with `lmqg`, an all-in-one python package for model fine-tuning, generation, and evaluation. We also release QAG models in eight languages fine-tuned on a few variants of pre-trained encoder-decoder language models, which can be used online via AutoQG or locally via `lmqg`. With these resources, practitioners of any level can benefit from a toolkit that includes a web interface for end users, and easy-to-use code for developers who require custom models or fine-grained controls for generation.

[Demo] LIDA: A Tool for Automatic Generation of Grammar-Agnostic Visualizations and Infographics using Large Language Models
Victor Dibia   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Systems that support users in the automatic creation of visualizations must address several subtasks - understand the semantics of data, enumerate relevant visualization goals and generate visualization specifications. In this work, we pose visualization generation as a multi-stage generation problem and argue that well-orchestrated pipelines based on large language models (LLMs) and image generation models (IGMs) are suitable to addressing these tasks. We present LIDA, a novel tool for generating grammar-agnostic visualizations and infographics. LIDA comprises of 4 modules - A SUMMARIZER that converts data into a rich but compact natural language summary, a GOAL EXPLORER that enumerates visualization goals given the data, a VISGENERATOR that generates, refines, executes and filters visualization code and an INFOGRAPHER module that yields data-faithful stylized graphics using IGMs. LIDA provides a python api, and a hybrid user interface

(direct manipulation and multilingual natural language) for interactive chart, infographics and data story generation. Code and demo are available at this url - https://microsoft.github.io/lida/

[Demo] XMD: An End-to-End Framework for Interactive Explanation-Based Debugging of NLP Models
Xiang Ren, Jay Pujara, Toshiyuki Sekiya, Ryosuke Mitani, Takashi Shibuya, Kiran Narahari, Ziyi Liu, Aaron Chan, Brihi Joshi, Akshen Kadakia and Dong-Ho Lee   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
NLP models are susceptible to learning spurious biases (i.e., bugs) that work on some datasets but do not properly reflect the underlying task. Explanation-based model debugging aims to resolve spurious biases by showing human users explanations of model behavior, asking users to give feedback based on the behavior, then using the feedback to update the model. While existing model debugging methods have shown promise, their prototype-level implementations provide limited practical utility. Thus, we propose XMD: the first open-source, end-to-end framework for explanation-based model debugging. Given task- or instance-level explanations, users can flexibly provide various forms of feedback via an intuitive, web-based UI. After receiving user feedback, XMD automatically updates the model in real time, by regularizing the model so that its explanations align with the user feedback. The new model can then be easily deployed into real-world applications via Hugging Face. Using XMD, we can improve the model's OOD performance on text classification tasks by up to 18

[Demo] The OPUS-MT Dashboard – A Toolkit for a Systematic Evaluation of Open Machine Translation Models
Ona de Gibert and Jörg Tiedemann   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
The OPUS-MT dashboard is a web-based platform that provides a comprehensive overview of open translation models. We focus on a systematic collection of benchmark results with verifiable translation performance and large coverage in terms of languages and domains. We provide results for in-house OPUS-MT and Tatoeba models as well as external models from the Huggingface repository and user-contributed translations. The functionalities of the evaluation tool include summaries of benchmarks for over 2,300 models covering 4,560 language directions and 294 languages, as well as the inspection of predicted translations against their human reference. We focus on centralization, reproducibility and coverage of MT evaluation combined with scalability. The dashboard can be accessed live at https://opus.nlpl.eu/dashboard/.

[Demo] Petals: Collaborative Inference and Fine-tuning of Large Models
Colin Raffel, Pavel Samygin, Artem Chumachenko, Younes Belkada, Maksim Riabinin, Tim Dettmers, Dmitry Baranchuk and Alexander Borzunov   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Many NLP tasks benefit from using large language models (LLMs) that often have more than 100 billion parameters. With the release of BLOOM-176B and OPT-175B, everyone can download pretrained models of this scale. Still, using these models requires high-end hardware unavailable to many researchers. In some cases, LLMs can be used more affordably via RAM offloading or hosted APIs. However, these techniques have innate limitations: offloading is too slow for interactive inference, while APIs are not flexible enough for research that requires access to weights, attention or logits. In this work, we propose Petals - a system for inference and fine-tuning of large models collaboratively by joining the resources of multiple parties. We demonstrate that this strategy outperforms offloading for very large models, running inference of BLOOM-176B on consumer GPUs with 1 step per second, which is enough for many interactive LLM applications. Unlike most inference APIs, Petals also natively exposes hidden states of served models, allowing to train and share custom model extensions based on efficient fine-tuning methods. The system, its source code, and documentation are available at https://petals.ml

Video (2 min): https://youtu.be/F4muLI-0hTE

[Demo] UKP-SQuARE v3: A Platform for Multi-Agent QA Research
Iryna Gurevych, Kexin Wang, Sewin Tariverdian, Hao Zhang, Haishuo Fang, Rachneet Sachdeva, Tim Baumgärtner and Haritz Puerto 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
The continuous development of Question Answering (QA) datasets has drawn the research community's attention toward multi-domain models. A popular approach is to use multi-dataset models, which are models trained on multiple datasets to learn their regularities and prevent overfitting to a single dataset. However, with the proliferation of QA models in online repositories such as GitHub or Hugging Face, an alternative is becoming viable. Recent works have demonstrated that combining expert agents can yield large performance gains over multi-dataset models. To ease research in multi-agent models, we extend UKP-SQuARE, an online platform for QA research, to support three families of multi-agent systems: i) agent selection, ii) early-fusion of agents, and iii) late-fusion of agents. We conduct experiments to evaluate their inference speed and discuss the performance vs. speed trade-off compared to multi-dataset models. UKP-SQuARE is open-source and publicly available.

bgGLUE: A Bulgarian General Language Understanding Evaluation Benchmark
Momchil Hardalov, Pepa Atanasova, Todor Mihaylov, Galia Angelova, Kiril Simov, Petya Osenova, Veselin Stoyanov, Ivan K. Koychev, Preslav Nakov and Dragomir Radev   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We present bgGLUE (Bulgarian General Language Understanding Evaluation), a benchmark for evaluating language models on Natural Language Understanding (NLU) tasks in Bulgarian. Our benchmark includes NLU tasks targeting a variety of NLP problems (e.g., natural language inference, fact-checking, named entity recognition, sentiment analysis, question answering, etc.) and machine learning tasks (sequence labeling, document-level classification, and regression). We run the first systematic evaluation of pre-trained language models for Bulgarian, comparing and contrasting results across the nine tasks in the benchmark. The evaluation results show strong performance on sequence labeling tasks, but there is a lot of room for improvement for tasks that require more complex reasoning. We make bgGLUE publicly available together with the fine-tuning and the evaluation code, as well as a public leaderboard at https://bgglue.github.io, and we hope that it will enable further advancements in developing NLU models for Bulgarian.

Can Large Language Models Be an Alternative to Human Evaluations?
Cheng-Han Chiang and Hung-yi Lee   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Human evaluation is indispensable and inevitable for assessing the quality of texts generated by machine learning models or written by humans. However, human evaluation is very difficult to reproduce and its quality is notoriously unstable, hindering fair comparisons among different natural language processing (NLP) models and algorithms. Recently, large language models (LLMs) have demonstrated exceptional performance on unseen tasks when only the task instructions are provided. In this paper, we explore if such an ability of the LLMs can be used as an alternative to human evaluation. We present the LLMs with the exact same instructions, samples to be evaluated, and questions used to conduct human evaluation, and then ask the LLMs to generate responses to those questions; we dub this LLM evaluation. We use human evaluation and LLM evaluation to evaluate the texts in two NLP tasks: open-ended story generation and adversarial attacks. We show that the result of LLM evaluation is consistent with the results obtained by expert human evaluation: the texts rated higher by human experts are also rated higher by the LLMs. We also find that the results of LLM evaluation are stable over different formatting of the task instructions and the sampling algorithm used to generate the answer. We are the first to show the potential of using LLMs to assess the quality of texts and discuss the limitations and ethical considerations of LLM evaluation.

BIG-C: a Multimodal Multi-Purpose Dataset for Bemba
Claytone Sikasote, Eunice Mukonde, Md Mahfuz Ibn Alam and Antonios Anastasopoulos   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We present BIG-C (Bemba Image Grounded Conversations), a large multimodal dataset for Bemba. While Bemba is the most populous

language of Zambia, it exhibits a dearth of resources which render the development of language technologies or language processing research almost impossible. The dataset is comprised of multi-turn dialogues between Bemba speakers based on images, transcribed and translated into English. There are more than 92,000 utterances/sentences, amounting to more than 180 hours of audio data with corresponding transcriptions and English translations. We also provide baselines on speech recognition (ASR), machine translation (MT) and speech translation (ST) tasks, and sketch out other potential future multimodal uses of our dataset. We hope that by making the dataset available to the research community, this work will foster research and encourage collaboration across the language, speech, and vision communities especially for languages outside the "traditionally" used high-resourced ones. All data and code are publicly available: https://github.com/csikasote/bigc.

Evaluating Zero-Shot Event Structures: Recommendations for Automatic Content Extraction (ACE) Annotations
Erica Cai and Brendan O'Connor   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Zero-shot event extraction (EE) methods infer richly structured event records from text, based only on a minimal user specification and no training examples, which enables flexibility in exploring and developing applications. Most event extraction research uses the Automatic Content Extraction (ACE) annotated dataset to evaluate supervised EE methods, but can it be used to evaluate zero-shot and other low-supervision EE? We describe ACE's event structures and identify significant ambiguities and issues in current evaluation practice, including (1) coreferent argument mentions, (2) conflicting argument head conventions, and (3) ignorance of modality and event class details. By sometimes mishandling these subtleties, current work may dramatically understate the actual performance of zero-shot and other low-supervision EE, considering up to 32% of correctly identified arguments and 25% of correctly ignored event mentions as false negatives. For each issue, we propose recommendations for future evaluations so the research community can better utilize ACE as an event evaluation resource.

StoryWars: A Dataset and Instruction Tuning Baselines for Collaborative Story Understanding and Generation
Yulun Du and Lydia Chilton   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Collaborative stories, which are texts created through the collaborative efforts of multiple authors with different writing styles and intentions, pose unique challenges for NLP models. Understanding and generating such stories remains an underexplored area due to the lack of open-domain corpora. To address this, we introduce StoryWars, a new dataset of over 40,000 collaborative stories written by 9,400 different authors from an online platform. We design 12 task types, comprising 7 understanding and 5 generation task types, on {pasted macro 'STORYWARS'}, deriving 101 diverse story-related tasks in total as a multi-task benchmark covering all fully-supervised, few-shot, and zero-shot scenarios. Furthermore, we present our instruction-tuned model, InstructStory, for the story tasks showing that instruction tuning, in addition to achieving superior results in zero-shot and few-shot scenarios, can also obtain the best performance on the fully-supervised tasks in StoryWars, establishing strong multi-task benchmark performances on StoryWars.

Is GPT-3 a Good Data Annotator?
BOSHENG DING, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty and Lidong Bing   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Data annotation is the process of labeling data that could be used to train machine learning models. Having high quality annotation is crucial, as it allows the model to learn the relationship between the input data and the desired output. GPT-3, a large-scale language model developed by OpenAI, has demonstrated im- impressive zero- and few-shot performance on a wide range of NLP tasks. It is therefore natural to wonder whether it can be used to effectively annotate data for NLP tasks. In this paper, we evaluate the performance of GPT-3 as a data annotator by comparing it with traditional data annotation methods and analyzing its output on a range of tasks. Through this analysis, we aim to provide insight into the potential of GPT-3 as a general-purpose data annotator in NLP.

On Complementarity Objectives for Hybrid Retrieval
Dohyeon Lee, Seung-won Hwang, Kyungjae Lee, Seungtaek Choi and Sunghyun Park   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Dense retrieval has shown promising results in various information retrieval tasks, and hybrid retrieval, combined with the strength of sparse retrieval, has also been actively studied. A key challenge in hybrid retrieval is to make sparse and dense complementary to each other. Existing models have focused on dense models to capture "residual" features neglected in the sparse models. Our key distinction is to show how this notion of residual complementarity is limited, and propose a new objective, denoted as RoC (Ratio of Complementarity), which captures a fuller notion of complementarity. We propose a two-level orthogonality designed to improve RoC, then show that the improved RoC of our model, in turn, improves the performance of hybrid retrieval. Our method outperforms all state-of-the-art methods on three representative IR benchmarks: MSMARCO-Passage, Natural Questions, and TREC Robust04, with statistical significance. Our finding is also consistent in various adversarial settings.

DisorBERT: A Double Domain Adaptation Model for Detecting Signs of Mental Disorders in Social Media
Mario Ezra Aragon, Adrian Pastor Lopez Monroy, Luis C Gonzalez, David E. Losada and Manuel Montes 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Mental disorders affect millions of people worldwide and cause interference with their thinking and behavior. Through the past years, awareness created by health campaigns and other sources motivated the study of these disorders using information extracted from social media platforms. In this work, we aim to contribute to the study of these disorders and to the understanding of how mental problems reflect on social media. To achieve this goal, we propose a double-domain adaptation of a language model. First, we adapted the model to social media language, and then, we adapted it to the mental health domain. In both steps, we incorporated a lexical resource to guide the masking process of the language model and, therefore, to help it in paying more attention to words related to mental disorders. We have evaluated our model in the detection of signs of three major mental disorders: Anorexia, Self-harm, and Depression. Results are encouraging as they show that the proposed adaptation enhances the classification performance and yields competitive results against state-of-the-art methods.

ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis
Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo and Yu Xu   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Multimodal Sentiment Analysis aims to predict the sentiment of video content. Recent research suggests that multimodal sentiment analysis critically depends on learning a good representation of multimodal information, which should contain both modality-invariant representations that are consistent across modalities as well as modality-specific representations. In this paper, we propose ConFEDE, a unified learning framework that jointly performs contrastive representation learning and contrastive feature decomposition to enhance the representation of multimodal information. It decomposes each of the three modalities of a video sample, including text, video frames, and audio, into a similarity feature and a dissimilarity feature, which are learned by a contrastive relation centered around the text. We conducted extensive experiments on CH-SIMS, MOSI and MOSEI to evaluate various state-of-the-art multimodal sentiment analysis methods. Experimental results show that ConFEDE outperforms all baselines on these datasets on a range of metrics.

Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis
Agam Shah, Suvan Satya Paturi and Sudheer Chava   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Monetary policy pronouncements by Federal Open Market Committee (FOMC) are a major driver of financial market returns. We construct the largest tokenized and annotated dataset of FOMC speeches, meeting minutes, and press conference transcripts in order to understand

how monetary policy influences financial markets. In this study, we develop a novel task of hawkish-dovish classification and benchmark various pre-trained language models on the proposed dataset. Using the best-performing model (RoBERTa-large), we construct a measure of monetary policy stance for the FOMC document release days. To evaluate the constructed measure, we study its impact on the treasury market, stock market, and macroeconomic indicators. Our dataset, models, and code are publicly available on Huggingface and GitHub under CC BY-NC 4.0 license.

Cross-Modal Attribute Insertions for Assessing the Robustness of Vision-and-Language Learning
Shivaen Ramshetty, Gaurav Verma and Srijan Kumar   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
The robustness of multimodal deep learning models to realistic changes in the input text is critical for applicability on important tasks such as text-to-image retrieval and cross-modal entailment. To measure robustness, several existing approaches edit the text data, but without leveraging the cross-modal information present in multimodal data. Such information from the visual modality, such as color, size, and shape, provides additional attributes that users can include in their inputs. Thus, we propose cross-modal attribute insertions as a realistic perturbation strategy for vision-and-language data that inserts visual attributes of the objects in the image into the corresponding text (e.g., "girl on a chair" to "little girl on a wooden chair"). Our proposed approach for cross-modal attribute insertions is modular, controllable, and task-agnostic. We find that augmenting input text using cross-modal insertions causes state-of-the-art approaches for text-to-image retrieval and cross-modal entailment to perform poorly, resulting in relative drops of 15% in MRR and 20% in F1 score, respectively. Crowd-sourced annotations demonstrate that cross-modal insertions lead to higher quality augmentations for multimodal data than augmentations using text-only data, and are equivalent in quality to original examples. We release the code to encourage robustness evaluations of deep vision-and-language models: https://github.com/claws-lab/multimodal-robustness-xmai

DataFinder: Scientific Dataset Recommendation from Natural Language Descriptions
Vijay Viswanathan, Luyu Gao, Tongshuang Wu, Pengfei Liu and Graham Neubig   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Modern machine learning relies on datasets to develop and validate research ideas. Given the growth of publicly available data, finding the right dataset to use is increasingly difficult. Any research question imposes explicit and implicit constraints on how well a given dataset will enable researchers to answer this question, such as dataset size, modality, and domain. We operationalize the task of recommending datasets given a short natural language description of a research idea, to help people find relevant datasets for their needs. Dataset recommendation poses unique challenges as an information retrieval problem; datasets are hard to directly index for search and there are no corpora readily available for this task. To facilitate this task, we build the DataFinder Dataset which consists of a larger automatically-constructed training set (17.5K queries) and a smaller expert-annotated evaluation set (392 queries). Using this data, we compare various information retrieval algorithms on our test set and present a superior bi-encoder retriever for text-based dataset recommendation. This system, trained on the DataFinder Dataset, finds more relevant search results than existing third-party dataset search engines. To encourage progress on dataset recommendation, we release our dataset and models to the public.

Curriculum Learning for Graph Neural Networks: A Multiview Competence-based Approach
Nidhi Vakil and Hadi Amiri   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
A curriculum is a planned sequence of learning materials and an effective one can make learning efficient and effective for both humans and machines. Recent studies developed effective data-driven curriculum learning approaches for training graph neural networks in language applications. However, existing curriculum learning approaches often employ a single criterion of difficulty in their training paradigms. In this paper, we propose a new perspective on curriculum learning by introducing a novel approach that builds on graph complexity formalisms (as difficulty criteria) and model competence during training. The model consists of a scheduling scheme which derives effective curricula by accounting for different views of sample difficulty and model competence during training. The proposed solution advances existing research in curriculum learning for graph neural networks with the ability to incorporate a fine-grained spectrum of graph difficulty criteria in their training paradigms. Experimental results on real-world link prediction and node classification tasks illustrate the effectiveness of the proposed approach.

Multitask Pretraining with Structured Knowledge for Text-to-SQL Generation
Robert Giaquinto, Dejiao Zhang, Benjamin Kleiner, Yang Li, Ming Tan, Parminder Bhatia, Ramesh Nallapati and Xiaofei Ma 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Many machine learning-based low-code or no-code applications involve generating code that interacts with structured knowledge. For example, one of the most studied tasks in this area is generating SQL code from a natural language statement. Prior work shows that incorporating context information from the database schema, such as table and column names, is beneficial to model performance on this task. In this work we present a large pretraining dataset and strategy for learning representations of text, tables, and SQL code that leverages the entire context of the problem. Specifically, we build on existing encoder-decoder architecture by introducing a multitask pretraining framework that complements the unique attributes of our diverse pretraining data. Our work represents the first study on large-scale pretraining of encoder-decoder models for interacting with structured knowledge, and offers a new state-of-the-art foundation model in text-to-SQL generation. We validate our approach with experiments on two SQL tasks, showing improvement over existing methods, including a 1.7 and 2.2 percentage point improvement over prior state-of-the-arts on Spider and CoSQL.

Tree-Based Representation and Generation of Natural and Mathematical Language
Alexander Scarlatos and Andrew Lan   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Mathematical language in scientific communications and educational scenarios is important yet relatively understudied compared to natural languages. Recent works on mathematical language focus either on representing stand-alone mathematical expressions, especially in their natural tree format, or mathematical reasoning in pre-trained natural language models. Existing works on jointly modeling and generating natural and mathematical languages simply treat mathematical expressions as text, without accounting for the rigid structural properties of mathematical expressions. In this paper, we propose a series of modifications to existing language models to jointly represent and generate text and math: representing mathematical expressions as sequences of node tokens in their operator tree format, using math symbol and tree position embeddings to preserve the semantic and structural properties of mathematical expressions, and using a constrained decoding method to generate mathematically valid expressions. We ground our modifications in GPT-2, resulting in a model MathGPT, and demonstrate that it outperforms baselines on mathematical expression generation tasks.

CHBias: Bias Evaluation and Mitigation of Chinese Conversational Language Models
Jiaxu Zhao, Meng Fang, Zijing Shi, Yitong Li, Ling Chen and Mykola Pechenizkiy   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Warning: This paper contains content that may be offensive or upsetting.

Pretrained conversational agents have been exposed to safety issues, exhibiting a range of stereotypical human biases such as gender bias. However, there are still limited bias categories in current research, and most of them only focus on English. In this paper, we introduce a new Chinese dataset, CHBias, for bias evaluation and mitigation of Chinese conversational language models. Apart from those previous well-explored bias categories, CHBias includes under-explored bias categories, such as ageism and appearance biases, which received less attention. We evaluate two popular pretrained Chinese conversational models, CDial-GPT and EVA2.0, using CHBias. Furthermore, to mitigate different biases, we apply several debiasing methods to the Chinese pretrained models. Experimental results show that these Chinese

pretrained models are potentially risky for generating texts that contain social biases, and debiasing methods using the proposed dataset can make response generation less biased while preserving the models' conversational capabilities.

RECAP: Retrieval-Enhanced Context-Aware Prefix Encoder for Personalized Dialogue Response Generation
Shuai Liu, Hyundong J Cho, Marjorie Freedman, Xuezhe Ma and Jonathan May   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Endowing chatbots with a consistent persona is essential to an engaging conversation, yet it remains an unresolved challenge. In this work, we propose a new retrieval-enhanced approach for personalized response generation. Specifically, we design a hierarchical transformer retriever trained on dialogue domain data to perform personalized retrieval and a context-aware prefix encoder that fuses the retrieved information to the decoder more effectively. Extensive experiments on a real-world dataset demonstrate the effectiveness of our model at generating more fluent and personalized responses. We quantitatively evaluate our model's performance under a suite of human and automatic metrics and find it to be superior compared to state-of-the-art baselines on English Reddit conversations.

Modeling User Satisfaction Dynamics in Dialogue via Hawkes Process
Fanghua Ye, zhiyuan hu and Emine Yilmaz   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Dialogue systems have received increasing attention while automatically evaluating their performance remains challenging. User satisfaction estimation (USE) has been proposed as an alternative. It assumes that the performance of a dialogue system can be measured by user satisfaction and uses an estimator to simulate users. The effectiveness of USE depends heavily on the estimator. Existing estimators independently predict user satisfaction at each turn and ignore satisfaction dynamics across turns within a dialogue. In order to fully simulate users, it is crucial to take satisfaction dynamics into account. To fill this gap, we propose a new estimator ASAP (sAtisfaction eStimation via HAwkes Process) that treats user satisfaction across turns as an event sequence and employs a Hawkes process to effectively model the dynamics in this sequence. Experimental results on four benchmark dialogue datasets demonstrate that ASAP can substantially outperform state-of-the-art baseline estimators.

Toward Interactive Dictation
Belinda Z. Li, Jason Eisner, Adam Pauls and Sam Thomson   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Voice dictation is an increasingly important text input modality. Existing systems that allow both dictation and editing-by-voice restrict their command language to flat templates invoked by trigger words. In this work, we study the feasibility of allowing users to interrupt their dictation with spoken editing commands in open-ended natural language. We introduce a new task and dataset, TERTiUS, to experiment with such systems. To support this flexibility in real-time, a system must incrementally segment and classify spans of speech as either dictation or command, and interpret the spans that are commands. We experiment with using large pre-trained language models to predict the edited text, or alternatively, to predict a small text-editing program. Experiments show a natural trade-off between model accuracy and latency: a smaller model achieves 30% end-state accuracy with 1.3 seconds of latency, while a larger model achieves 55% end-state accuracy with 7 seconds of latency.

RADE: Reference-Assisted Dialogue Evaluation for Open-Domain Dialogue
*Zhengliang Shi, Weiwei Sun, Shuo Zhang, Zhen Zhang, Pengjie Ren and Zhaochun Ren*09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Evaluating open-domain dialogue systems is challenging for reasons such as the one-to-many problem, i.e., many appropriate responses other than just the golden response. As of now, automatic evaluation methods need better consistency with humans, while reliable human evaluation can be time- and cost-intensive. To this end, we propose the Reference-Assisted Dialogue Evaluation (RADE) approach under the multi-task learning framework, which leverages the pre-created utterance as reference other than the gold response to relief the one-to-many problem. Specifically, RADE explicitly compares reference and the candidate response to predict their overall scores. Moreover, an auxiliary response generation task enhances prediction via a shared encoder. To support RADE, we extend three datasets with additional rated responses other than just a golden response by human annotation. Experiments on our three datasets and two existing benchmarks demonstrate the effectiveness of our method, where Pearson, Spearman, and Kendall correlations with human evaluation outperform state-of-the-art baselines.

Towards Faithful Dialogues via Focus Learning
Yifan Deng, Xingsheng Zhang, Heyan Huang and Yue Hu   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Maintaining faithfulness between responses and knowledge is an important research topic for building reliable knowledge-grounded dialogue systems. Existing models heavily rely on elaborate data engineering or increasing the model's parameters ignoring to track the tokens that significantly influence losses, which is decisive for the optimization direction of the model in each iteration. To address this issue, we propose Focus Learning (FocusL), a novel learning approach that adjusts the contribution of each token to the optimization direction by directly scaling the corresponding objective loss. Specifically, we first introduce a positioning method by utilizing similarity distributions between knowledge and each response token to locate knowledge-aware tokens. Then, we further design a similarity-to-weight transformation to provide dynamic token-level weights for the cross-entropy loss. Finally, we use the weighted loss to encourage the model to pay special attention to the knowledge utilization. Experimental results demonstrate that our method achieves the new state-of-the-art results and generates more reliable responses while maintaining training stability.

Modeling What-to-ask and How-to-ask for Answer-unaware Conversational Question Generation
Xuan Long Do, Bowei Zou, Shafiq Joty, Tran Anh Tai, Liangming Pan, Nancy Chen and Ai Ti Aw   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Conversational Question Generation (CQG) is a critical task for machines to assist humans in fulfilling their information needs through conversations. The task is generally cast into two different settings: answer-aware and answer-unaware. While the former facilitates the models by exposing the expected answer, the latter is more realistic and receiving growing attentions recently. What-to-ask and how-to-ask are the two main challenges in the answer-unaware setting. To address the first challenge, existing methods mainly select sequential sentences in context as the rationales. We argue that the conversation generated using such naive heuristics may not be natural enough as in reality, the interlocutors often talk about the relevant contents that are not necessarily sequential in context. Additionally, previous methods decide the type of question to be generated (boolean/span-based) implicitly. Modeling the question type explicitly is crucial as the answer, which hints the models to generate a boolean or span-based question, is unavailable. To this end, we present SG-CQG, a two-stage CQG framework. For the what-to-ask stage, a sentence is selected as the rationale from a semantic graph that we construct, and extract the answer span from it. For the how-to-ask stage, a classifier determines the target answer type of the question via two explicit control signals before generating and filtering. In addition, we propose Conv-Distinct, a novel evaluation metric for CQG, to evaluate the diversity of the generated conversation from a context. Compared with the existing answer-unaware CQG models, the proposed SG-CQG achieves state-of-the-art performance.

ACCENT: An Automatic Event Commonsense Evaluation Metric for Open-Domain Dialogue Systems
Sarik Ghazarian, Yijia Shao, Rujun Han, Aram Galstyan and Nanyun Peng   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Commonsense reasoning is omnipresent in human communications and thus is an important feature for open-domain dialogue systems. However, evaluating commonsense in dialogue systems is still an open challenge. We take the first step by focusing on *event commonsense* that considers events and their relations, and is crucial in both dialogues and general commonsense reasoning. We propose **ACCENT**, an event commonsense evaluation metric empowered by commonsense knowledge bases (CSKBs). ACCENT first extracts event-relation tuples from

a dialogue, and then evaluates the response by scoring the tuples in terms of their compatibility with the CSKB. To evaluate ACCENT, we construct the first public event commonsense evaluation dataset for open-domain dialogues. Our experiments show that ACCENT is an efficient metric for event commonsense evaluation, which achieves higher correlations with human judgments than existing baselines.

Prompter: Zero-shot Adaptive Prefixes for Dialogue State Tracking Domain Adaptation
Ibrahim Taha Aksu, Min-Yen Kan and Nancy Chen   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
A challenge in the Dialogue State Tracking (DST) field is adapting models to new domains without using any supervised data — zero-shot domain adaptation. Parameter-Efficient Transfer Learning (PETL) has the potential to address this problem due to its robustness. However, it has yet to be applied to the zero-shot scenarios, as it is not clear how to apply it unsupervisedly.
Our method, Prompter, uses descriptions of target domain slots to generate dynamic prefixes that are concatenated to the key and values at each layer's self-attention mechanism. This allows for the use of prefix-tuning in zero-shot. Prompter outperforms previous methods on both the MultiWOZ and SGD benchmarks. In generating prefixes, our analyses find that Prompter not only utilizes the semantics of slot descriptions but also how often the slots appear together in conversation. Moreover, Prompter's gains are due to its improved ability to distinguish "none"-valued dialogue slots, compared against baselines.

Don't Generate, Discriminate: A Proposal for Grounding Language Models to Real-World Environments
Yu Gu, Xiang Deng and Yu Su   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
A key missing capacity of current language models (LMs) is grounding to real-world environments. Most existing work for grounded language understanding uses LMs to directly generate plans that can be executed in the environment to achieve the desired effects. It thereby casts the burden of ensuring grammaticality, faithfulness, and controllability all on the LMs. We propose Pangu, a generic framework for grounded language understanding that capitalizes on the discriminative ability of LMs instead of their generative ability. Pangu consists of a symbolic agent and a neural LM working in a concerted fashion: The agent explores the environment to incrementally construct valid plans, and the LM evaluates the plausibility of the candidate plans to guide the search process. A case study on the challenging problem of knowledge base question answering (KBQA), which features a massive environment, demonstrates the remarkable effectiveness and flexibility of Pangu: A BERT-base LM is sufficient for setting a new record on standard KBQA datasets, and larger LMs further bring substantial gains. Pangu also enables, for the first time, effective few-shot in-context learning for KBQA with large LMs such as Codex.

Probing Physical Reasoning with Counter-Commonsense Context
Kazushi Kondo, Saku Sugawara and Akiko Aizawa   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
In this study, we create a CConS (Counter-commonsense Contextual Size comparison) dataset to investigate how physical commonsense affects the contextualized size comparison task; the proposed dataset consists of both contexts that fit physical commonsense and those that do not. This dataset tests the ability of language models to predict the size relationship between objects under various contexts generated from our curated noun list and templates. We measure the ability of several masked language models and encoder-decoder models. The results show that while large language models can use prepositions such as "in" and "into" in the provided context to infer size relationships, they fail to use verbs and thus make incorrect judgments led by their prior physical commonsense.

To Revise or Not to Revise: Learning to Detect Improvable Claims for Argumentative Writing Support
Gabriella Skitalinskaya and Henning Wachsmuth   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Optimizing the phrasing of argumentative text is crucial in higher education and professional development. However, assessing whether and how the different claims in a text should be revised is a hard task, especially for novice writers. In this work, we explore the main challenges to identifying argumentative claims in need of specific revisions. By learning from collaborative editing behaviors in online debates, we seek to capture implicit revision patterns in order to develop approaches aimed at guiding writers in how to further improve their arguments. We systematically compare the ability of common word embedding models to capture the differences between different versions of the same text, and we analyze their impact on various types of writing issues. To deal with the noisy nature of revision-based corpora, we propose a new sampling strategy based on revision distance. Opposed to approaches from prior work, such sampling can be done without employing additional annotations and judgments. Moreover, we provide evidence that using contextual information and domain knowledge can further improve prediction results. How useful a certain type of context is, depends on the issue the claim is suffering from, though.

A Simple and Flexible Modeling for Mental Disorder Detection by Learning from Clinical Questionnaires
Hoyun Song, Jisu Shin, Huije Lee and Jong Park   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Social media is one of the most highly sought resources for analyzing characteristics of the language by its users. In particular, many researchers utilized various linguistic features of mental health problems from social media. However, existing approaches to detecting mental disorders face critical challenges, such as the scarcity of high-quality data or the trade-off between addressing the complexity of models and presenting interpretable results grounded in expert domain knowledge. To address these challenges, we design a simple but flexible model that preserves domain-based interpretability. We propose a novel approach that captures the semantic meanings directly from the text and compares them to symptom-related descriptions. Experimental results demonstrate that our model outperforms relevant baselines on various mental disorder detection tasks. Our detailed analysis shows that the proposed model is effective at leveraging domain knowledge, transferable to other mental disorders, and providing interpretable detection results.

StoryARG: a corpus of narratives and personal experiences in argumentative texts
Neele Falk and Gabriella Lapesa   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Humans are storytellers, even in communication scenarios which are assumed to be more rationality-oriented, such as argumentation. Indeed, supporting arguments with narratives or personal experiences (henceforth, stories) is a very natural thing to do – and yet, this phenomenon is largely unexplored in computational argumentation. Which role do stories play in an argument? Do they make the argument more effective? What are their narrative properties? To address these questions, we collected and annotated StoryARG, a dataset sampled from well-established corpora in computational argumentation (ChangeMyView and RegulationRoom), and the Social Sciences (Europolis), as well as comments to New York Times articles. StoryARG contains 2451 textual spans annotated at two levels. At the argumentative level, we annotate the function of the story (e.g., clarification, disclosure of harm, search for a solution, establishing speaker's authority), as well as its impact on the effectiveness of the argument and its emotional load. At the level of narrative properties, we annotate whether the story has a plot-like development, is factual or hypothetical, and who the protagonist is.
What makes a story effective in an argument? Our analysis of the annotations in StoryARG uncover a positive impact on effectiveness for stories which illustrate a solution to a problem, and in general, annotator-specific preferences that we investigate with regression analysis.

Holistic Prediction on a Time-Evolving Attributed Graph
Shohei Yamasaki, Yuya Sasaki, Panagiotis Karras and Makoto Onizuka   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Graph-based prediction is essential in NLP tasks such as temporal knowledge graph completion. A cardinal question in this field is, how to predict the future links, nodes, and attributes of a time-evolving attributed graph? Unfortunately, existing techniques assume that each link, node, and attribute prediction is independent, and fall short of predicting the appearance of new nodes that were not observed in the past. In this paper, we address two interrelated questions; (1) can we exploit task interdependence to improve prediction accuracy? and (2) can

we predict new nodes with their attributes? We propose a unified framework that predicts node attributes and topology changes such as the appearance and disappearance of links and the emergence and loss of nodes. This frame-work comprises components for independent and interactive prediction and for predicting new nodes. Our experimental study using real-world data confirms that our interdependent prediction framework achieves higher accuracy than methods based on independent prediction.

The Benefits of Bad Advice: Autocontrastive Decoding across Model Layers
Ariel Gera, Roni Friedman, Ofir Arviv, Chulaka Gunasekara, Benjamin Sznajder, Noam Slonim and Eyal Shnarch   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Applying language models to natural language processing tasks typically relies on the representations in the final model layer, as intermediate hidden layer representations are presumed to be less informative. In this work, we argue that due to the gradual improvement across model layers, additional information can be gleaned from the contrast between higher and lower layers during inference. Specifically, in choosing between the probable next token predictions of a generative model, the predictions of lower layers can be used to highlight which candidates are best avoided. We propose a novel approach that utilizes the contrast between layers to improve text generation outputs, and show that it mitigates degenerative behaviors of the model in open-ended generation, significantly improving the quality of generated texts. Furthermore, our results indicate that contrasting between model layers at inference time can yield substantial benefits to certain aspects of general language model capabilities, more effectively extracting knowledge during inference from a given set of model parameters.

Linear Guardedness and its Implications
Shauli Ravfogel, Yoav Goldberg and Ryan Cotterell   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Methods for erasing human-interpretable concepts from neural representations that assume linearity have been found to be tractable and useful. However, the impact of this removal on the behavior of downstream classifiers trained on the modified representations is not fully understood. In this work, we formally define the notion of linear guardedness as the inability of an adversary to predict the concept directly from the representation, and study its implications. We show that, in the binary case, under certain assumptions, a downstream log-linear model cannot recover the erased concept. However, we constructively demonstrate that a multiclass log-linear model *can* be constructed that indirectly recovers the concept in some cases, pointing to the inherent limitations of linear guardedness as a downstream bias mitigation technique. These findings shed light on the theoretical limitations of linear erasure methods and highlight the need for further research on the connections between intrinsic and extrinsic bias in neural models.

Characterizing and Measuring Linguistic Dataset Drift
Tyler A Chang, Kishaloy Halder, Neha Anna John, Yogarshi Vyas, Yassine Benajiba, Miguel Ballesteros and Dan Roth   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
NLP models often degrade in performance when real world data distributions differ markedly from training data. However, existing dataset drift metrics in NLP have generally not considered specific dimensions of linguistic drift that affect model performance, and they have not been validated in their ability to predict model performance at the individual example level, where such metrics are often used in practice. In this paper, we propose three dimensions of linguistic dataset drift: vocabulary, structural, and semantic drift. These dimensions correspond to content word frequency divergences, syntactic divergences, and meaning changes not captured by word frequencies (e.g. lexical semantic change). We propose interpretable metrics for all three drift dimensions, and we modify past performance prediction methods to predict model performance at both the example and dataset level for English sentiment classification and natural language inference. We find that our drift metrics are more effective than previous metrics at predicting out-of-domain model accuracies (mean 16.8% root mean square error decrease), particularly when compared to popular fine-tuned embedding distances (mean 47.7% error decrease). Fine-tuned embedding distances are much more effective at ranking individual examples by expected performance, but decomposing into vocabulary, structural, and semantic drift produces the best example rankings of all considered model-agnostic drift metrics (mean 6.7% ROC AUC increase).

Hidden Schema Networks
*Ramses J Sanchez, Lukas Alexander Conrads, Pascal Welke, Kostadin Cvejoski and Cesar Ali Ojeda Marin*09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Large, pretrained language models infer powerful representations that encode rich semantic and syntactic content, albeit implicitly. In this work we introduce a novel neural language model that enforces, via inductive biases, explicit relational structures which allow for compositionality onto the output representations of pretrained language models. Specifically, the model encodes sentences into sequences of symbols (composed representations), which correspond to the nodes visited by biased random walkers on a global latent graph, and infers the posterior distribution of the latter. We first demonstrate that the model is able to uncover ground-truth graphs from artificially generated datasets of random token sequences. Next, we leverage pretrained BERT and GPT-2 language models as encoder and decoder, respectively, to infer networks of symbols (schemata) from natural language datasets. Our experiments show that (i) the inferred symbols can be interpreted as encoding different aspects of language, as e.g. topics or sentiments, and that (ii) GPT-2-like models can effectively be conditioned on symbolic representations. Finally, we explore training autoregressive, random walk "reasoning" models on schema networks inferred from commonsense knowledge databases, and using the sampled paths to enhance the performance of pretrained language models on commonsense If-Then reasoning tasks.

ContraCLM: Contrastive Learning For Causal Language Model
Nihal Jain, Dejiao Zhang, Wasi Uddin Ahmad, Zijian Wang, Feng Nan, Xiaopeng Li, Ming Tan, Ramesh Nallapati, Baishakhi Ray, Parminder Bhatia, Xiaofei Ma and Bing Xiang   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Despite exciting progress in causal language models, the expressiveness of their representations is largely limited due to poor discrimination ability. To remedy this issue, we present CONTRACLM, a novel contrastive learning framework at both the token-level and the sequence-level. We assess CONTRACLM on a variety of downstream tasks. We show that CONTRACLM enhances the discrimination of representations and bridges the gap with encoder-only models, which makes causal language models better suited for tasks beyond language generation. Specifically, we attain 44% relative improvement on the Semantic Textual Similarity tasks and 34% on Code-to-Code Search tasks. Furthermore, by improving the expressiveness of representations, CONTRACLM also boosts the source code generation capability with 9% relative improvement on execution accuracy on the HumanEval benchmark.

Is Fine-tuning Needed? Pre-trained Language Models Are Near Perfect for Out-of-Domain Detection
Rheeya Uppaal, Junjie Hu and Yixuan Li   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Out-of-distribution (OOD) detection is a critical task for reliable predictions over text. Fine-tuning with pre-trained language models has been a de facto procedure to derive OOD detectors with respect to in-distribution (ID) data. Despite its common use, the understanding of the role of fine-tuning and its necessity for OOD detection is largely unexplored. In this paper, we raise the question: is fine-tuning necessary for OOD detection? We present a study investigating the efficacy of directly leveraging pre-trained language models for OOD detection, without any model fine-tuning on the ID data. We compare the approach with several competitive fine-tuning objectives, and offer new insights under various types of distributional shifts. Extensive experiments demonstrate near-perfect OOD detection performance (with 0% FPR95 in many cases), strongly outperforming the fine-tuned counterpart.

Learning Neuro-Symbolic World Models with Conversational Proprioception

Don Joven Agravante, Daiki Kimura, Michiaki Tatsubori, Asim Munawar and Alexander Gray   09:00-10:30 (Frontenac Ballroom and Queen's Quay)

The recent emergence of Neuro-Symbolic Agent (NeSA) approaches to natural language-based interactions calls for the investigation of model-based approaches. In contrast to model-free approaches, which existing NeSAs take, learning an explicit world model has an interesting potential especially in the explainability, which is one of the key selling points of NeSA. To learn useful world models, we leverage one of the recent neuro-symbolic architectures, Logical Neural Networks (LNN). Here, we describe a method that can learn neuro-symbolic world models on the TextWorld-Commonsense set of games. We then show how this can be improved further by taking inspiration from the concept of proprioception, but for conversation. This is done by enhancing the internal logic state with a memory of previous actions while also guiding future actions by augmenting the learned model with constraints based on this memory. This greatly improves the game-solving agents performance in a TextWorld setting, where the advantage over the baseline is an 85% average steps reduction and x2.3 average score.

Reasoning with Language Model Prompting: A Survey

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang and Huajun Chen   09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Reasoning, as an essential ability for complex problem-solving, can provide back-end support for various real-world applications, such as medical diagnosis, negotiation, etc. This paper provides a comprehensive survey of cutting-edge research on reasoning with language model prompting. We introduce research works with comparisons and summaries and provide systematic resources to help beginners. We also discuss the potential reasons for emerging such reasoning abilities and highlight future research directions. Resources are available at https://github.com/zjunlp/Prompt4ReasoningPapers (updated periodically).

Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer and Huan Sun   09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Chain-of-Thought (CoT) prompting can dramatically improve the multi-step reasoning abilities of large language models (LLMs). CoT explicitly encourages the LLM to generate intermediate rationales for solving a problem, by providing a series of reasoning steps in the demonstrations. Despite its success, there is still little understanding of what makes CoT prompting effective and which aspects of the demonstrated reasoning steps contribute to its performance. In this paper, we show that CoT reasoning is possible even with invalid demonstrations - prompting with invalid reasoning steps can achieve over 80-90% of the performance obtained using CoT under various metrics, while still generating coherent lines of reasoning during inference. Further experiments show that other aspects of the rationales, such as being relevant to the query and correctly ordering the reasoning steps, are much more important for effective CoT reasoning. Overall, these findings both deepen our understanding of CoT prompting, and open up new questions regarding LLMs' capability to learn to reason in context.

Z-ICL: Zero-Shot In-Context Learning with Pseudo-Demonstrations

Xinxi Lyu, Sewon Min, Iz Beltagy, Luke Zettlemoyer and Hannaneh Hajishirzi   09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Although large language models can be prompted for both zero- and few-shot learning, performance drops significantly when no demonstrations are available. In this paper, we introduce Z-ICL, a new zero-shot method that closes the gap by constructing pseudo-demonstrations for a given test input using a raw text corpus. Concretely, pseudo-demonstrations are constructed by (1) finding the nearest neighbors to the test input from the corpus and pairing them with random task labels, and (2) applying a set of techniques to reduce the amount of direct copying the model does from the resulting demonstrations. Evaluation on nine classification datasets shows that Z-ICL outperforms previous zero-shot methods by a significant margin, and is on par with in-context learning with labeled training data in the few-shot setting. Overall, Z-ICL provides a significantly higher estimate of the zero-shot performance levels of a model, and supports future efforts to develop better pseudo-demonstrations that further improve zero-shot results.

Matching Pairs: Attributing Fine-Tuned Models to their Pre-Trained Large Language Models

Myles Foley, Ambrish Rawat, Taesung Lee, Yufang Hou, Gabriele Picco and Giulio Zizzo   09:00-10:30 (Frontenac Ballroom and Queen's Quay)

The wide applicability and adaptability of generative large language models (LLMs) has enabled their rapid adoption. While the pre-trained models can perform many tasks, such models are often fine-tuned to improve their performance on various downstream applications. However, this leads to issues over violation of model licenses, model theft, and copyright infringement. Moreover, recent advances show that generative technology is capable of producing harmful content which exacerbates the problems of accountability within model supply chains. Thus, we need a method to investigate how a model was trained or a piece of text was generated and what their pre-trained base model was. In this paper we take the first step to address this open problem by tracing back the origin of a given fine-tuned LLM to its corresponding pre-trained base model. We consider different knowledge levels and attribution strategies, and find that we can correctly trace back 8 out of the 10 fine tuned models with our best method.

Explanation-based Finetuning Makes Models More Robust to Spurious Cues

Josh Magnus Ludan, Yixuan Meng, Tai Nguyen, Saurabh Shah, Qing Lyu, Marianna Apidianaki and Chris Callison-Burch   09:00-10:30 (Frontenac Ballroom and Queen's Quay)

Large Language Models (LLMs) are so powerful that they sometimes learn correlations between labels and features that are irrelevant to the task, leading to poor generalization on out-of-distribution data. We propose explanation-based finetuning as a general approach to mitigate LLMs' reliance on spurious correlations. Unlike standard finetuning where the model only predicts the answer given the input, we finetune the model to additionally generate a free-text explanation supporting its answer. To evaluate our method, we finetune the model on artificially constructed training sets containing different types of spurious cues, and test it on a test set without these cues. Compared to standard finetuning, our method makes GPT-3 (davinci) remarkably more robust against spurious cues in terms of accuracy drop across four classification tasks: ComVE (+1.2), CREAK (+9.1), e-SNLI (+15.4), and SBIC (+6.5). The efficacy generalizes across multiple model families and scales, with greater gains for larger models. Finally, our method also works well with explanations generated by the model, implying its applicability to more datasets without human-written explanations.

DOC: Improving Long Story Coherence With Detailed Outline Control

Kevin Yang, Dan Klein, Nanyun Peng and Yuandong Tian   09:00-10:30 (Frontenac Ballroom and Queen's Quay)

We propose the Detailed Outline Control (DOC) framework for improving long-range plot coherence when automatically generating severalthousand-word-long stories. DOC consists of two complementary components: a detailed outliner and a detailed controller. The detailed outliner creates a more detailed, hierarchically structured outline, shifting creative burden from the main drafting procedure to the planning stage. The detailed controller ensures the more detailed outline is still respected during generation by controlling story passages to align with outline details. In human evaluations of automatically generated stories, DOC substantially outperforms a strong Re3 baseline (Yang et al., 2022) on plot coherence (22.5% absolute gain), outline relevance (28.2%), and interestingness (20.7%). Humans also judged DOC to be much more controllable in an interactive generation setting.

SIMSUM: Document-level Text Simplification via Simultaneous Summarization
Sofia Blinova, Xinyu Zhou, Martin Jaggi, Carsten Eickhoff and Seyed Ali Bahrainian   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Document-level text simplification is a specific type of simplification which involves simplifying documents consisting of several sentences by rewriting them into fewer or more sentences. In this paper, we propose a new two-stage framework SIMSUM for automated document-level text simplification. Our model is designed with explicit summarization and simplification models and guides the generation using the main keywords of a source text. In order to evaluate our new model, we use two existing benchmark datasets for simplification, namely D-Wikipedia and Wiki-Doc. We compare our model's performance with state of the art and show that SIMSUM achieves top results on the D-Wikipedia dataset SARI (+1.20), D-SARI (+1.64), and FKGL (-0.35) scores, improving over the best baseline models. In order to evaluate the quality of the generated text, we analyze the outputs from different models qualitatively and demonstrate the merit of our new model. Our code and datasets are available.

Summarizing, Simplifying, and Synthesizing Medical Evidence using GPT-3 (with Varying Success)
*Chantal Shaib, Millicent L Li, Sebastian A Joseph, Iain Marshall, Junyi Jessy Li and Byron C. Wallace*09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Large language models, particularly GPT-3, are able to produce high quality summaries of general domain news articles in few- and zero-shot settings. However, it is unclear if such models are similarly capable in more specialized domains such as biomedicine. In this paper we enlist domain experts (individuals with medical training) to evaluate summaries of biomedical articles generated by GPT-3, given no supervision. We consider both single- and multi-document settings. In the former, GPT-3 is tasked with generating regular and plain-language summaries of articles describing randomized controlled trials; in the latter, we assess the degree to which GPT-3 is able to synthesize evidence reported across a collection of articles. We design an annotation scheme for evaluating model outputs, with an emphasis on assessing the factual accuracy of generated summaries. We find that while GPT-3 is able to summarize and simplify single biomedical articles faithfully, it struggles to provide accurate aggregations of findings over multiple documents. We release all data, code, and annotations used in this work.

Understanding Factual Errors in Summarization: Errors, Summarizers, Datasets, Error Detectors
Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin F Rousseau and Greg Durrett
09:00-10:30 (Frontenac Ballroom and Queen's Quay)
The propensity of abstractive summarization models to make factual errors has been studied extensively, including design of metrics to detect factual errors and annotation of errors in current systems' outputs. However, the ever-evolving nature of summarization systems, metrics, and annotated benchmarks makes factuality evaluation a moving target, and drawing clear comparisons among metrics has become increasingly difficult. In this work, we aggregate factuality error annotations from nine existing datasets and stratify them according to the underlying summarization model. We compare performance of state-of-the-art factuality metrics, including recent ChatGPT-based metrics, on this stratified benchmark and show that their performance varies significantly across different types of summarization models. Critically, our analysis shows that much of the recent improvement in the factuality detection space has been on summaries from older (pre-Transformer) models instead of more relevant recent summarization models. We further perform a finer-grained analysis per error-type and find similar performance variance across error types for different factuality metrics. Our results show that no one metric is superior in all settings or for all error types, and we provide recommendations for best practices given these insights.

Factually Consistent Summarization via Reinforcement Learning with Textual Entailment Feedback
Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Leonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos Garea, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin and Idan Szpektor   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Despite the seeming success of contemporary grounded text generation systems, they often tend to generate factually inconsistent text with respect to their input. This phenomenon is emphasized in tasks like summarization, in which the generated summaries should be corroborated by their source article. In this work we leverage recent progress on textual entailment models to directly address this problem for abstractive summarization systems. We use reinforcement learning with reference-free, textual-entailment rewards to optimize for factual consistency and explore the ensuing trade-offs, as improved consistency may come at the cost of less informative or more extractive summaries. Our results, according to both automatic metrics and human evaluation, show that our method considerably improves the faithfulness, salience and conciseness of the generated summaries.

Optimal Transport for Unsupervised Hallucination Detection in Neural Machine Translation
Nuno M. Guerreiro, Pierre Colombo, Pablo Piantanida and André Martins   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Neural machine translation (NMT) has become the de-facto standard in real-world machine translation applications. However, NMT models can unpredictably produce severely pathological translations, known as hallucinations, that seriously undermine user trust. It becomes thus crucial to implement effective preventive strategies to guarantee their proper functioning. In this paper, we address the problem of hallucination detection in NMT by following a simple intuition: as hallucinations are detached from the source content, they exhibit encoder-decoder attention patterns that are statistically different from those of good quality translations. We frame this problem with an optimal transport formulation and propose a fully unsupervised, plug-in detector that can be used with any attention-based NMT model. Experimental results show that our detector not only outperforms all previous model-based detectors, but is also competitive with detectors that employ external models trained on millions of samples for related tasks such as quality estimation and cross-lingual sentence similarity.

Memory-efficient NLLB-200: Language-specific Expert Pruning of a Massively Multilingual Machine Translation Model
Yeskendir Koishekenov, Alexandre Berard and Vassilina Nikoulina   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
The recently released NLLB-200 is a set of multilingual Neural Machine Translation models that cover 202 languages. The largest model is based on a Mixture of Experts architecture and achieves SoTA results across many language pairs. It contains 54.5B parameters and requires at least four 32GB GPUs just for inference. In this work, we propose a pruning method that enables the removal of up to 80% of experts without further finetuning and with a negligible loss in translation quality, which makes it feasible to run the model on a single 32GB GPU. Further analysis suggests that our pruning metrics can identify language-specific experts.

Towards Understanding and Improving Knowledge Distillation for Neural Machine Translation
Songming Zhang, Yunlong Liang, Shuaibo Wang, Yufeng Chen, Wenjuan Han, Jian Liu and Jinan Xu   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Knowledge distillation (KD) is a promising technique for model compression in neural machine translation. However, where the knowledge hides in KD is still not clear, which may hinder the development of KD. In this work, we first unravel this mystery from an empirical perspective and show that the knowledge comes from the top-1 predictions of teachers, which also helps us build a potential connection between word- and sequence-level KD. Further, we point out two inherent issues in vanilla word-level KD based on this finding. Firstly, the current objective of KD spreads its focus to whole distributions to learn the knowledge, yet lacks special treatment on the most crucial top-1 information. Secondly, the knowledge is largely covered by the golden information due to the fact that most top-1 predictions of teachers overlap with ground-truth tokens, which further restricts the potential of KD. To address these issues, we propose a new method named Top-1 Information Enhanced Knowledge Distillation (TIE-KD). Specifically, we design a hierarchical ranking loss to enforce the learning of the

top-1 information from the teacher. Additionally, we develop an iterative KD procedure to infuse more additional knowledge by distilling on the data without ground-truth targets. Experiments on WMT'14 English-German, WMT'14 English-French and WMT'16 English-Romanian demonstrate that our method can respectively boost Transformer$_{base}$ students by +1.04, +0.60 and +1.11 BLEU scores and significantly outperforms the vanilla word-level KD baseline. Besides, our method shows higher generalizability on different teacher-student capacity gaps than existing KD techniques.

When Does Translation Require Context? A Data-driven, Multilingual Exploration
Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins and Graham Neubig   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Although proper handling of discourse significantly contributes to the quality of machine translation (MT), these improvements are not adequately measured in common translation quality metrics. Recent works in context-aware MT attempt to target a small set of discourse phenomena during evaluation, however not in a fully systematic way. In this paper, we develop the Multilingual Discourse-Aware (MuDA) benchmark, a series of taggers that identify and evaluate model performance on discourse phenomena in any given dataset. The choice of phenomena is inspired by a novel methodology to systematically identify translations that require context. This methodology confirms the difficulty of previously studied phenomena while uncovering others which were not previously addressed. We find that commonly studied context-aware MT models make only marginal improvements over context-agnostic models, which suggests these models do not handle these ambiguities effectively. We release code and data for 14 language pairs to encourage the MT community to focus on accurately capturing discourse phenomena. Code available at https://github.com/neulab/contextual-mt

Code4Struct: Code Generation for Few-Shot Event Structure Prediction
Xingyao Wang, Sha Li and Heng Ji   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Large Language Model (LLM) trained on a mixture of text and code has demonstrated impressive capability in translating natural language (NL) into structured code. We observe that semantic structures can be conveniently translated into code and propose Code4Struct to leverage such text-to-structure translation capability to tackle structured prediction tasks. As a case study, we formulate Event Argument Extraction (EAE) as converting text into event-argument structures that can be represented as a class object using code. This alignment between structures and code enables us to take advantage of Programming Language (PL) features such as inheritance and type annotation to introduce external knowledge or add constraints. We show that, with sufficient in-context examples, formulating EAE as a code generation problem is advantageous over using variants of text-based prompts. Despite only using 20 training event instances for each event type, Code4Struct is comparable to supervised models trained on 4,202 instances and outperforms current state-of-the-art (SOTA) trained on 20-shot data by 29.5% absolute F1. Code4Struct can use 10-shot training data from a sibling event type to predict arguments for zero-resource event types and outperforms the zero-shot baseline by 12% absolute F1.

CodeIE: Large Code Generation Models are Better Few-Shot Information Extractors
Peng Li, Tianxiang Sun, Qiong Tang, Hang Yan, Yuanbin Wu, Xuanjing Huang and Xipeng Qiu   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Large language models (LLMs) pre-trained on massive corpora have demonstrated impressive few-shot learning ability on many NLP tasks. A common practice is to recast the task into a text-to-text format such that generative LLMs of natural language (NL-LLMs) like GPT-3 can be prompted to solve it. However, it is nontrivial to perform information extraction (IE) tasks with NL-LLMs since the output of the IE task is usually structured and therefore is hard to be converted into plain text. In this paper, we propose to recast the structured output in the form of code instead of natural language and utilize generative LLMs of code (Code-LLMs) such as Codex to perform IE tasks, in particular, named entity recognition and relation extraction. In contrast to NL-LLMs, we show that Code-LLMs can be well-aligned with these IE tasks by designing code-style prompts and formulating these IE tasks as code generation tasks. Experiment results on seven benchmarks show that our method consistently outperforms fine-tuning moderate-size pre-trained models specially designed for IE tasks (e.g., UIE) and prompting NL-LLMs under few-shot settings. We further conduct a series of in-depth analyses to demonstrate the merits of leveraging Code-LLMs for IE tasks.

Document-Level Multi-Event Extraction with Event Proxy Nodes and Hausdorff Distance Minimization
Xinyu Wang, Lin Gui and Yulan He   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Document-level multi-event extraction aims to extract the structural information from a given document automatically. Most recent approaches usually involve two steps: (1) modeling entity interactions; (2) decoding entity interactions into events. However, such approaches ignore a global view of inter-dependency of multiple events. Moreover, an event is decoded by iteratively merging its related entities as arguments, which might suffer from error propagation and is computationally inefficient. In this paper, we propose an alternative approach for document-level multi-event extraction with event proxy nodes and Hausdorff distance minimization. The event proxy nodes, representing pseudo-events, are able to build connections with other event proxy nodes, essentially capturing global information. The Hausdorff distance makes it possible to compare the similarity between the set of predicted events and the set of ground-truth events. By directly minimizing Hausdorff distance, the model is trained towards the global optimum directly, which improves performance and reduces training time. Experimental results show that our model outperforms previous state-of-the-art method in F1-score on two datasets with only a fraction of training time.

Debiasing Generative Named Entity Recognition by Calibrating Sequence Likelihood
Yu Xia, Yongwei Zhao, Wenhao Wu and Sujian Li   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Recognizing flat, overlapped and discontinuous entities uniformly has been paid increasing attention. Among these works, Seq2Seq formulation prevails for its flexibility and effectiveness. It arranges the output entities into a specific target sequence. However, it introduces bias by assigning all the probability mass to the observed sequence. To alleviate the bias, previous works either augment the data with possible sequences or resort to other formulations. In this paper, we stick to the Seq2Seq formulation and propose a reranking-based approach. It redistributes the likelihood among candidate sequences depending on their performance via a contrastive loss. Extensive experiments show that our simple yet effective method consistently boosts the baseline, and yields competitive or better results compared with the state-of-the-art methods on 8 widely-used datasets for Named Entity Recognition.

GENEVA: Benchmarking Generalizability for Event Argument Extraction with Hundreds of Event Types and Argument Roles
Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang and Nanyun Peng   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Recent works in Event Argument Extraction (EAE) have focused on improving model generalizability to cater to new events and domains. However, standard benchmarking datasets like ACE and ERE cover less than 40 event types and 25 entity-centric argument roles. Limited diversity and coverage hinder these datasets from adequately evaluating the generalizability of EAE models. In this paper, we first contribute by creating a large and diverse EAE ontology. This ontology is created by transforming FrameNet, a comprehensive semantic role labeling (SRL) dataset for EAE, by exploiting the similarity between these two tasks. Then, exhaustive human expert annotations are collected to build the ontology, concluding with 115 events and 220 argument roles, with a significant portion of roles not being entities. We utilize this ontology to further introduce GENEVA, a diverse generalizability benchmarking dataset comprising four test suites aimed at evaluating models' ability to handle limited data and unseen event type generalization. We benchmark six EAE models from various families. The results show that owing to non-entity argument roles, even the best-performing model can only achieve 39% F1 score, indicating how GENEVA provides new

challenges for generalization in EAE. Overall, our large and diverse EAE ontology can aid in creating more comprehensive future resources, while GENEVA is a challenging benchmarking dataset encouraging further research for improving generalizability in EAE. The code and data can be found at https://github.com/PlusLabNLP/GENEVA.

AMPERE: AMR-Aware Prefix for Generation-Based Event Argument Extraction Model
I-Hung Hsu, Zhiyu Xie, Kuan-Hao Huang, Prem Natarajan and Nanyun Peng   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Event argument extraction (EAE) identifies event arguments and their specific roles for a given event. Recent advancement in generation-based EAE models has shown great performance and generalizability over classification-based models. However, existing generation-based EAE models mostly focus on problem re-formulation and prompt design, without incorporating additional information that has been shown to be effective for classification-based models, such as the abstract meaning representation (AMR) of the input passages. Incorporating such information into generation-based models is challenging due to the heterogeneous nature of the natural language form prevalently used in generation-based models and the structured form of AMRs. In this work, we study strategies to incorporate AMR into generation-based EAE models. We propose AMPERE, which generates AMR-aware prefixes for every layer of the generation model. Thus, the prefix introduces AMR information to the generation-based EAE model and then improves the generation. We also introduce an adjusted copy mechanism to AMPERE to help overcome potential noises brought by the AMR graph. Comprehensive experiments and analyses on ACE2005 and ERE datasets show that AMPERE can get 4% - 10% absolute F1 score improvements with reduced training data and it is in general powerful across different training sizes.

MM-SHAP: A Performance-agnostic Metric for Measuring Multimodal Contributions in Vision and Language Models & Tasks
Letitia Parcalabescu and Anette Frank   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Vision and language models (VL) are known to exploit unrobust indicators in individual modalities (e.g., introduced by distributional biases) instead of focusing on relevant information in each modality. That a unimodal model achieves similar accuracy on a VL task to a multimodal one, indicates that so-called unimodal collapse occurred. However, accuracy-based tests fail to detect e.g., when the model prediction is wrong, while the model used relevant information from a modality.

Instead, we propose MM-SHAP, a performance-agnostic multimodality score based on Shapley values that reliably quantifies in which proportions a multimodal model uses individual modalities. We apply MM-SHAP in two ways: (1) to compare models for their average degree of multimodality, and (2) to measure for individual models the contribution of individual modalities for different tasks and datasets.

Experiments with six VL models – LXMERT, CLIP and four ALBEF variants – on four VL tasks highlight that unimodal collapse can occur to different degrees and in different directions, contradicting the wide-spread assumption that unimodal collapse is one-sided. Based on our results, we recommend MM-SHAP for analysing multimodal tasks, to diagnose and guide progress towards multimodal integration. Code available at https://github.com/Heidelberg-NLP/MM-SHAP.

Weakly Supervised Vision-and-Language Pre-training with Relative Representations
Chi Chen, Peng Li, Maosong Sun and Yang Liu   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Weakly supervised vision-and-language pre-training (WVLP), which learns cross-modal representations with limited cross-modal supervision, has been shown to effectively reduce the data cost of pre-training while maintaining decent performance on downstream tasks. However, current WVLP methods use only local descriptions of images, i.e., object tags, as cross-modal anchors to construct weakly-aligned image-text pairs for pre-training. This affects the data quality and thus the effectiveness of pre-training. In this paper, we propose to directly take a small number of aligned image-text pairs as anchors, and represent each unaligned image and text by its similarities to these anchors, i.e., relative representations. We build a WVLP framework based on the relative representations, namely RELIT, which collects high-quality weakly-aligned image-text pairs from large-scale image-only and text-only data for pre-training through relative representation-based retrieval and generation. Experiments on four downstream tasks show that RELIT achieves new state-of-the-art results under the weakly supervised setting.

End-to-end Knowledge Retrieval with Multi-modal Queries
Man Luo, Zhiyuan Fang, Tejas Gokhale, Yezhou Yang and Chitta Baral   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We investigate knowledge retrieval with multi-modal queries, i.e. queries containing information split across image and text inputs, a challenging task that differs from previous work on cross-modal retrieval. We curate a new dataset called ReMuQ for benchmarking progress on this task. ReMuQ requires a system to retrieve knowledge from a large corpus by integrating contents from both text and image queries. We introduce a retriever model "ReViz" that can directly process input text and images to retrieve relevant knowledge in an end-to-end fashion without being dependent on intermediate modules such as object detectors or caption generators. We introduce a new pretraining task that is effective for learning knowledge retrieval with multimodal queries and also improves performance on downstream tasks. We demonstrate superior performance in retrieval on two datasets (ReMuQ and OK-VQA) under zero-shot settings as well as further improvements when finetuned on these datasets.

Multilingual Conceptual Coverage in Text-to-Image Models
Michael S Saxon and William Yang Wang   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We propose "Conceptual Coverage Across Languages" (CoCo-CroLa), a technique for benchmarking the degree to which any generative text-to-image system provides multilingual parity to its training language in terms of tangible nouns. For each model we can assess "conceptual coverage" of a given target language relative to a source language by comparing the population of images generated for a series of tangible nouns in the source language to the population of images generated for each noun under translation in the target language. This technique allows us to estimate how well-suited a model is to a target language as well as identify model-specific weaknesses, spurious correlations, and biases without a-priori assumptions. We demonstrate how it can be used to benchmark T2I models in terms of multilinguality, and how despite its simplicity it is a good proxy for impressive generalization.

Modular Visual Question Answering via Code Generation
Sanjay Subramanian, Medhini Narasimhan, Kushal M Khangaonkar, Kevin Yang, Arsha Nagrani, Cordelia Schmid, Andy Zeng, Trevor Darrell and Dan Klein   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We present a framework that formulates visual question answering as modular code generation. In contrast to prior work on modular approaches to VQA, our approach requires no additional training and relies on pre-trained language models (LMs), visual models pre-trained on image-caption pairs, and fifty VQA examples used for in-context learning. The generated Python programs invoke and compose the outputs of the visual models using arithmetic and conditional logic. Our approach improves accuracy on the COVR dataset by at least 3% and on the GQA dataset by 2% compared to the few-shot baseline that does not employ code generation.

Multimodal Relation Extraction with Cross-Modal Retrieval and Synthesis
Xuming Hu, Zhijiang Guo, ZHIYANG TENG, Irwin King and Philip S. Yu   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Multimodal relation extraction (MRE) is the task of identifying the semantic relationships between two entities based on the context of the sentence image pair. Existing retrieval-augmented approaches mainly focused on modeling the retrieved textual knowledge, but this may not be able to accurately identify complex relations. To improve the prediction, this research proposes to retrieve textual and visual evidence based on the object, sentence, and whole image. We further develop a novel approach to synthesize the object-level, image-level,

and sentence-level information for better reasoning between the same and different modalities. Extensive experiments and analyses show that the proposed method is able to effectively select and compare evidence across modalities and significantly outperforms state-of-the-art models.

A Theory of Unsupervised Speech Recognition
Liming Wang, Mark Hasegawa-Johnson and Chang D. Yoo   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Unsupervised speech recognition ({pasted macro 'ASRU'}/) is the problem of learning automatic speech recognition (ASR) systems from *unpaired* speech-only and text-only corpora. While various algorithms exist to solve this problem, a theoretical framework is missing to study their properties and address such issues as sensitivity to hyperparameters and training instability. In this paper, we proposed a general theoretical framework to study the properties of {pasted macro 'ASRU'}/ systems based on random matrix theory and the theory of neural tangent kernels. Such a framework allows us to prove various learnability conditions and sample complexity bounds of {pasted macro 'ASRU'}/. Extensive {pasted macro 'ASRU'}/ experiments on synthetic languages with three classes of transition graphs provide strong empirical evidence for our theory (code available at https://github.com/cactuswiththoughts/UnsupASRTheory.gitcactuswiththoughts/UnsupASRTheory.git).

When to Use Efficient Self Attention? Profiling Text, Speech and Image Transformer Variants
Anuj Diwan, Eunsol Choi and David Harwath   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
We present the first unified study of the efficiency of self-attention-based Transformer variants spanning text, speech and vision. We identify input length thresholds (tipping points) at which efficient Transformer variants become more efficient than vanilla models, using a variety of efficiency metrics (latency, throughput, and memory). To conduct this analysis for speech, we introduce L-HuBERT, a novel local-attention variant of a self-supervised speech model. We observe that these thresholds are (a) much higher than typical dataset sequence lengths and (b) dependent on the metric and modality, showing that choosing the right model depends on modality, task type (long-form vs. typical context) and resource constraints (time vs. memory). By visualising the breakdown of the computational costs for transformer components, we also show that non-self-attention components exhibit significant computational costs. We release our profiling toolkit at https://github.com/ajd12342/profiling-transformers .

Introducing Semantics into Speech Encoders
Derek Q Xu, Shuyan Annie Dong, Changhan Wang, Suyoun Kim, Zhaojiang Lin, Bing Liu, Akshat Shrivastava, Shang-Wen Li, Liang-Hsuan Tseng, Guan-Ting Lin, Alexei Baevski, Hung-yi Lee, Yizhou Sun and Wei Wang   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Recent studies find existing self-supervised speech encoders contain primarily acoustic rather than semantic information. As a result, pipelined supervised automatic speech recognition (ASR) to large language model (LLM) systems achieve state-of-the-art results on semantic spoken language tasks by utilizing rich semantic representations from the LLM. These systems come at the cost of labeled audio transcriptions, which is expensive and time-consuming to obtain. We propose a task-agnostic unsupervised way of incorporating semantic information from LLMs into self-supervised speech encoders without labeled audio transcriptions. By introducing semantics, we improve existing speech encoder spoken language understanding (SLU) performance by over 5% on intent classification (IC), with modest gains in named entity resolution (NER) and slot filling (SF), and spoken question answering (SQA) FF1 score by over 2%. Our approach, which uses no ASR data, achieves similar performance as methods trained on over 100 hours of labeled audio transcripts, demonstrating the feasibility of unsupervised semantic augmentations to existing speech encoders.

Back to Patterns: Efficient Japanese Morphological Analysis with Feature-Sequence Trie
Naoki Yoshinaga   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Accurate neural models are much less efficient than non-neural models and are useless for processing billions of social media posts or handling user queries in real time with a limited budget. This study revisits the fastest pattern-based NLP methods to make them as accurate as possible, thus yielding a strikingly simple yet surprisingly accurate morphological analyzer for Japanese. The proposed method induces reliable patterns from a morphological dictionary and annotated data. Experimental results on two standard datasets confirm that the method exhibits comparable accuracy to learning-based models, while boasting a remarkable throughput of over 1,000,000 sentences per second on a single modern CPU. The source code is available at https://www.tkl.iis.u-tokyo.ac.jp/ ynaga/jagger/

Federated Learning for Semantic Parsing: Task Formulation, Evaluation Setup, New Algorithms
Tianshu Zhang, Changchang Liu, Wei-Han Lee, Yu Su and Huan Sun   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
This paper studies a new task of federated learning (FL) for semantic parsing, where multiple clients collaboratively train one global model without sharing their semantic parsing data. By leveraging data from multiple clients, the FL paradigm can be especially beneficial for clients that have little training data to develop a data-hungry neural semantic parser on their own. We propose an evaluation setup to study this task, where we re-purpose widely-used single-domain text-to-SQL datasets as clients to form a realistic heterogeneous FL setting and collaboratively train a global model. As standard FL algorithms suffer from the high client heterogeneity in our realistic setup, we further propose a novel LOss Reduction Adjusted Re-weighting (Lorar) mechanism, which adjusts each client's contribution to the global model update based on its training loss reduction during each round. Our intuition is that the larger the loss reduction, the further away the current global model is from the client's local optimum, and the larger weight the client should get. By applying Lorar to three widely adopted FL algorithms (FedAvg, FedOPT and FedProx), we observe that their performance can be improved substantially on average (4%-20% absolute gain under MacroAvg) and that clients with smaller datasets enjoy larger performance gains. In addition, the global model converges faster for almost all the clients.

NatLogAttack: A Framework for Attacking Natural Language Inference Models with Natural Logic
Zi'ou Zheng and Xiaodan Zhu   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Reasoning has been a central topic in artificial intelligence from the beginning. The recent progress made on distributed representation and neural networks continues to improve the state-of-the-art performance of natural language inference. However, it remains an open question whether the models perform real reasoning to reach their conclusions or rely on spurious correlations. Adversarial attacks have proven to be an important tool to help evaluate the Achilles' heel of the victim models. In this study, we explore the fundamental problem of developing attack models based on logic formalism. We propose NatLogAttack to perform systematic attacks centring around natural logic, a classical logic formalism that is traceable back to Aristotle's syllogism and has been closely developed for natural language inference. The proposed framework renders both label-preserving and label-flipping attacks. We show that compared to the existing attack models, NatLogAttack generates better adversarial examples with fewer visits to the victim models. The victim models are found to be more vulnerable under the label-flipping setting. NatLogAttack provides a tool to probe the existing and future NLI models' capacity from a key viewpoint and we hope more logic-based attacks will be further explored for understanding the desired property of reasoning.

I2D2: Inductive Knowledge Distillation with NeuroLogic and Self-Imitation
Chandra Bhagavatula, Jena D. Hwang, Doug Downey, Ronan Le Bras, Ximing Lu, Lianhui Qin, Keisuke Sakaguchi, Swabha Swayamdipta, Peter West and Yejin Choi   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Commonsense capabilities of pre-trained language models dramatically improve with scale, leading many to believe that scale is the only winning recipe. But is it? Here, we investigate an alternative that a priori seems impossible: can smaller language models (e.g., GPT-2) win over models that are orders of magnitude larger and better (e.g., GPT-3), if powered with novel commonsense distillation algorithms? The

key intellectual challenge is to design a learning algorithm that achieve a competitive level of commonsense acquisition, without relying on the benefits of scale. In particular, we study generative models of commonsense knowledge, focusing on the task of generating generics, statements of commonsense facts about everyday concepts, e.g., birds can fly.

We introduce I2D2, a novel commonsense distillation framework that loosely follows the Symbolic Knowledge Distillation of West et al. but breaks the dependence on the extreme-scale teacher model with two innovations: (1) the novel adaptation of NeuroLogic Decoding to enhance the generation quality of the weak, off-the-shelf language models, and (2) self-imitation learning to iteratively learn from the model's own enhanced commonsense acquisition capabilities. Empirical results suggest that scale is not the only way, as novel algorithms can be a promising alternative. Moreover, our study leads to a new corpus of generics, Gen-A-tomic, that is the largest and highest quality available to date.

AMRs Assemble! Learning to Ensemble with Autoregressive Models for AMR Parsing
Abelardo Carlos Martínez Lorenzo, Pere Lluís Huguet Cabot and Roberto Navigli   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
In this paper, we examine the current state-of-the-art in AMR parsing, which relies on ensemble strategies by merging multiple graph predictions. Our analysis reveals that the present models often violate AMR structural constraints. To address this issue, we develop a validation method, and show how ensemble models can exploit SMATCH metric weaknesses to obtain higher scores, but sometimes result in corrupted graphs. Additionally, we highlight the demanding need to compute the SMATCH score among all possible predictions. To overcome these challenges, we propose two novel ensemble strategies based on Transformer models, improving robustness to structural constraints, while also reducing the computational time. Our methods provide new insights for enhancing AMR parsers and metrics. Our code is available at https://www.github.com/babelscape/AMRs-Assemble.

LAIT: Efficient Multi-Segment Encoding in Transformers with Layer-Adjustable Interaction
Jeremiah Milbauer, Annie Louis, Mohammad Javad Hosseini, Alex Fabrikant, Donald Metzler and Tal Schuster   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Transformer encoders contextualize token representations by attending to all other tokens at each layer, leading to quadratic increase in compute effort with the input length. In practice, however, the input text of many NLP tasks can be seen as a sequence of related segments (e.g., the sequence of sentences within a passage, or the hypothesis and premise in NLI). While attending across these segments is highly beneficial for many tasks, we hypothesize that this interaction can be delayed until later encoding stages.

To this end, we introduce Layer-Adjustable Interactions in Transformers (LAIT). Within LAIT, segmented inputs are first encoded independently, and then jointly. This partial two-tower architecture bridges the gap between a Dual Encoder's ability to pre-compute representations for segments and a fully self-attentive Transformer's capacity to model cross-segment attention. The LAIT framework effectively leverages existing pretrained Transformers and converts them into the hybrid of the two aforementioned architectures, allowing for easy and intuitive control over the performance-efficiency tradeoff. Experimenting on a wide range of NLP tasks, we find LAIT able to reduce 30-50% of the attention FLOPs on many tasks, while preserving high accuracy; in some practical settings, LAIT could reduce actual latency by orders of magnitude.

Ethical Considerations for Machine Translation of Indigenous Languages: Giving a Voice to the Speakers
Manuel Mager, Elisabeth Albine Mager, Katharina Kann and Ngoc Thang Vu   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
In recent years machine translation has become very successful for high-resource language pairs. This has also sparked new interest in research on the automatic translation of low-resource languages, including Indigenous languages. However, the latter are deeply related to the ethnic and cultural groups that speak (or used to speak) them. The data collection, modeling and deploying machine translation systems thus result in new ethical questions that must be addressed. Motivated by this, we first survey the existing literature on ethical considerations for the documentation, translation, and general natural language processing for Indigenous languages. Afterward, we conduct and analyze an interview study to shed light on the positions of community leaders, teachers, and language activists regarding ethical concerns for the automatic translation of their languages. Our results show that the inclusion, at different degrees, of native speakers and community members is vital to performing better and more ethical research on Indigenous languages.

A Crosslingual Investigation of Conceptualization in 1335 Languages
Yihong Liu, Haotian Ye, Leonie Weissweiler, Philipp Wicke, Renhao Pei, Robert Zangenfeind and Hinrich Schütze   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Languages differ in how they divide up the world into concepts and words; e.g., in contrast to English, Swahili has a single concept for 'belly' and 'womb'. We investigate these differences in conceptualization across 1,335 languages by aligning concepts in a parallel corpus. To this end, we propose Conceptualizer, a method that creates a bipartite directed alignment graph between source language concepts and sets of target language strings. In a detailed linguistic analysis across all languages for one concept ('bird') and an evaluation on gold standard data for 32 Swadesh concepts, we show that Conceptualizer has good alignment accuracy. We demonstrate the potential of research on conceptualization in NLP with two experiments. (1) We define crosslingual stability of a concept as the degree to which it has 1-1 correspondences across languages, and show that concreteness predicts stability. (2) We represent each language by its conceptualization pattern for 83 concepts, and define a similarity measure on these representations. The resulting measure for the conceptual similarity between two languages is complementary to standard genealogical, typological, and surface similarity measures. For four out of six language families, we can assign languages to their correct family based on conceptual similarity with accuracies between 54% and 87%

SLABERT Talk Pretty One Day: Modeling Second Language Acquisition with BERT
Aditya Yadavalli, Alekhya Yadavalli and Vera Tobin   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Second language acquisition (SLA) research has extensively studied cross-linguistic transfer, the influence of linguistic structure of a speaker's native language [L1] on the successful acquisition of a foreign language [L2]. Effects of such transfer can be positive (facilitating acquisition) or negative (impeding acquisition). We find that NLP literature has not given enough attention to the phenomenon of negative transfer. To understand patterns of both positive and negative transfer between L1 and L2, we model sequential second language acquisition in LMs. Further, we build a Mutlilingual Age Ordered CHILDES (MAO-CHILDES)—a dataset consisting of 5 typologically diverse languages, i.e., German, French, Polish, Indonesian, and Japanese—to understand the degree to which native Child-Directed Speech (CDS) [L1] can help or conflict with English language acquisition [L2]. To examine the impact of native CDS, we use the TILT-based cross lingual transfer learning approach established by Papadimitriou and Jurafsky (2020) and find that, as in human SLA, language family distance predicts more negative transfer. Additionally, we find that conversational speech data shows greater facilitation for language acquisition than scripted speech data. Our findings call for further research using our novel Transformer-based SLA models and we would like to encourage it by releasing our code, data, and models.

Diversity-Aware Coherence Loss for Improving Neural Topic Models
Raymond Li, Felipe Gonzalez-Pizarro, Linzi Xing, Gabriel Murray and Giuseppe Carenini   09:00-10:30 (Frontenac Ballroom and Queen's Quay)
The standard approach for neural topic modeling uses a variational autoencoder (VAE) framework that jointly minimizes the KL divergence between the estimated posterior and prior, in addition to the reconstruction loss. Since neural topic models are trained by recreating indi-

vidual input documents, they do not explicitly capture the coherence between words on the corpus level. In this work, we propose a novel diversity-aware coherence loss that encourages the model to learn corpus-level coherence scores while maintaining high diversity between topics. Experimental results on multiple datasets show that our method significantly improves the performance of neural topic models without requiring any pretraining or additional parameters.

DecompX: Explaining Transformers Decisions by Propagating Token Decomposition
Ali Modarressi, Mohsen Fayyaz, Ehsan Aghazadeh, Yadollah Yaghoobzadeh and Mohammad Taher Pilehvar 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
An emerging solution for explaining Transformer-based models is to use vector-based analysis on how the representations are formed. However, providing a faithful vector-based explanation for a multi-layer model could be challenging in three aspects: (1) Incorporating all components into the analysis, (2) Aggregating the layer dynamics to determine the information flow and mixture throughout the entire model, and (3) Identifying the connection between the vector-based analysis and the model's predictions. In this paper, we present DecompX to tackle these challenges. DecompX is based on the construction of decomposed token representations and their successive propagation throughout the model without mixing them in between layers. Additionally, our proposal provides multiple advantages over existing solutions for its inclusion of all encoder components (especially nonlinear feed-forward networks) and the classification head. The former allows acquiring precise vectors while the latter transforms the decomposition into meaningful prediction-based values, eliminating the need for norm- or summation-based vector aggregation. According to the standard faithfulness evaluations, DecompX consistently outperforms existing gradient-based and vector-based approaches on various datasets. Our code is available at https://github.com/mohsenfayyaz/DecompX.

Language Detoxification with Attribute-Discriminative Latent Space
Jin Myung Kwak, Minseon Kim and Sung Ju Hwang 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Transformer-based Language Models (LMs) have achieved impressive results on natural language understanding tasks, but they can also generate toxic text such as insults, threats, and profanity, limiting their real-world applications. To overcome this issue, a few text generation approaches aim to detoxify toxic texts using additional LMs or perturbations. However, previous methods require excessive memory, computations, and time which are serious bottlenecks in their real-world application. To address such limitations, we propose an effective yet efficient method for language detoxification using an attribute-discriminative latent space. Specifically, we project the latent space of an original Transformer LM onto a discriminative latent space that well-separates texts by their attributes using a projection block and an attribute discriminator. This allows the LM to control the text generation to be non-toxic with minimal memory and computation overhead. We validate our model, Attribute-Discriminative Language Model (ADLM) on detoxified language and dialogue generation tasks, on which our method significantly outperforms baselines both in performance and efficiency.

Morphological Inflection: A Reality Check
Jordan Kodner, Sarah Payne, Salam Khalifa and Zoey Liu 09:00-10:30 (Frontenac Ballroom and Queen's Quay)
Morphological inflection is a popular task in sub-word NLP with both practical and cognitive applications. For years now, state-of-the-art systems have reported high, but also highly variable, performance across data sets and languages. We investigate the causes of this high performance and high variability; we find several aspects of data set creation and evaluation which systematically inflate performance and obfuscate differences between languages. To improve generalizability and reliability of results, we propose new data sampling and evaluation strategies that better reflect likely use-cases. Using these new strategies, we make new observations on the generalization abilities of current inflection systems.

Semantics: Sentence-level Semantics, Textual Inference, and Other Areas

09:00-10:30 (Pier 2&3)

[TACL] Compositional Evaluation on Japanese Textual Entailment and Similarity
Hitomi Yanaka and Koji Mineshima 09:00-09:15 (Pier 2&3)
Natural Language Inference (NLI) and Semantic Textual Similarity (STS) are widely used benchmark tasks for compositional evaluation of pre-trained language models. Despite growing interest in linguistic universals, most NLI/STS studies have focused almost exclusively on English. In particular, there are no available multilingual NLI/STS datasets in Japanese, which is typologically different from English and can shed light on the currently controversial behavior of language models in matters such as sensitivity to word order and case particles. Against this background, we introduce JSICK, a Japanese NLI/STS dataset that was manually translated from the English dataset SICK. We also present a stress-test dataset for compositional inference, created by transforming syntactic structures of sentences in JSICK to investigate whether language models are sensitive to word order and case particles. We conduct baseline experiments on different pre-trained language models and compare the performance of multilingual models when applied to Japanese and other languages. The results of the stress-test experiments suggest that the current pre-trained language models are insensitive to word order and case marking.

[TACL] Scientia Potentia Est – On the Role of Knowledge in Computational Argumentation
Anne Lauscher, Henning Wachsmuth, Iryna Gurevych and Goran Glavaš 09:15-09:30 (Pier 2&3)
Despite extensive research efforts in recent years, computational argumentation (CA) remains one of the most challenging areas of natural language processing. The reason for this is the inherent complexity of the cognitive processes behind human argumentation, which integrate a plethora of different types of knowledge, ranging from topic-specific facts and common sense to rhetorical knowledge. The integration of knowledge from such a wide range in CA requires modeling capabilities far beyond many other natural language understanding tasks. Existing research on mining, assessing, reasoning over, and generating arguments largely acknowledges that much more knowledge is needed to accurately model argumentation computationally. However, a systematic overview of the types of knowledge introduced in existing CA models is missing, hindering targeted progress in the field. Adopting the operational definition of knowledge as any task-relevant normative information not provided as input, the survey paper at hand fills this gap by (1) proposing a taxonomy of types of knowledge required in CA tasks, (2) systematizing the large body of CA work according to the reliance on and exploitation of these knowledge types for the four main research areas in CA, and (3) outlining and discussing directions for future research efforts in CA.

Dense-ATOMIC: Towards Densely-connected ATOMIC with High Knowledge Coverage and Massive Multi-hop Paths
Xiangqing Shen, Siwei Wu and Rui Xia 09:30-09:45 (Pier 2&3)
ATOMIC is a large-scale commonsense knowledge graph (CSKG) containing everyday if-then knowledge triplets, i.e., head event, relation, tail event. The one-hop annotation manner made ATOMIC a set of independent bipartite graphs, which ignored the numerous links between events in different bipartite graphs and consequently caused shortages in knowledge coverage and multi-hop paths. In this work, we aim to construct Dense-ATOMIC with high knowledge coverage and massive multi-hop paths. The events in ATOMIC are normalized to a consistent pattern at first. We then propose a CSKG completion method called Rel-CSKGC to predict the relation given the head event and the tail event

of a triplet, and train a CSKG completion model based on existing triplets in ATOMIC. We finally utilize the model to complete the missing links in ATOMIC and accordingly construct Dense-ATOMIC. Both automatic and human evaluation on an annotated subgraph of ATOMIC demonstrate the advantage of Rel-CSKGC over strong baselines. We further conduct extensive evaluations on Dense-ATOMIC in terms of statistics, human evaluation, and simple downstream tasks, all proving Dense-ATOMIC's advantages in Knowledge Coverage and Multi-hop Paths. Both the source code of Rel-CSKGC and Dense-ATOMIC are publicly available on https://github.com/NUSTM/Dense-ATOMIC.

COLA: Contextualized Commonsense Causal Reasoning from the Causal Inference Perspective
Zhaowei Wang, Quyet V. Do, Hongming Zhang, Jiayao Zhang, Weiqi Wang, Tianqing Fang, Yangqiu Song, Ginny Y. Wong and Simon See
09:45-10:00 (Pier 2&3)
Detecting commonsense causal relations (causation) between events has long been an essential yet challenging task. Given that events are complicated, an event may have different causes under various contexts. Thus, exploiting context plays an essential role in detecting causal relations. Meanwhile, previous works about commonsense causation only consider two events and ignore their context, simplifying the task formulation. This paper proposes a new task to detect commonsense causation between two events in an event sequence (i.e., context), called contextualized commonsense causal reasoning. We also design a zero-shot framework: COLA (Contextualized Commonsense Causality Reasoner) to solve the task from the causal inference perspective. This framework obtains rich incidental supervision from temporality and balances covariates from multiple timestamps to remove confounding effects. Our extensive experiments show that COLA can detect commonsense causality more accurately than baselines.

CAT: A Contextualized Conceptualization and Instantiation Framework for Commonsense Reasoning
Weiqi Wang, Tianqing Fang, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song and Lei Chen
10:00-10:15 (Pier 2&3)
Commonsense reasoning, aiming at endowing machines with a human-like ability to make situational presumptions, is extremely challenging to generalize. For someone who barely knows about "meditation," while is knowledgeable about "singing," he can still infer that "meditation makes people relaxed" from the existing knowledge that "singing makes people relaxed" by first conceptualizing "singing" as a "relaxing event" and then instantiating that event to "meditation." This process, known as conceptual induction and deduction, is fundamental to commonsense reasoning while lacking both labeled data and methodologies to enhance commonsense modeling. To fill such a research gap, we propose CAT (Contextualized ConceptuAlization and InsTantiation), a semi-supervised learning framework that integrates event conceptualization and instantiation to conceptualize commonsense knowledge bases at scale. Extensive experiments show that our framework achieves state-of-the-art performances on two conceptualization tasks, and the acquired abstract commonsense knowledge can significantly improve commonsense inference modeling. Our code, data, and fine-tuned models are publicly available at https://github.com/HKUST-KnowComp/CAT.

[CL] Curing the SICK and other NLI maladies
Aikaterini-Lida Kalouli, Hai Hu, Alexander Webb, Lawrence Moss and Valeria Paiva
10:15-10:30 (Pier 2&3)
Against the backdrop of the ever-improving Natural Language Inference (NLI) models, recent efforts have focused on the suitability of the current NLI datasets and on the feasibility of the NLI task as it is currently approached. Many of the recent studies have exposed the inherent human disagreements of the inference task and have proposed a shift from categorical labels to human subjective probability assessments, capturing human uncertainty. In this work, we show how neither the current task formulation nor the proposed uncertainty gradient are entirely suitable for solving the NLI challenges. Instead, we propose an ordered sense space annotation, which distinguishes between logical and common-sense inference. One end of the space captures non-sensical inferences, while the other end represents strictly logical scenarios. In the middle of the space, we find a continuum of common-sense, namely, the subjective and graded opinion of a "person on the street." To arrive at the proposed annotation scheme, we perform a careful investigation of the SICK corpus and we create a taxonomy of annotation issues and guidelines. We re-annotate the corpus with the proposed annotation scheme, utilizing four symbolic inference systems, and then perform a thorough evaluation of the scheme by fine-tuning and testing commonly used pre-trained language models on the re-annotated SICK within various settings. We also pioneer a crowd annotation of a small portion of the MultiNLI corpus, showcasing that it is possible to adapt our scheme for annotation by non-experts on another NLI corpus. Our work shows the efficiency and benefits of the proposed mechanism and opens the way for a careful NLI task refinement.

Industry track: Interactive Systems, Speech

09:00-10:30 (Pier 4&5)

[Industry] Accurate Training of Web-based Question Answering Systems with Feedback from Ranked Users
Liang Wang, Ivano Lauriola and Alessandro Moschitti
09:00-09:15 (Pier 4&5)
Recent work has shown that large-scale annotated datasets are essential for training state-of-the-art Question Answering (QA) models. Unfortunately, creating this data is expensive and requires a huge amount of annotation work. An alternative and cheaper source of supervision is given by feedback data collected from deployed QA systems. This data can be collected from tens of millions of user with no additional cost, for real-world QA services, e.g., Alexa, Google Home, and users. The main drawback is the noise affecting feedback on individual examples. Recent literature on QA systems has shown the benefit of training models even with noisy feedback. However, these studies have multiple limitations: (i) they used uniform random noise to simulate feedback responses, which is typically an unrealistic approximation as noise follows specific patterns, depending on target examples and users; and (ii) they do not show how to aggregate feedback for improving training signals. In this paper, we first collect a large scale (16M) QA dataset with real feedback sampled from the QA traffic of a popular Virtual Assistant. Second, we use this data to develop two strategies for filtering unreliable users and thus de-noise feedback: (i) ranking users with an automatic classifier, and (ii) aggregating feedback over similar instances and comparing users between each other. Finally, we train QA models on our filtered feedback data, showing a significant improvement over the state of the art.

[Industry] Reliable and Interpretable Drift Detection in Streams of Short Texts
Ella Rabinovich, Matan Vetzler, Samuel Ackerman and Ateret Anaby Tavor
09:15-09:30 (Pier 4&5)
Data drift is the change in model input data that is one of the key factors leading to machine learning models performance degradation over time. Monitoring drift helps detecting these issues and preventing their harmful consequences. Meaningful drift interpretation is a fundamental step towards effective re-training of the model. In this study we propose an end-to-end framework for reliable model-agnostic change-point detection and interpretation in large task-oriented dialog systems, proven effective in multiple customer deployments. We evaluate our approach and demonstrate its benefits with a novel variant of intent classification training dataset, simulating customer requests to a dialog system. We make the data publicly available.

[Industry] Answering Unanswered Questions through Semantic Reformulations in Spoken QA
Pedro Faustini, Zhiyu Chen, Besnik Fetahu, Oleg Rokhlenko and Shervin Malmasi
09:30-09:45 (Pier 4&5)

Spoken Question Answering (QA) is a key feature of voice assistants, usually backed by multiple QA systems. Users ask questions via spontaneous speech that can contain disfluencies, errors, and informal syntax or phrasing. This is a major challenge in QA, causing unanswered questions or irrelevant answers, leading to bad user experiences. We analyze failed QA requests to identify core challenges: lexical gaps, proposition types, complex syntactic structure, and high specificity. We propose a Semantic Question Reformulation (SURF) model offering three linguistically-grounded operations (repair, syntactic reshaping, generalization) to rewrite questions to facilitate answering. Offline evaluation on 1M unanswered questions from a leading voice assistant shows that SURF significantly improves answer rates: up to 24% of previously unanswered questions obtain relevant answers (75%). Live deployment shows positive impact for millions of customers with unanswered questions; explicit relevance feedback shows high user satisfaction.

[Industry] Sharing Encoder Representations across Languages, Domains and Tasks in Large-Scale Spoken Language Understanding
Jonathan Hueser, Judith Gaspers, Thomas Gueudre, Chandana Prakash, Jin Cao, Daniil Sorokin, Quynh Do, Nicolas Anastassacos, Tobias Falke and Turan Gojayev   09:45-10:00 (Pier 4&5)
Leveraging representations from pre-trained transformer-based encoders achieves state-of-the-art performance on numerous NLP tasks. Larger encoders can improve accuracy for spoken language understanding (SLU) but are challenging to use given the inference latency constraints of online systems (especially on CPU machines). We evaluate using a larger 170M parameter BERT encoder that shares representations across languages, domains and tasks for SLU compared to using smaller 17M parameter BERT encoders with language-, domain- and task-decoupled finetuning. Running inference with a larger shared encoder on GPU is latency neutral and reduces infrastructure cost compared to running inference for decoupled smaller encoders on CPU machines. The larger shared encoder reduces semantic error rates by 4.62% for test sets representing user requests to voice-controlled devices and 5.79% on the tail of the test sets on average across four languages.

[Industry] Regression-Free Model Updates for Spoken Language Understanding
Andrea Caciolai, Verena Weber, Tobias Falke, Alessandro Pedrani and Davide Bernardi   10:00-10:15 (Pier 4&5)
In real-world systems, an important requirement for model updates is to avoid regressions in user experience caused by flips of previously correct classifications to incorrect ones. Multiple techniques for that have been proposed in the recent literature. In this paper, we apply one such technique, focal distillation, to model updates in a goal-oriented dialog system and assess its usefulness in practice. In particular, we evaluate its effectiveness for key language understanding tasks, including sentence classification and sequence labeling tasks, we further assess its effect when applied to repeated model updates over time, and test its compatibility with mislabeled data. Our experiments on a public benchmark and data from a deployed dialog system demonstrate that focal distillation can substantially reduce regressions, at only minor drops in accuracy, and that it further outperforms naive supervised training in challenging mislabeled data and label expansion settings.

[Industry] "Let's not Quote out of Context": Unified Vision-Language Pretraining for Context Assisted Image Captioning
Abisek Rajakumar Kalarani, Pushpak Bhattacharyya, Niyati Chhaya and Sumit Shekhar   10:15-10:30 (Pier 4&5)
Well-formed context aware image captions and tags in enterprise content such as marketing material are critical to ensure their brand presence and content recall. Manual creation and updates to ensure the same is non trivial given the scale and the tedium towards this task. We propose a new unified Vision-Language (VL) model based on the One For All (OFA) model, with a focus on context-assisted image captioning where the caption is generated based on both the image and its context. Our approach aims to overcome the context-independent (image and text are treated independently) nature of the existing approaches. We exploit context by pretraining our model with datasets of three tasks- news image captioning where the news article is the context, contextual visual entailment, and keyword extraction from the context. The second pretraining task is a new VL task, and we construct and release two datasets for the task with 1.1M and 2.2K data instances. Our system achieves state-of-the-art results with an improvement of up to 8.34 CIDEr score on the benchmark news image captioning datasets. To the best of our knowledge, ours is the first effort at incorporating contextual information in pretraining the models for the VL tasks.

Phonology, Morphology, and Word Segmentation

09:00-10:30 (Pier 7&8)

What is the best recipe for character-level encoder-only modelling?
Kris Cao
                                                                                                                    09:00-09:15 (Pier 7&8)
This paper aims to benchmark recent progress in language understanding models that output contextualised representations at the character level. Many such modelling architectures and methods to train those architectures have been proposed, but it is currently unclear what the relative contributions of the architecture vs. the pretraining objective are to final model performance. We explore the design space of such models, comparing architectural innovations (Clark et al., 2022, Jaegle et al., 2022, Tay et al., 2021) and a variety of different pretraining objectives on a suite of evaluation tasks with a fixed training procedure in order to find the currently optimal way to build and train character-level BERT-like models. We find that our best performing character-level model exceeds the performance of a token-based model trained with the same settings on the same data, suggesting that character-level models are ready for more widespread adoption. Unfortunately, the best method to train character-level models still relies on a subword-level tokeniser during pretraining, and final model performance is highly dependent on tokeniser quality. We believe our results demonstrate the readiness of character-level models for multilingual language representation, and encourage NLP practitioners to try them as drop-in replacements for token-based models.

Bi-Phone: Modeling Inter Language Phonetic Influences in Text
Abhirut Gupta, Ananya B. Sai, Richard Sproat, Yuri Vasilevski, James S Ren, Ambarish Jash, Sukhdeep S Sodhi and Aravindan Raghuveer
09:15-09:30 (Pier 7&8)
A large number of people are forced to use the Web in a language they have low literacy in due to technology asymmetries. Written text in the second language (L2) from such users often contains a large number of errors that are influenced by their native language (L1). We propose a method to mine phoneme confusions (sounds in L2 that an L1 speaker is likely to conflate) for pairs of L1 and L2. These confusions are then plugged into a generative model (Bi-Phone) for synthetically producing corrupted L2 text. Through human evaluations, we show that Bi-Phone generates plausible corruptions that differ across L1s and also have widespread coverage on the Web. We also corrupt the popular language understanding benchmark SuperGLUE with our technique (FunGLUE for Phonetically Noised GLUE) and show that SoTA language understating models perform poorly. We also introduce a new phoneme prediction pre-training task which helps byte models to recover performance close to SuperGLUE. Finally, we also release the SuperGLUE benchmark to promote further research in phonetically robust language models. To the best of our knowledge, FunGLUE is the first benchmark to introduce L1-L2 interactions in text.

Transformed Protoform Reconstruction
Young Min Kim, Kalvin Chang, Chenxuan Cui and David R. Mortensen   09:30-09:45 (Pier 7&8)
Protoform reconstruction is the task of inferring what morphemes or words appeared like in the ancestral languages of a set of daughter languages. Meloni et al (2021) achieved the state-of-the-art on Latin protoform reconstruction with an RNN-based encoder-decoder with

attention model. We update their model with the state-of-the-art seq2seq model: the Transformer. Our model outperforms their model on a suite of different metrics on two different datasets: their Romance data of 8,000 cognates spanning 5 languages and a Chinese dataset (Hou 2004) of 800+ cognates spanning 39 varieties. We also probe our model for potential phylogenetic signal contained in the model. Our code is publicly available at https://github.com/cmu-llab/acl-2023.

Session 7 - 11:00-12:30

Resources and Evaluation

11:00-12:30 (Metropolitan East)

WikiHowQA: A Comprehensive Benchmark for Multi-Document Non-Factoid Question Answering

Valeriia Bolotova-Baranova, Vladislav Blinov, Sofya Filippova, Falk Scholer and Mark Sanderson   11:00-11:15 (Metropolitan East)
Answering non-factoid questions (NFQA) is a challenging task, requiring passage-level answers that are difficult to construct and evaluate. Search engines may provide a summary of a single web page, but many questions require reasoning across multiple documents. Meanwhile, modern models can generate highly coherent and fluent, but often factually incorrect answers that can deceive even non-expert humans. There is a critical need for high-quality resources for multi-document NFQA (MD-NFQA) to train new models and evaluate answers' grounding and factual consistency in relation to supporting documents.
To address this gap, we introduce WikiHowQA, a new multi-document NFQA benchmark built on WikiHow, a website dedicated to answering "how-to" questions. The benchmark includes 11,746 human-written answers along with 74,527 supporting documents. We describe the unique challenges of the resource, provide strong baselines, and propose a novel human evaluation framework that utilizes highlighted relevant supporting passages to mitigate issues such as assessor unfamiliarity with the question topic. All code and data, including the automatic code for preparing the human evaluation, are publicly available.

Toward Human-Like Evaluation for Natural Language Generation with Error Analysis

Qingyu Lu, Liang Ding, Liping Xie, Kanjian Zhang, Derek F. Wong and Dacheng Tao   11:15-11:30 (Metropolitan East)
The pretrained language model (PLM) based metrics have been successfully used in evaluating language generation tasks. Recent studies of the human evaluation community show that considering both major errors (e.g. mistranslated tokens) and minor errors (e.g. imperfections in fluency) can produce high-quality judgments. This inspires us to approach the final goal of the automatic metrics (human-like evaluations) by fine-grained error analysis. In this paper, we argue that the ability to estimate sentence confidence is the tip of the iceberg for PLM-based metrics. And it can be used to refine the generated sentence toward higher confidence and more reference-grounded, where the costs of refining and approaching reference are used to determine the major and minor errors, respectively. To this end, we take BARTScore as the testbed and present an innovative solution to marry the unexploited sentence refining capacity of BARTScore and human-like error analysis, where the final score consists of both the evaluations of major and minor errors. Experiments show that our solution consistently and significantly improves BARTScore, and outperforms top-scoring metrics in 19/25 test settings. Analyses demonstrate our method robustly and efficiently approaches human-like evaluations, enjoying better interpretability. Our code and scripts will be publicly released in https://github.com/Coldmist-Lu/ErrorAnalysis_NLGEvaluation.

SQuARe: A Large-Scale Dataset of Sensitive Questions and Acceptable Responses Created through Human-Machine Collaboration

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Meeyoung Cha, Yejin Choi, BYOUNGPIL KIM, Gunhee Kim, Eun-Ju Lee, Yong Lim, Alice Oh, Sangchul Park and Jung-Woo Ha   11:30-11:45 (Metropolitan East)
The potential social harms that large language models pose, such as generating offensive content and reinforcing biases, are steeply rising. Existing works focus on coping with this concern while interacting with ill-intentioned users, such as those who explicitly make hate speech or elicit harmful responses. However, discussions on sensitive issues can become toxic even if the users are well-intentioned. For safer models in such scenarios, we present the Sensitive Questions and Acceptable Response (SQuARe) dataset, a large-scale Korean dataset of 49k sensitive questions with 42k acceptable and 46k non-acceptable responses. The dataset was constructed leveraging HyperCLOVA in a human-in-the-loop manner based on real news headlines. Experiments show that acceptable response generation significantly improves for HyperCLOVA and GPT-3, demonstrating the efficacy of this dataset.

QUEST: A Retrieval Dataset of Entity-Seeking Queries with Implicit Set Operations

Chaitanya Malaviya, Peter Shaw, Ming-Wei Chang, Kenton Lee and Kristina Toutanova   11:45-12:00 (Metropolitan East)
Formulating selective information needs results in queries that implicitly specify set operations, such as intersection, union, and difference. For instance, one might search for "shorebirds that are not sandpipers" or "science-fiction films shot in England". To study the ability of retrieval systems to meet such information needs, we construct QUEST, a dataset of 3357 natural language queries with implicit set operations, that map to a set of entities corresponding to Wikipedia documents. The dataset challenges models to match multiple constraints mentioned in queries with corresponding evidence in documents and correctly perform various set operations. The dataset is constructed semi-automatically using Wikipedia category names. Queries are automatically composed from individual categories, then paraphrased and further validated for naturalness and fluency by crowdworkers. Crowdworkers also assess the relevance of entities based on their documents and highlight attribution of query constraints to spans of document text. We analyze several modern retrieval systems, finding that they often struggle on such queries. Queries involving negation and conjunction are particularly challenging and systems are further challenged with combinations of these operations.

A Critical Evaluation of Evaluations for Long-form Question Answering

Fangyuan Xu, Yixiao Song, Mohit Iyyer and Eunsol Choi   12:00-12:15 (Metropolitan East)
Long-form question answering (LFQA) enables answering a wide range of questions, but its flexibility poses enormous challenges for evaluation. We perform the first targeted study of the evaluation of long-form answers, covering both human and automatic evaluation practices. We hire domain experts in seven areas to provide preference judgments over pairs of answers, along with free-form justifications for their choices. We present a careful analysis of experts' evaluation, which focuses on new aspects such as the comprehensiveness of the answer. Next, we examine automatic text generation metrics, finding that no existing metrics are predictive of human preference judgments. However, some metrics correlate with fine-grained aspects of answers (e.g., coherence). We encourage future work to move away from a single "overall score" of the answer and adopt a multi-faceted evaluation, targeting aspects such as factuality and completeness. We publicly release all of our annotations and code to spur future work into LFQA evaluation.

Are Human Explanations Always Helpful? Towards Objective Evaluation of Human Natural Language Explanations

Bingsheng Yao, Prithviraj Sen, Lucian Popa, James Hendler and Dakuo Wang   12:15-12:30 (Metropolitan East)
Human-annotated labels and explanations are critical for training explainable NLP models. However, unlike human-annotated labels whose quality is easier to calibrate (e.g., with a majority vote), human-crafted free-form explanations can be quite subjective. Before blindly using them as ground truth to train ML models, a vital question needs to be asked: How do we evaluate a human-annotated explanation's quality? In this paper, we build on the view that the quality of a human-annotated explanation can be measured based on its helpfulness (or impairment) to the ML models' performance for the desired NLP tasks for which the annotations were collected. In comparison to the commonly used Simulatability score, we define a new metric that can take into consideration the helpfulness of an explanation for model performance at both fine-tuning and inference. With the help of a unified dataset format, we evaluated the proposed metric on five datasets (e.g., e-SNLI) against two model architectures (T5 and BART), and the results show that our proposed metric can objectively evaluate the quality of human-annotated explanations, while Simulatability falls short.

Information Extraction / Generation

11:00-12:30 (Metropolitan Centre)

Consistent Prototype Learning for Few-Shot Continual Relation Extraction
Xiudi Chen, Hui Wu and xiaodong shi   11:00-11:15 (Metropolitan Centre)
Few-shot continual relation extraction aims to continually train a model on incrementally few-shot data to learn new relations while avoiding forgetting old ones. However, current memory-based methods are prone to overfitting memory samples, resulting in insufficient activation of old relations and limited ability to handle the confusion of similar classes. In this paper, we design a new N-way-K-shot Continual Relation Extraction (NK-CRE) task and propose a novel few-shot continual relation extraction method with Consistent Prototype Learning (ConPL) to address the aforementioned issues. Our proposed ConPL is mainly composed of three modules: 1) a prototype-based classification module that provides primary relation predictions under few-shot continual learning; 2) a memory-enhanced module designed to select vital samples and refined prototypical representations as a novel multi-information episodic memory; 3) a consistent learning module to reduce catastrophic forgetting by enforcing distribution consistency. To effectively mitigate catastrophic forgetting, ConPL ensures that the samples and prototypes in the episodic memory remain consistent in terms of classification and distribution. Additionally, ConPL uses prompt learning to extract better representations and adopts a focal loss to alleviate the confusion of similar classes. Experimental results on two commonly-used datasets show that our model consistently outperforms other competitive baselines.

Uncertainty Guided Label Denoising for Document-level Distant Relation Extraction
Qi Sun, Kun Huang, Xiaocui Yang, Pengfei Hong, Kun Zhang and Soujanya Poria   11:15-11:30 (Metropolitan Centre)
Document-level relation extraction (DocRE) aims to infer complex semantic relations among entities in a document. Distant supervision (DS) is able to generate massive auto-labeled data, which can improve DocRE performance. Recent works leverage pseudo labels generated by the pre-denoising model to reduce noise in DS data. However, unreliable pseudo labels bring new noise, e.g., adding false pseudo labels and losing correct DS labels. Therefore, how to select effective pseudo labels to denoise DS data is still a challenge in document-level distant relation extraction. To tackle this issue, we introduce uncertainty estimation technology to determine whether pseudo labels can be trusted. In this work, we propose a Document-level distant Relation Extraction framework with Uncertainty Guided label denoising, UGDRE. Specifically, we propose a novel instance-level uncertainty estimation method, which measures the reliability of the pseudo labels with overlapping relations. By further considering the long-tail problem, we design dynamic uncertainty thresholds for different types of relations to filter high-uncertainty pseudo labels. We conduct experiments on two public datasets. Our framework outperforms strong baselines by 1.91 F1 and 2.28 Ign F1 on the RE-DocRED dataset.

Improving Continual Relation Extraction by Distinguishing Analogous Semantics
Wenzheng Zhao, Yuanning Cui and Wei Hu   11:30-11:45 (Metropolitan Centre)
Continual relation extraction (RE) aims to learn constantly emerging relations while avoiding forgetting the learned relations. Existing works store a small number of typical samples to re-train the model for alleviating forgetting. However, repeatedly replaying these samples may cause the overfitting problem. We conduct an empirical study on existing works and observe that their performance is severely affected by analogous relations. To address this issue, we propose a novel continual extraction model for analogous relations. Specifically, we design memory-insensitive relation prototypes and memory augmentation to overcome the overfitting problem. We also introduce integrated training and focal knowledge distillation to enhance the performance on analogous relations. Experimental results show the superiority of our model and demonstrate its effectiveness in distinguishing analogous relations and overcoming overfitting.

On the Blind Spots of Model-Based Evaluation Metrics for Text Generation
Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass and Yulia Tsvetkov   11:45-12:00 (Metropolitan Centre)
In this work, we explore a useful but often neglected methodology for robustness analysis of text generation evaluation metrics: stress tests with synthetic data. Basically, we design and synthesize a wide range of potential errors and check whether they result in a commensurate drop in the metric scores. We examine a range of recently proposed evaluation metrics based on pretrained language models, for the tasks of open-ended generation, translation, and summarization. Our experiments reveal interesting insensitivities, biases, or even loopholes in existing metrics. For example, we find that BERTScore is confused by truncation errors in summarization, and MAUVE (built on top of GPT-2) is insensitive to errors at the beginning or middle of generations. Further, we investigate the reasons behind these blind spots and suggest practical workarounds for a more reliable evaluation of text generation. We have released our code and data at https://github.com/cloudygoose/blindspot_nlg.

COGEN: Abductive Commonsense Language Generation
rohola zandie, Diwanshu Shekhar and Mohammad Mahoor   12:00-12:15 (Metropolitan Centre)
Reasoning is one of the most important elements in achieving Artificial General Intelligence (AGI), specifically when it comes to Abductive and counterfactual reasoning. In order to introduce these capabilities of reasoning in Natural Language Processing (NLP) models, there have been recent advances towards training NLP models to better perform on two main tasks - Abductive Natural Language Inference (alphaNLI) and Abductive Natural Language Generation Task (alphaNLG). This paper proposes CoGen, a model for both alphaNLI and alphaNLG tasks that employ a novel approach of combining the temporal commonsense reasoning for each observation (before and after a real hypothesis) from pre-trained models with contextual filtering for training. Additionally, we use state-of-the-art semantic entailment to filter out the contradictory hypothesis during the inference. Our experimental results show that CoGen outperforms current models and set a new state of the art in regards to alphaNLI and alphaNLG tasks. We make the source code of CoGen model publicly available for reproducibility and to facilitate relevant future research.

Reward Gaming in Conditional Text Generation
Richard Yuanzhe Pang, Vishakh Padmakumar, Thibault Sellam, Ankur Parikh and He He   12:15-12:30 (Metropolitan Centre)
To align conditional text generation model outputs with desired behaviors, there has been an increasing focus on training the model using reinforcement learning (RL) with reward functions learned from human annotations. Under this framework, we identify three common cases where high rewards are incorrectly assigned to undesirable patterns: noise-induced spurious correlation, naturally occurring spurious correlation, and covariate shift. We show that even though learned metrics achieve high performance on the distribution of the data used to train the reward function, the undesirable patterns may be amplified during RL training of the text generation model. While there has been discussion about reward gaming in the RL or safety community, in this discussion piece, we would like to highlight reward gaming in the natural language generation (NLG) community using concrete conditional text generation examples and discuss potential fixes and areas for future work.

Information Retrieval and Text Mining

11:00-12:30 (Metropolitan West)

BERM: Training the Balanced and Extractable Representation for Matching to Improve Generalization Ability of Dense Retrieval
Shicheng Xu, Liang Pang, Huawei Shen and Xueqi Cheng   11:00-11:15 (Metropolitan West)
Dense retrieval has shown promise in the first-stage retrieval process when trained on in-domain labeled datasets. However, previous studies have found that dense retrieval is hard to generalize to unseen domains due to its weak modeling of domain-invariant and interpretable feature (i.e., matching signal between two texts, which is the essence of information retrieval). In this paper, we propose a novel method to improve the generalization of dense retrieval via capturing matching signal called BERM. Fully fine-grained expression and query-oriented saliency are two properties of the matching signal. Thus, in BERM, a single passage is segmented into multiple units and two unit-level requirements are proposed for representation as the constraint in training to obtain the effective matching signal. One is semantic unit balance and the other is essential matching unit extractability. Unit-level view and balanced semantics make representation express the text in a fine-grained manner. Essential matching unit extractability makes passage representation sensitive to the given query to extract the pure matching information from the passage containing complex context. Experiments on BEIR show that our method can be effectively combined with different dense retrieval training methods (vanilla, hard negatives mining and knowledge distillation) to improve its generalization ability without any additional inference overhead and target domain data.

ConvGQR: Generative Query Reformulation for Conversational Search
Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang and Jian-Yun Nie   11:15-11:30 (Metropolitan West)
In conversational search, the user's real search intent for the current conversation turn is dependent on the previous conversation history. It is challenging to determine a good search query from the whole conversation context. To avoid the expensive re-training of the query encoder, most existing methods try to learn a rewriting model to de-contextualize the current query by mimicking the manual query rewriting. However, manually rewritten queries are not always the best search queries. Thus, training a rewriting model on them would lead to sub-optimal queries. Another useful information to enhance the search query is the potential answer to the question. In this paper, we propose ConvGQR, a new framework to reformulate conversational queries based on generative pre-trained language models (PLMs), one for query rewriting and another for generating potential answers. By combining both, ConvGQR can produce better search queries. In addition, to relate query reformulation to the retrieval task, we propose a knowledge infusion mechanism to optimize both query reformulation and retrieval. Extensive experiments on four conversational search datasets demonstrate the effectiveness of ConvGQR.

Precise Zero-Shot Dense Retrieval without Relevance Labels
Luyu Gao, Xueguang Ma, Jimmy Lin and Jamie Callan   11:30-11:45 (Metropolitan West)
While dense retrieval has been shown to be effective and efficient across tasks and languages, it remains difficult to create effective fully zero-shot dense retrieval systems when no relevance labels are available. In this paper, we recognize the difficulty of zero-shot learning and encoding relevance. Instead, we propose to pivot through Hypothetical Document Embeddings (HyDE). Given a query, HyDE first zero-shot prompts an instruction-following language model (e.g., InstructGPT) to generate a hypothetical document. The document captures relevance patterns but is "fake" and may contain hallucinations. Then, an unsupervised contrastively learned encoder (e.g., Contriever) encodes the document into an embedding vector. This vector identifies a neighborhood in the corpus embedding space, from which similar real documents are retrieved based on vector similarity. This second step grounds the generated document to the actual corpus, with the encoder's dense bottleneck filtering out the hallucinations. Our experiments show that HyDE significantly outperforms the state-of-the-art unsupervised dense retriever Contriever and shows strong performance comparable to fine-tuned retrievers across various tasks (e.g. web search, QA, fact verification) and in non-English languages (e.g., sw, ko, ja, bn).

What Are You Token About? Dense Retrieval as Distributions Over the Vocabulary
Ori Ram, Liat Bezalel, Adi Zicher, Yonatan Belinkov, Jonathan Berant and Amir Globerson   11:45-12:00 (Metropolitan West)
Dual encoders are now the dominant architecture for dense retrieval. Yet, we have little understanding of how they represent text, and why this leads to good performance. In this work, we shed light on this question via distributions over the vocabulary. We propose to interpret the vector representations produced by dual encoders by projecting them into the model's vocabulary space. We show that the resulting projections contain rich semantic information, and draw connection between them and sparse retrieval. We find that this view can offer an explanation for some of the failure cases of dense retrievers. For example, we observe that the inability of models to handle tail entities is correlated with a tendency of the token distributions to forget some of the tokens of those entities. We leverage this insight and propose a simple way to enrich query and passage representations with lexical information at inference time, and show that this significantly improves performance compared to the original model in zero-shot settings, and specifically on the BEIR benchmark.

FAA: Fine-grained Attention Alignment for Cascade Document Ranking
Zhen Li, Chongyang Tao, Jiazhan Feng, Tao Shen, Dongyan Zhao, Xiubo Geng and Daxin Jiang   12:00-12:15 (Metropolitan West)
Document ranking aims at sorting a collection of documents with their relevance to a query. Contemporary methods explore more efficient transformers or divide long documents into passages to handle the long input. However, intensive query-irrelevant content may lead to harmful distraction and high query latency. Some recent works further propose cascade document ranking models that extract relevant passages with an efficient selector before ranking, however, their selection and ranking modules are almost independently optimized and deployed, leading to selecting error reinforcement and sub-optimal performance. In fact, the document ranker can provide fine-grained supervision to make the selector more generalizable and compatible, and the selector built upon a different structure can offer a distinct perspective to assist in document ranking. Inspired by this, we propose a fine-grained attention alignment approach to jointly optimize a cascade document ranking model. Specifically, we utilize the attention activations over the passages from the ranker as fine-grained attention feedback to optimize the selector. Meanwhile, we fuse the relevance scores from the passage selector into the ranker to assist in calculating the cooperative matching representation. Experiments on MS MARCO and TREC DL demonstrate the effectiveness of our method.

CITADEL: Conditional Token Interaction via Dynamic Lexical Routing for Efficient and Effective Multi-Vector Retrieval
Minghan Li, Sheng-Chieh Lin, Barlas Oguz, Asish Ghoshal, Jimmy Lin, Yashar Mehdad, Wen-tau Yih and Xilun Chen   12:15-12:30
(Metropolitan West)
Multi-vector retrieval methods combine the merits of sparse (e.g. BM25) and dense (e.g. DPR) retrievers and have achieved state-of-the-art performance on various retrieval tasks. These methods, however, are orders of magnitude slower and need much more space to store their indices compared to their single-vector counterparts. In this paper, we unify different multi-vector retrieval models from a token routing viewpoint and propose conditional token interaction via dynamic lexical routing, namely CITADEL, for efficient and effective multi-vector retrieval. CITADEL learns to route different token vectors to the predicted lexical keys such that a query token vector only interacts with document token vectors routed to the same key. This design significantly reduces the computation cost while maintaining high accuracy. Notably, CITADEL achieves the same or slightly better performance than the previous state of the art, ColBERT-v2, on both in-domain (MS MARCO) and out-of-domain (BEIR) evaluations, while being nearly 40 times faster. Source code and data are available at https://github.com/facebookresearch/dpr-scale/tree/citadel.

Posters

11:00-12:30 (Frontenac Ballroom and Queen's Quay)

[TACL] An Empirical Survey of Data Augmentation for Limited Data Learning in NLP
Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal and Diyi Yang   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
NLP has achieved great progress in the past decade through the use of neural models and large labeled datasets. The dependence on abundant data prevents NLP models from being applied to low-resource or novel tasks where significant time, money, or expertise is required to label massive amounts of textual data. Recently, data augmentation methods have been explored as a means of improving data efficiency in NLP. To date, there has been no systematic empirical overview of data augmentation for NLP in the limited labeled data setting, making it difficult to understand which methods work in which . In this paper, we provide an empirical survey of recent progress on data augmentation for NLP in the limited labeled data setting, summarizing the landscape of methods (including token-level augmentations, sentence-level augmentations, adversarial augmentations and hidden-space augmentations) and carrying out experiments on 11 datasets covering topics/news classification, inference tasks, paraphrasing tasks, and single-sentence tasks. Based on the results, we draw several conclusions to help practitioners choose appropriate augmentations in different and discuss the current challenges and future directions for limited data learning in NLP.

[TACL] Abstractive Meeting Summarization: A Survey
Virgile Rennard, Guokan Shang, Julie Hunter and Michalis Vazirgiannis   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Recent advances in deep learning, and especially the invention of encoder-decoder architectures, has significantly improved the performance of abstractive summarization. While the majority of research has focused on written documents, we have observed an increasing interest in the summarization of dialogues and multi-party conversation over the past few years. A system that could reliably transform the audio or transcript of a human conversation into an abridged version that homes in on the most important points of the discussion would be valuable in a wide variety of real-world contexts, from business meetings to medical consultations to customer service calls. This paper focuses on abstractive summarization for multi-party meetings, providing a survey of the challenges, datasets and systems relevant to this task and a discussion of promising directions for future study.

[TACL] Meta-Learning a Cross-lingual Manifold for Semantic Parsing
Tom Sherborne and Mirella Lapata   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Localizing a semantic parser to support new languages requires effective cross-lingual generalization. Recent work has found success with machine-translation or zero-shot methods although these approaches can struggle to model how native speakers ask questions. We consider how to effectively leverage minimal annotated examples in new languages for few-shot cross-lingual semantic parsing. We introduce a first-order meta-learning algorithm to train a semantic parser with maximal sample efficiency during cross-lingual transfer. Our algorithm uses high-resource languages to train the parser and simultaneously optimizes for cross-lingual generalization for lower-resource languages. Results across six languages on ATIS demonstrate that our combination of generalization steps yields accurate semantic parsers sampling 10

[SRW] How do different tokenizers perform on downstream tasks in scriptio continua languages?: A case study in Japanese
Takuro Fujii, Koki Shibata, Atsuki Yamaguchi, Terufumi Morishita and Yasuhiro Sogawa   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
We investigate the impact of different tokenizers on downstream performance in Japanese NLP, with the case of BERT architecture.

[SRW] Kanbun-LM: Reading and Translating Classical Chinese in Japanese Methods by Language Models
Hao Wang, Hirofumi Shimizu and Daisuke Kawahara   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Reading and Translating Classical Chinese in Japanese Methods by Language Models

[SRW] Aligning Code-Switching Metrics with Bilingual Behavior
Rebecca Pattichis, Sonya Trawick, Dora Lacasse and Rena Torres Cacoullos   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
While models of bilingual code-switching have utilized word-level tokens, this work advocates for the Intonation Unit, a multi-word prosodic unit, by demonstrating how metrics of code-switching complexity are impacted by the distinction between other-language single-word items and multi-word strings.

[SRW] Enhancing Ancient Chinese Understanding with Derived Noisy Syntax Trees
Ping Wang, Shitou Zhang, Zuchao Li and Jingrui Hou   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
This paper introduces a confidence-based syntax encoding network (cSEN) to incorporate syntax in ancient Chinese understanding tasks, effectively improving performance by mitigating noise and incompatibility issues.

[SRW] Theoretical Linguistics Rivals Embeddings in Language Clustering for Multilingual Named Entity Recognition
Sakura Imai, Daisuke Kawahara, Naho Orita and Hiromune Oda   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
This study investigates whether and how theoretical linguistics improves language clustering for multilingual named entity recognition (NER), with the two types of language groupings proposed: one based on morpho-syntactic features in a nominal domain and one based on a head parameter.

[SRW] EvoGrad: An Online Platform for an Evolving Winograd Schema Challenge using Adversarial Human Perturbations
Jing Han Sun, Jad Kabbara and Ali Emami   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

An open-source, user-friendly platform for the continual evaluation and development of models for the Winograd Schema Challenge, based on iterations of human-adversarial perturbations.

[SRW] SWEET: Weakly Supervised Person Name Extraction for Fighting Human Trafficking
Javin Liu, Peter Yu, Vidya Sujaya, Pratheeksha Nair, Kellin Pelrine and Reihaneh Rabbany   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Weak supervision for entity extraction from noisy text

[Demo] Finspector: A Human-Centered Visual Inspection Tool for Exploring and Comparing Biases among Foundation Models
Nandana Mihindukulasooriya and Bum Chul Kwon   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Pre-trained transformer-based language models are becoming increasingly popular due to their exceptional performance on various benchmarks. However, concerns persist regarding the presence of hidden biases within these models, which can lead to discriminatory outcomes and reinforce harmful stereotypes. To address this issue, we propose Finspector, a human-centered visual inspection tool designed to detect biases in different categories through log-likelihood scores generated by language models. The goal of the tool is to enable researchers to easily identify potential biases using visual analytics, ultimately contributing to a fairer and more just deployment of these models in both academic and industrial settings. Finspector is available at https://github.com/IBM/finspector.

[Demo] PEEP-Talk: A Situational Dialogue-based Chatbot for English Education
Heuiseok Lim, Bernardo Yahya, Seounghoon Lee, Sugyeong Eo, Hyeonseok Moon, Jaehyung Seo, Jungseob Lee, Chanjun Park, Yoonna Jang and Seungjun Lee   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
English is acknowledged worldwide as a mode of communication. However, due to the absence of realistic practicing scenarios, students learning English as a foreign language (EFL) typically have limited chances to converse and share feedback with others. In this paper, we propose PEEP-Talk, a real-world situational dialogue-based chatbot designed for English education. It also naturally switches to a new topic or situation in response to out-of-topic utterances, which are common among English beginners. Furthermore, PEEP-Talk provides feedback score on conversation and grammar error correction. We performed automatic and user evaluations to validate performance and education efficiency of our system. The results show that PEEP-Talk generates appropriate responses in various real-life situations while providing accurate feedback to learners. Moreover, we demonstrate a positive impact on English-speaking, grammar, and English learning anxiety, implying that PEEP-Talk can lower the barrier to learning natural conversation in effective ways.

[Demo] Ranger: A Toolkit for Effect-Size Based Multi-Task Evaluation
Sebastian Hofstätter, Sophia Althammer and Mete Sertkan   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
In this paper, we introduce Ranger - a toolkit to facilitate the easy use of effect-size-based meta-analysis for multi-task evaluation in NLP and IR. We observed that our communities often face the challenge of aggregating results over incomparable metrics and scenarios, which makes conclusions and take-away messages less reliable. With Ranger, we aim to address this issue by providing a task-agnostic toolkit that combines the effect of a treatment on multiple tasks into one statistical evaluation, allowing for comparison of metrics and computation of an overall summary effect. Our toolkit produces publication-ready forest plots that enable clear communication of evaluation results over multiple tasks. Our goal with the ready-to-use Ranger toolkit is to promote robust, effect-size-based evaluation and improve evaluation standards in the community. We provide two case studies for common IR and NLP settings to highlight Ranger's benefits.

FIREBALL: A Dataset of Dungeons and Dragons Actual-Play with Structured Game State Information
Andrew Zhu, Karmanya Aggarwal, Alexander H Feng, Lara J. Martin and Chris Callison-Burch   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Dungeons & Dragons (D&D) is a tabletop roleplaying game with complex natural language interactions between players and hidden state information. Recent work has shown that large language models (LLMs) that have access to state information can generate higher quality game turns than LLMs that use dialog history alone. However, previous work used game state information that was heuristically created and was not a true gold standard game state. We present FIREBALL, a large dataset containing nearly 25,000 unique sessions from real D&D gameplay on Discord with true game state info. We recorded game play sessions of players who used the Avrae bot, which was developed to aid people in playing D&D online, capturing language, game commands and underlying game state information. We demonstrate that FIREBALL can improve natural language generation (NLG) by using Avrae state information, improving both automated metrics and human judgments of quality. Additionally, we show that LLMs can generate executable Avrae commands, particularly after finetuning.

Movie101: A New Movie Understanding Benchmark
Zihao Yue, Qi Zhang, Anwen Hu, Liang Zhang, Ziheng Wang and Qin Jin   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
To help the visually impaired enjoy movies, automatic movie narrating systems are expected to narrate accurate, coherent, and role-aware plots when there are no speaking lines of actors. Existing works benchmark this challenge as a normal video captioning task using some simplifications, such as removing role names and evaluating narrations with ngram-based metrics, which makes it difficult for automatic systems to meet the needs of real application scenarios. To narrow this gap, we construct a large-scale Chinese movie benchmark, named Movie101. Closer to real scenarios, the Movie Clip Narrating (MCN) task in our benchmark asks models to generate role-aware narration paragraphs for complete movie clips where no actors are speaking. External knowledge, such as role information and movie genres, is also provided for better movie understanding. Besides, we propose a new metric called Movie Narration Score (MNScore) for movie narrating evaluation, which achieves the best correlation with human evaluation. Our benchmark also supports the Temporal Narration Grounding (TNG) task to investigate clip localization given text descriptions. For both two tasks, our proposed methods well leverage external knowledge and outperform carefully designed baselines. The dataset and codes are released at https://github.com/yuezih/Movie101.

FactKG: Fact Verification via Reasoning on Knowledge Graphs
Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne and Edward Choi   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
In real world applications, knowledge graphs (KG) are widely used in various domains (e.g. medical applications and dialogue agents). However, for fact verification, KGs have not been adequately utilized as a knowledge source. KGs can be a valuable knowledge source in fact verification due to their reliability and broad applicability. A KG consists of nodes and edges which makes it clear how concepts are linked together, allowing machines to reason over chains of topics. However, there are many challenges in understanding how these machine-readable concepts map to information in text. To enable the community to better use KGs, we introduce a new dataset, FactKG: Fact Verification via Reasoning on Knowledge Graphs. It consists of 108k natural language claims with five types of reasoning: One-hop, Conjunction, Existence, Multi-hop, and Negation. Furthermore, FactKG contains various linguistic patterns, including colloquial style claims as well as written style claims to increase practicality. Lastly, we develop a baseline approach and analyze FactKG over these reasoning types. We believe FactKG can advance both reliability and practicality in KG-based fact verification.

IndicMT Eval: A Dataset to Meta-Evaluate Machine Translation Metrics for Indian Languages
Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra and Raj Dabre   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

The rapid growth of machine translation (MT) systems necessitates meta-evaluations of evaluation metrics to enable selection of those that best reflect MT quality. Unfortunately, most meta-evaluation studies focus on European languages, the observations for which may not always apply to other languages. Indian languages, having over a billion speakers, are linguistically different from them, and to date, there are no such systematic studies focused solely on English to Indian language MT. This paper fills this gap through a Multidimensional Quality Metric (MQM) dataset consisting of 7000 fine-grained annotations, spanning 5 Indian languages and 7 MT systems. We evaluate 16 metrics and show that, pre-trained metrics like COMET have the highest correlations with annotator scores as opposed to n-gram metrics like BLEU. We further leverage our MQM annotations to develop an Indic-COMET metric and show that it outperforms COMET counterparts in both human scores correlations and robustness scores in Indian languages. Additionally, we show that the Indic-COMET can outperform COMET on some unseen Indian languages. We hope that our dataset and analysis will facilitate further research in Indic MT evaluation.

STT4SG-350: A Speech Corpus for All Swiss German Dialect Regions

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel and Mark Cieliebak   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
We present STT4SG-350, a corpus of Swiss German speech, annotated with Standard German text at the sentence level. The data is collected using a web app in which the speakers are shown Standard German sentences, which they translate to Swiss German and record. We make the corpus publicly available. It contains 343 hours of speech from all dialect regions and is the largest public speech corpus for Swiss German to date. Application areas include automatic speech recognition (ASR), text-to-speech, dialect identification, and speaker recognition. Dialect information, age group, and gender of the 316 speakers are provided. Genders are equally represented and the corpus includes speakers of all ages. Roughly the same amount of speech is provided per dialect region, which makes the corpus ideally suited for experiments with speech technology for different dialects. We provide training, validation, and test splits of the data. The test set consists of the same spoken sentences for each dialect region and allows a fair evaluation of the quality of speech technologies in different dialects. We train an ASR model on the training set and achieve an average BLEU score of 74.7 on the test set. The model beats the best published BLEU scores on 2 other Swiss German ASR test sets, demonstrating the quality of the corpus.

Effective Contrastive Weighting for Dense Query Expansion

Xiao Wang, Sean MacAvaney, Craig Macdonald and Iadh Ounis   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Verbatim queries submitted to search engines often do not sufficiently describe the user's search intent. Pseudo-relevance feedback (PRF) techniques, which modify a query's representation using the top-ranked documents, have been shown to overcome such inadequacies and improve retrieval effectiveness for both lexical methods (e.g., BM25) and dense methods (e.g., ANCE, ColBERT). For instance, the recent ColBERT-PRF approach heuristically chooses new embeddings to add to the query representation using the inverse document frequency (IDF) of the underlying tokens. However, this heuristic potentially ignores the valuable context encoded by the embeddings. In this work, we present a contrastive solution that learns to select the most useful embeddings for expansion. More specifically, a deep language model-based contrastive weighting model, called CWPRF, is trained to learn to discriminate between relevant and non-relevant documents for semantic search. Our experimental results show that our contrastive weighting model can aid to select useful expansion embeddings and outperform various baselines. In particular, CWPRF can improve nDCG@10 by upto to 4.1% compared to an existing PRF approach for ColBERT while maintaining its efficiency.

Pivotal Role of Language Modeling in Recommender Systems: Enriching Task-specific and Task-agnostic Representation Learning

Kyuyong Shin, Hanock Kwak, Wonjae Kim, Jisu Jeong, Seungjae Jung, Kyumgin Kim, Jung-Woo Ha and Sang-Woo Lee   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Recent studies have proposed unified user modeling frameworks that leverage user behavior data from various applications. Many of them benefit from utilizing users' behavior sequences as plain texts, representing rich information in any domain or system without losing generality. Hence, a question arises: Can language modeling for user history corpus help improve recommender systems? While its versatile usability has been widely investigated in many domains, its applications to recommender systems still remain underexplored. We show that language modeling applied directly to task-specific user histories achieves excellent results on diverse recommendation tasks. Also, leveraging additional task-agnostic user histories delivers significant performance benefits. We further demonstrate that our approach can provide promising transfer learning capabilities for a broad spectrum of real-world recommender systems, even on unseen domains and services.

Detecting Contradictory COVID-19 Drug Efficacy Claims from Biomedical Literature

Daniel N Sosa, Malavika Suresh, Christopher Potts and Russ B Altman   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The COVID-19 pandemic created a deluge of questionable and contradictory scientific claims about drug efficacy – an "infodemic" with lasting consequences for science and society. In this work, we argue that NLP models can help domain experts distill and understand the literature in this complex, high-stakes area. Our task is to automatically identify contradictory claims about COVID-19 drug efficacy. We frame this as a natural language inference problem and offer a new NLI dataset created by domain experts. The NLI framing allows us to create curricula combining existing datasets and our own. The resulting models are useful investigative tools. We provide a case study of how these models help a domain expert summarize and assess evidence concerning remdisivir and hydroxychloroquine.

Shrinking Embeddings for Hyper-Relational Knowledge Graphs

Bo Xiong, Mojtaba Nayyeri, Shirui Pan and Steffen Staab   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Link prediction on knowledge graphs (KGs) has been extensively studied on binary relational KGs, wherein each fact is represented by a triple. A significant amount of important knowledge, however, is represented by hyper-relational facts where each fact is composed of a primal triple and a set of qualifiers comprising a key-value pair that allows for expressing more complicated semantics. Although some recent works have proposed to embed hyper-relational KGs, these methods fail to capture essential inference patterns of hyper-relational facts such as qualifier monotonicity, qualifier implication, and qualifier mutual exclusion, limiting their generalization capability. To unlock this, we present ShrinkE, a geometric hyper-relational KG embedding method aiming to explicitly model these patterns. ShrinkE models the primal triple as a spatial-functional transformation from the head into a relation-specific box. Each qualifier "shrinks" the box to narrow down the possible answer set and, thus, realizes qualifier monotonicity. The spatial relationships between the qualifier boxes allow for modeling core inference patterns of qualifiers such as implication and mutual exclusion. Experimental results demonstrate ShrinkE's superiority on three benchmarks of hyper-relational KGs.

Efficient Diagnosis Assignment Using Unstructured Clinical Notes

Louis Blankemeier, Jason Fries, Robert Tinn, Joseph S Preston, Nigam Shah and Akshay Chaudhari   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Electronic phenotyping entails using electronic health records (EHRs) to identify patients with specific health outcomes and determine when those outcomes occurred. Unstructured clinical notes, which contain a vast amount of information, are a valuable resource for electronic phenotyping. However, traditional methods, such as rule-based labeling functions or neural networks, require significant manual effort to tune and may not generalize well to multiple indications. To address these challenges, we propose *HyDE* (hybrid diagnosis extractor). HyDE is a simple framework for electronic phenotyping that integrates labeling functions and a disease-agnostic neural network to assign diagnoses to

patients. By training HyDE's model to correct predictions made by labeling functions, we are able to disambiguate hypertension true positives and false positives with a supervised area under the precision-recall curve (AUPRC) of 0.85. We extend this hypertension-trained model to zero-shot evaluation of four other diseases, generating AUPRC values ranging from 0.82 - 0.95 and outperforming a labeling function baseline by 44 points in F1 score and a Word2Vec baseline by 24 points in F1 score on average. Furthermore, we demonstrate a speedup of >4x by pruning the length of inputs into our language model to 2.3% of the full clinical notes, with negligible impact to the AUPRC. HyDE has the potential to improve the efficiency and efficacy of interpreting large-scale unstructured clinical notes for accurate EHR phenotyping.

A Compare-and-contrast Multistage Pipeline for Uncovering Financial Signals in Financial Reports
Jia-Huei Ju, Yu-Shiang Huang, Cheng-Wei Lin, Che Lin and Chuan-Ju Wang   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
In this paper, we address the challenge of discovering financial signals in narrative financial reports. As these documents are often lengthy and tend to blend routine information with new information, it is challenging for professionals to discern critical financial signals. To this end, we leverage the inherent nature of the year-to-year structure of reports to define a novel signal-highlighting task; more importantly, we propose a compare-and-contrast multistage pipeline that recognizes different relationships between the reports and locates relevant rationales for these relationships. We also create and publicly release a human-annotated dataset for our task. Our experiments on the dataset validate the effectiveness of our pipeline, and we provide detailed analyses and ablation studies to support our findings.

DT-Solver: Automated Theorem Proving with Dynamic-Tree Sampling Guided by Proof-level Value Function
Haiming Wang, Ye Yuan, Zhengying Liu, Jianhao Shen, Yichun Yin, Jing Xiong, Enze Xie, Han Shi, Yujun Li, lin li, Jian Yin, Zhenguo Li and Xiaodan Liang   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Recent advances in neural theorem-proving resort to large language models and tree searches. When proving a theorem, a language model advises single-step actions based on the current proving state and the tree search finds a sequence of correct steps using actions given by the language model. However, prior works often conduct constant computation efforts for each proving state while ignoring that the hard states often need more exploration than easy states. Moreover, they evaluate and guide the proof search solely depending on the current proof state instead of considering the whole proof trajectory as human reasoning does. Here, to accommodate general theorems, we propose a novel Dynamic-Tree Driven Theorem Solver (DT-Solver) by guiding the search procedure with state confidence and proof-level values. Specifically, DT-Solver introduces a dynamic-tree Monte-Carlo search algorithm, which dynamically allocates computing budgets for different state confidences, guided by a new proof-level value function to discover proof states that require substantial exploration. Experiments on two popular theorem-proving datasets, PISA and Mathlib, show significant performance gains by our DT-Solver over the state-of-the-art approaches, with a 6.65% improvement on average in terms of success rate. And especially under low computing resource settings (11.03% improvement on average).

U-CREAT: Unsupervised Case Retrieval using Events extrAcTion
Abhinav Joshi, Akshat Sharma, Sai Kiran Tanikella and Ashutosh Modi   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The task of Prior Case Retrieval (PCR) in the legal domain is about automatically citing relevant (based on facts and precedence) prior legal cases in a given query case. To further promote research in PCR, in this paper, we propose a new large benchmark (in English) for the PCR task: IL-PCR (Indian Legal Prior Case Retrieval) corpus. Given the complex nature of case relevance and the long size of legal documents, BM25 remains a strong baseline for ranking the cited prior documents. In this work, we explore the role of events in legal case retrieval and propose an unsupervised retrieval method-based pipeline U-CREAT (Unsupervised Case Retrieval using Events Extraction). We find that the proposed unsupervised retrieval method significantly increases performance compared to BM25 and makes retrieval faster by a considerable margin, making it applicable to real-time case retrieval systems. Our proposed system is generic, we show that it generalizes across two different legal systems (Indian and Canadian), and it shows state-of-the-art performance on the benchmarks for both the legal systems (IL-PCR and COLIEE corpora).

Natural Language to Code Generation in Interactive Data Science Notebooks
Pengcheng Yin, Wen-Ding Li, Kefan Xiao, Abhishek K Rao, Yeming Wen, Kensen Shi, Joshua Howland, Paige Bailey, Michele Catasta, Henryk Michalewski, Oleksandr Polozov and Charles Sutton   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Computational notebooks, such as Jupyter notebooks, are interactive computing environments that are ubiquitous among data scientists to perform data wrangling and analytic tasks. To measure the performance of AI pair programmers that automatically synthesize programs for those tasks given natural language (NL) intents from users, we build ARCADE, a benchmark of 1078 code generation problems using the pandas data analysis framework in data science notebooks. ARCADE features multiple rounds of NL-to-code problems from the same notebook. It requires a model to understand rich multi-modal contexts, such as existing notebook cells and their execution states as well as previous turns of interaction. To establish a strong baseline on this challenging task, we develop PaChiNCo, a 62B code language model (LM) for Python computational notebooks, which significantly outperforms public code LMs. Finally, we explore few-shot prompting strategies to elicit better code with step-by-step decomposition and NL explanation, showing the potential to improve the diversity and explainability of model predictions. Arcade is publicly available at https://github.com/google-research/arcade-nl2code/.

Synthetic Text Generation with Differential Privacy: A Simple and Practical Recipe
Xiang Yue, Huseyin A Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan and Robert Sim 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Privacy concerns have attracted increasing attention in data-driven products due to the tendency of machine learning models to memorize sensitive training data. Generating synthetic versions of such data with a formal privacy guarantee, such as differential privacy (DP), provides a promising path to mitigating these privacy concerns, but previous approaches in this direction have typically failed to produce synthetic data of high quality. In this work, we show that a simple and practical recipe in the text domain is effective: simply fine-tuning a pretrained generative language model with DP enables the model to generate useful synthetic text with strong privacy protection. Through extensive empirical analyses on both benchmark and private customer data, we demonstrate that our method produces synthetic text that is competitive in terms of utility with its non-private counterpart, meanwhile providing strong protection against potential privacy leakages.

Exploring Continual Learning for Code Generation Models
Prateek Yadav, Qing Sun, Hantian Ding, Xiaopeng Li, Dejiao Zhang, Ming Tan, Parminder Bhatia, Xiaofei Ma, Ramesh Nallapati, Murali Krishna Ramanathan, Mohit Bansal and Bing Xiang   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Large-scale code generation models such as Copilot and CodeT5 have achieved impressive performance. However, libraries are upgraded or deprecated very frequently and re-training large-scale language models is computationally expensive. Therefore, Continual Learning (CL) is an important aspect that remains under-explored in the code domain. In this paper, we introduce a benchmark called CodeTask-CL that covers a wide range of tasks, including code generation, translation, summarization, and refinement, with different input and output programming languages. Next, on our CodeTask-CL benchmark, we compare popular CL techniques from NLP and Vision domains. We find that effective methods like Prompt Pooling (PP) suffer from catastrophic forgetting due to the unstable training of the prompt selection mechanism caused by stark distribution shifts in coding tasks. We address this issue with our proposed method, Prompt Pooling with Teacher Forcing (PP-TF), that stabilizes training by enforcing constraints on the prompt selection mechanism and leads to a 21.54% improvement over Prompt Pooling. Along with the benchmark, we establish a training pipeline that can be used for CL on code models, which we believe can motivate further

development of CL methods for code models.

VendorLink: An NLP approach for Identifying & Linking Vendor Migrants & Potential Aliases on Darknet Markets
Vageesh Kumar Saxena, Nils Rethmeier, Gijs van Dijck and Gerasimos Spanakis   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The anonymity on the Darknet allows vendors to stay undetected by using multiple vendor aliases or frequently migrating between markets. Consequently, illegal markets and their connections are challenging to uncover on the Darknet. To identify relationships between illegal markets and their vendors, we propose VendorLink, an NLP-based approach that examines writing patterns to verify, identify, and link unique vendor accounts across text advertisements (ads) on seven public Darknet markets. In contrast to existing literature, VendorLink utilizes the strength of supervised pre-training to perform closed-set vendor verification, open-set vendor identification, and low-resource market adaption tasks. Through VendorLink, we uncover (i) 15 migrants and 71 potential aliases in the Alphabay-Dreams-Silk dataset, (ii) 17 migrants and 3 potential aliases in the Valhalla-Berlusconi dataset, and (iii) 75 migrants and 10 potential aliases in the Traderoute-Agora dataset. Altogether, our approach can help Law Enforcement Agencies (LEA) make more informed decisions by verifying and identifying migrating vendors and their potential aliases on existing and Low-Resource (LR) emerging Darknet markets.

Compounding Geometric Operations for Knowledge Graph Completion
Xiou Ge, Yun Cheng Wang, Bin Wang and C.-C. Jay Kuo   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Geometric transformations including translation, rotation, and scaling are commonly used operations in image processing. Besides, some of them are successfully used in developing effective knowledge graph embedding (KGE). Inspired by the synergy, we propose a new KGE model by leveraging all three operations in this work. Since translation, rotation, and scaling operations are cascaded to form a composite one, the new model is named CompoundE. By casting CompoundE in the framework of group theory, we show that quite a few distanced-based KGE models are special cases of CompoundE. CompoundE extends the simple distance-based scoring functions to relation-dependent compound operations on head and/or tail entities. To demonstrate the effectiveness of CompoundE, we perform three prevalent KG prediction tasks including link prediction, path query answering, and entity typing, on a range of datasets. CompoundE outperforms extant models consistently, demonstrating its effectiveness and flexibility.

Learning Multi-Step Reasoning by Solving Arithmetic Tasks
Tianduo Wang and Wei Lu   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Mathematical reasoning is regarded as a necessary ability for Language Models (LMs). Recent works demonstrate large LMs' impressive performance in solving math problems. The success is attributed to their Chain-of-Thought (CoT) reasoning abilities, i.e., the ability to decompose complex questions into step-by-step reasoning chains, but such ability seems only to emerge from models with abundant parameters. This work investigates how to incorporate relatively small LMs with the capabilities of multi-step reasoning. We propose to inject such abilities by continually pre-training LMs on a synthetic dataset MsAT which is composed of Multi-step Arithmetic Tasks. Our experiments on four math word problem datasets show the effectiveness of the proposed method in enhancing LMs' math reasoning abilities.

Robust Multi-bit Natural Language Watermarking through Invariant Features
KiYoon Yoo, Wonhyuk Ahn, Jiho Jang and Nojun Kwak   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Recent years have witnessed a proliferation of valuable original natural language contents found in subscription-based media outlets, web novel platforms, and outputs of large language models. However, these contents are susceptible to illegal piracy and potential misuse without proper security measures. This calls for a secure watermarking system to guarantee copyright protection through leakage tracing or ownership identification. To effectively combat piracy and protect copyrights, a multi-bit watermarking framework should be able to embed adequate bits of information and extract the watermarks in a robust manner despite possible corruption. In this work, we explore ways to advance both payload and robustness by following a well-known proposition from image watermarking and identify features in natural language that are invariant to minor corruption. Through a systematic analysis of the possible sources of errors, we further propose a corruption-resistant infill model. Our full method improves upon the previous work on robustness by +16.8% point on average on four datasets, three corruption types, and two corruption ratios

MoralDial: A Framework to Train and Evaluate Moral Dialogue Systems via Moral Discussions
Hao Sun, Zhexin Zhang, Fei Mi, Yasheng Wang, Wei Liu, Jianwei Cui, Bin Wang, Qun Liu and Minlie Huang   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Morality in dialogue systems has raised great attention in research recently. A moral dialogue system aligned with users' values could enhance conversation engagement and user connections. In this paper, we propose a framework, MoralDial to train and evaluate moral dialogue systems. In our framework, we first explore the communication mechanisms of morality and resolve expressed morality into three parts, which indicate the roadmap for building a moral dialogue system. Based on that, we design a simple yet effective method: constructing moral discussions between simulated specific users and the dialogue system. The constructed discussions consist of expressing, explaining, revising, and inferring moral views in dialogue exchanges, which makes conversational models learn morality well in a natural manner. Furthermore, we propose a novel evaluation method under the framework. We evaluate the multiple aspects of morality by judging the relation between dialogue responses and human values in discussions, where the multifaceted nature of morality is particularly considered. Automatic and manual experiments demonstrate that our framework is promising to train and evaluate moral dialogue systems.

MidMed: Towards Mixed-Type Dialogues for Medical Consultation
Xiaoming Shi, Zeming Liu, Chuan Wang, Haitao Leng, Kui Xue, Xiaofan Zhang and Shaoting Zhang   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Most medical dialogue systems assume that patients have clear goals (seeking a diagnosis, medicine querying, etc.) before medical consultation. However, in many real situations, due to the lack of medical knowledge, it is usually difficult for patients to determine clear goals with all necessary slots. In this paper, we identify this challenge as how to construct medical consultation dialogue systems to help patients clarify their goals. For further study, we create a novel human-to-human mixed-type medical consultation dialogue corpus, termed MidMed, covering four dialogue types: task-oriented dialogue for diagnosis, recommendation, QA, and chitchat. MidMed covers four departments (otorhinolaryngology, ophthalmology, skin, and digestive system), with 8,309 dialogues. Furthermore, we build benchmarking baselines on MidMed and propose an instruction-guiding medical dialogue generation framework, termed InsMed, to handle mixed-type dialogues. Experimental results show the effectiveness of InsMed.

Laziness Is a Virtue When It Comes to Compositionality in Neural Semantic Parsing
Maxwell Crouse, Pavan Kapanipathi, Subhajit Chaudhury, Tahira Naseem, Ramon Fernandez Astudillo, Achille Fokoue and Tim Klinger
11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Nearly all general-purpose neural semantic parsers generate logical forms in a strictly top-down autoregressive fashion. Though such systems have achieved impressive results across a variety of datasets and domains, recent works have called into question whether they are ultimately limited in their ability to compositionally generalize. In this work, we approach semantic parsing from, quite literally, the opposite direction; that is, we introduce a neural semantic parsing generation method that constructs logical forms from the bottom up, beginning from the logical form's leaves. The system we introduce is lazy in that it incrementally builds up a set of potential semantic parses, but only expands and pro-

cesses the most promising candidate parses at each generation step. Such a parsimonious expansion scheme allows the system to maintain an arbitrarily large set of parse hypotheses that are never realized and thus incur minimal computational overhead. We evaluate our approach on compositional generalization; specifically, on the challenging CFQ dataset and two other Text-to-SQL datasets where we show that our novel, bottom-up semantic parsing technique outperforms general-purpose semantic parsers while also being competitive with semantic parsers that have been tailored to each task.

Learning Answer Generation using Supervision from Automatic Question Answering Evaluators
Matteo Gabburo, Siddhant Garg, Rik Koncel-Kedziorski and Alessandro Moschitti   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Recent studies show that sentence-level extractive QA, i.e., based on Answer Sentence Selection (AS2), is outperformed by Generation-based QA (GenQA) models, which generate answers using the top-k answer sentences ranked by AS2 models (a la retrieval-augmented generation style). In this paper, we propose a novel training paradigm for GenQA using supervision from automatic QA evaluation models (GAVA). Specifically, we propose three strategies to transfer knowledge from these QA evaluation models to a GenQA model: (i) augmenting training data with answers generated by the GenQA model and labelled by GAVA (either statically, before training, or (ii) dynamically, at every training epoch); and (iii) using the GAVA score for weighting the generator loss during the learning of the GenQA model. We evaluate our proposed methods on two academic and one industrial dataset, obtaining a significant improvement in answering accuracy over the previous state of the art.

An Inner Table Retriever for Robust Table Question Answering
Weizhe Lin, Rexhina Bloshmi, Bill Byrne, Adria de Gispert and Gonzalo Iglesias   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Recent years have witnessed the thriving of pretrained Transformer-based language models for understanding semi-structured tables, with several applications, such as Table Question Answering (TableQA). These models are typically trained on joint tables and surrounding natural language text, by linearizing table content into sequences comprising special tokens and cell information. This yields very long sequences which increase system inefficiency, and moreover, simply truncating long sequences results in information loss for downstream tasks. We propose Inner Table Retriever (ITR), a general-purpose approach for handling long tables in TableQA that extracts sub-tables to preserve the most relevant information for a question. We show that ITR can be easily integrated into existing systems to improve their accuracy with up to 1.3-4.8% and achieve state-of-the-art results in two benchmarks, i.e., 63.4% in WikiTableQuestions and 92.1% in WikiSQL. Additionally, we show that ITR makes TableQA systems more robust to reduced model capacity and to different ordering of columns and rows. We make our code available at: https://github.com/amazon-science/robust-tableqa.

Minding Language Models' (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker
Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi and Yulia Tsvetkov   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Theory of Mind (ToM)—the ability to reason about the mental states of other people—is a key element of our social intelligence. Yet, despite their ever more impressive performance, large-scale neural language models still lack basic theory of mind capabilities out-of-the-box. We posit that simply scaling up models will not imbue them with theory of mind due to the inherently symbolic and implicit nature of the phenomenon, and instead investigate an alternative: can we design a decoding-time algorithm that enhances theory of mind of off-the-shelf neural language models without explicit supervision? We present SymbolicToM, a plug-and-play approach to reason about the belief states of multiple characters in reading comprehension tasks via explicit symbolic representation. More concretely, our approach tracks each entity's beliefs, their estimation of other entities' beliefs, and higher-order levels of reasoning, all through graphical representations, allowing for more precise and interpretable reasoning than previous approaches. Empirical results on the well-known ToMi benchmark (Le et al., 2019) demonstrate that SymbolicToM dramatically enhances off-the-shelf neural networks' theory of mind in a zero-shot setting while showing robust out-of-distribution performance compared to supervised baselines. Our work also reveals spurious patterns in existing theory of mind benchmarks, emphasizing the importance of out-of-distribution evaluation and methods that do not overfit a particular dataset.

From Key Points to Key Point Hierarchy: Structured and Expressive Opinion Summarization
Arie Cattan, Lilach Eden, Yoav Kantor and Roy Bar-Haim   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Key Point Analysis (KPA) has been recently proposed for deriving fine-grained insights from collections of textual comments. KPA extracts the main points in the data as a list of concise sentences or phrases, termed Key Points, and quantifies their prevalence. While key points are more expressive than word clouds and key phrases, making sense of a long, flat list of key points, which often express related ideas in varying levels of granularity, may still be challenging. To address this limitation of KPA, we introduce the task of organizing a given set of key points into a hierarchy, according to their specificity. Such hierarchies may be viewed as a novel type of Textual Entailment Graph. We develop ThinkP, a high quality benchmark dataset of key point hierarchies for business and product reviews, obtained by consolidating multiple annotations. We compare different methods for predicting pairwise relations between key points, and for inferring a hierarchy from these pairwise predictions. In particular, for the task of computing pairwise key point relations, we achieve significant gains over existing strong baselines by applying directional distributional similarity methods to a novel distributional representation of key points, and further boost performance via weak supervision.

ArgU: A Controllable Factual Argument Generator
Sougata Saha and Rohini K Srihari   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Effective argumentation is essential towards a purposeful conversation with a satisfactory outcome. For example, persuading someone to reconsider smoking might involve empathetic, well founded arguments based on facts and expert opinions about its ill-effects and the consequences on one's family. However, the automatic generation of high-quality factual arguments can be challenging. Addressing existing controllability issues can make the recent advances in computational models for argument generation a potential solution. In this paper, we introduce ArgU: a neural argument generator capable of producing factual arguments from input facts and real-world concepts that can be explicitly controlled for stance and argument structure using Walton's argument scheme-based control codes. Unfortunately, computational argument generation is a relatively new field and lacks datasets conducive to training. Hence, we have compiled and released an annotated corpora of 69,428 arguments spanning six topics and six argument schemes, making it the largest publicly available corpus for identifying argument schemes; the paper details our annotation and dataset creation framework. We further experiment with an argument generation strategy that establishes an inference strategy by generating an "argument template" before actual argument generation. Our results demonstrate that it is possible to automatically generate diverse arguments exhibiting different inference patterns for the same set of facts by using control codes based on argument schemes and stance.

Language of Bargaining
Mourad Heddaya, Solomon E Dworkin, Chenhao Tan, Rob Voigt and Alexander K Zentefis   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Leveraging an established exercise in negotiation education, we build a novel dataset for studying how the use of language shapes bilateral bargaining. Our dataset extends existing work in two ways: 1) we recruit participants via behavioral labs instead of crowdsourcing platforms and allow participants to negotiate through audio, enabling more naturalistic interactions; 2) we add a control setting where participants negotiate only through alternating, written numeric offers. Despite the two contrasting forms of communication, we find that the average agreed prices of the two treatments are identical. But when subjects can talk, fewer offers are exchanged, negotiations finish faster, the likelihood of

reaching agreement rises, and the variance of prices at which subjects agree drops substantially. We further propose a taxonomy of speech acts in negotiation and enrich the dataset with annotated speech acts. We set up prediction tasks to predict negotiation success and find that being reactive to the arguments of the other party is advantageous over driving the negotiation.

What does a Text Classifier Learn about Morality? An Explainable Method for Cross-Domain Comparison of Moral Rhetoric
Enrico Liscio, Oscar Araque, Lorenzo Gatti, Ionut L. Constantinescu, Catholijn M Jonker, Kyriaki Kalimeri and Pradeep Kumar Murukannaiah   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Moral rhetoric influences our judgement. Although social scientists recognize moral expression as domain specific, there are no systematic methods for analyzing whether a text classifier learns the domain-specific expression of moral language or not. We propose Tomea, a method to compare a supervised classifier's representation of moral rhetoric across domains. Tomea enables quantitative and qualitative comparisons of moral rhetoric via an interpretable exploration of similarities and differences across moral concepts and domains. We apply Tomea on moral narratives in thirty-five thousand tweets from seven domains. We extensively evaluate the method via a crowd study, a series of cross-domain moral classification comparisons, and a qualitative analysis of cross-domain moral expression.

Detoxifying Text with MaRCo: Controllable Revision with Experts and Anti-Experts
Skyler R Hallinan, Alisa Liu, Yejin Choi and Maarten Sap   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Text detoxification has the potential to mitigate the harms of toxicity by rephrasing text to remove offensive meaning, but subtle toxicity remains challenging to tackle. We introduce MaRCo, a detoxification algorithm that combines controllable generation and text rewriting methods using a Product of Experts with autoencoder language models (LMs). MaRCo uses likelihoods under a non-toxic LM (expert) and a toxic LM (anti-expert) to find candidate words to mask and potentially replace. We evaluate our method on several subtle toxicity and microaggressions datasets, and show that it not only outperforms baselines on automatic metrics, but MaRCo's rewrites are preferred 2.1 times more in human evaluation. Its applicability to instances of subtle toxicity is especially promising, demonstrating a path forward for addressing increasingly elusive online hate.

On the Interpretability and Significance of Bias Metrics in Texts: a PMI-based Approach
Francisco Valentini, Germán Federico Rosati, Damián Blasi, Diego Fernandez Slezak and Edgar Altszyler 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
In recent years, word embeddings have been widely used to measure biases in texts. Even if they have proven to be effective in detecting a wide variety of biases, metrics based on word embeddings lack transparency and interpretability. We analyze an alternative PMI-based metric to quantify biases in texts. It can be expressed as a function of conditional probabilities, which provides a simple interpretation in terms of word co-occurrences. We also prove that it can be approximated by an odds ratio, which allows estimating confidence intervals and statistical significance of textual biases. This approach produces similar results to metrics based on word embeddings when capturing gender gaps in the real world embedded in large corpora.

Knowledge of cultural moral norms in large language models
Aida Ramezani and Yang Xu   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Moral norms vary across cultures. A recent line of work suggests that English large language models contain human-like moral biases, but these studies typically do not examine moral variation in a diverse cultural setting. We investigate the extent to which monolingual English language models capture knowledge about moral norms in different countries. We consider two levels of analysis: 1) whether language models capture fine-grained moral variation across countries over a variety of topics such as "homosexuality" and "divorce"; 2) whether language models capture cultural diversity and shared tendencies in which topics people around the globe tend to diverge or agree on in their moral judgment. We perform our analyses with two public datasets from the World Values Survey (across 55 countries) and PEW global surveys (across 40 countries) on morality. We find that pre-trained English language models predict empirical moral norms across countries worse than the English moral norms reported previously. However, fine-tuning language models on the survey data improves inference across countries at the expense of a less accurate estimate of the English moral norms. We discuss the relevance and challenges of incorporating cultural knowledge into the automated inference of moral norms.

Deriving Language Models from Masked Language Models
Lucas Torroba Hennigen and Yoon Kim   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Masked language models (MLM) do not explicitly define a distribution over language, i.e., they are not language models per se. However, recent work has implicitly treated them as such for the purposes of generation and scoring. This paper studies methods for deriving explicit joint distributions from MLMs, focusing on distributions over two tokens, which makes it possible to calculate exact distributional properties. We find that an approach based on identifying joints whose conditionals are closest to those of the MLM works well and outperforms existing Markov random field-based approaches. We further find that this derived model's conditionals can even occasionally outperform the original MLM's conditionals.

An Invariant Learning Characterization of Controlled Text Generation
Carolina Zheng, Claudia Shi, Keyon Vafa, Amir Feder and David Blei   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Controlled generation refers to the problem of creating text that contains stylistic or semantic attributes of interest. Many approaches reduce this problem to training a predictor of the desired attribute. For example, researchers hoping to deploy a large language model to produce non-toxic content may use a toxicity classifier to filter generated text. In practice, the generated text to classify, which is determined by user prompts, may come from a wide range of distributions.

In this paper, we show that the performance of controlled generation may be poor if the distributions of text in response to user prompts differ from the distribution the predictor was trained on. To address this problem, we cast controlled generation under distribution shift as an invariant learning problem: the most effective predictor should be invariant across multiple text environments. We then discuss a natural solution that arises from this characterization and propose heuristics for selecting natural environments.

We study this characterization and the proposed method empirically using both synthetic and real data. Experiments demonstrate both the challenge of distribution shift in controlled generation and the potential of invariance methods in this setting.

One Network, Many Masks: Towards More Parameter-Efficient Transfer Learning
Guangtao Zeng, Peiyuan Zhang and Wei Lu   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Fine-tuning pre-trained language models for multiple tasks can be expensive in terms of storage. Parameter-efficient transfer learning (PETL) methods have been proposed to address this issue, but they still require a significant number of parameters when being applied to broader ranges of tasks. To achieve even greater storage reduction, we propose ProPETL, a novel method that enables efficient sharing of a single prototype PETL network (e.g. adapter, LoRA, and prefix-tuning) across layers and tasks. We learn binary masks to select different sub-networks from the prototype network and apply them as PETL modules into different layers. We find that the binary masks can determine crucial structural information from the network, which is often ignored in previous studies. Our work can also be seen as a type of pruning method, where we find that overparameterization also exists in the seemingly small PETL modules. We evaluate ProPETL on various downstream tasks and show that it can outperform other PETL methods with around 10% parameters required by the latter.

In and Out-of-Domain Text Adversarial Robustness via Label Smoothing

Yahan Yang, Soham Dan, Dan Roth and Insup Lee   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Recently it has been shown that state-of-the-art NLP models are vulnerable to adversarial attacks, where the predictions of a model can be drastically altered by slight modifications to the input (such as synonym substitutions). While several defense techniques have been proposed, and adapted, to the discrete nature of text adversarial attacks, the benefits of general-purpose regularization methods such as label smoothing for language models, have not been studied. In this paper, we study the adversarial robustness provided by label smoothing strategies in foundational models for diverse NLP tasks in both in-domain and out-of-domain settings. Our experiments show that label smoothing significantly improves adversarial robustness in pre-trained models like BERT, against various popular attacks. We also analyze the relationship between prediction confidence and robustness, showing that label smoothing reduces over-confident errors on adversarial examples.

Dataset Distillation with Attention Labels for Fine-tuning BERT

Aru Maekawa, Naoki Kobayashi, Kotaro Funakoshi and Manabu Okumura   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Dataset distillation aims to create a small dataset of informative synthetic samples to rapidly train neural networks that retain the performance of the original dataset. In this paper, we focus on constructing distilled few-shot datasets for natural language processing (NLP) tasks to fine-tune pre-trained transformers. Specifically, we propose to introduce attention labels, which can efficiently distill the knowledge from the original dataset and transfer it to the transformer models via attention probabilities. We evaluated our dataset distillation methods in four various NLP tasks and demonstrated that it is possible to create distilled few-shot datasets with the attention labels, yielding impressive performances for fine-tuning BERT. Specifically, in AGNews, a four-class news classification task, our distilled few-shot dataset achieved up to 93.2% accuracy, which is 98.5% performance of the original dataset even with only one sample per class and only one gradient step.

Improving the robustness of NLI models with minimax training

Michalis Korakakis and Andreas Vlachos   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Natural language inference (NLI) models are susceptible to learning shortcuts, i.e. decision rules that spuriously correlate with the label. As a result, they achieve high in-distribution performance, but fail to generalize to out-of-distribution samples where such correlations do not hold. In this paper, we present a training method to reduce the reliance of NLI models on shortcuts and improve their out-of-distribution performance without assuming prior knowledge of the shortcuts being targeted. To this end, we propose a minimax objective between a learner model being trained for the NLI task, and an auxiliary model aiming to maximize the learner's loss by up-weighting examples from regions of the input space where the learner incurs high losses. This process incentivizes the learner to focus on under-represented "hard" examples with patterns that contradict the shortcuts learned from the prevailing "easy" examples. Experimental results on three NLI datasets demonstrate that our method consistently outperforms other robustness enhancing techniques on out-of-distribution adversarial test sets, while maintaining high in-distribution accuracy.

Are Message Passing Neural Networks Really Helpful for Knowledge Graph Completion?

Juanhui Li, Harry Aaron Shomer, Jiayuan Ding, Yiqi Wang, Yao Ma, Neil Shah, Jiliang Tang and Dawei Yin   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Knowledge graphs (KGs) facilitate a wide variety of applications. Despite great efforts in creation and maintenance, even the largest KGs are far from complete. Hence, KG completion (KGC) has become one of the most crucial tasks for KG research. Recently, considerable literature in this space has centered around the use of Message Passing (Graph) Neural Networks (MPNNs), to learn powerful embeddings. The success of these methods is naturally attributed to the use of MPNNs over simpler multi-layer perceptron (MLP) models, given their additional message passing (MP) component. In this work, we find that surprisingly, simple MLP models are able to achieve comparable performance to MPNNs, suggesting that MP may not be as crucial as previously believed. With further exploration, we show careful scoring function and loss function design has a much stronger influence on KGC model performance. This suggests a conflation of scoring function design, loss function design, and MP in prior work, with promising insights regarding the scalability of state-of-the-art KGC methods today, as well as careful attention to more suitable MP designs for KGC tasks tomorrow.

Dissecting Transformer Length Extrapolation via the Lens of Receptive Field Analysis

Ta-Chung Chi, Ting-Han Fan, alexander rudnicky and Peter J Ramadge   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Length extrapolation permits training a transformer language model on short sequences that preserves perplexities when tested on substantially longer sequences. A relative positional embedding design, ALiBi, has had the widest usage to date. We dissect ALiBi via the lens of receptive field analysis empowered by a novel cumulative normalized gradient tool. The concept of receptive field further allows us to modify the vanilla Sinusoidal positional embedding to create **Sandwich**, the first parameter-free relative positional embedding design that truly length information uses longer than the training sequence. Sandwich shares with KERPLE and T5 the same logarithmic decaying temporal bias pattern with learnable relative positional embeddings; these elucidate future extrapolatable positional embedding design.

WebCPM: Interactive Web Search for Chinese Long-form Question Answering

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, Ruobing Xie, Fanchao Qi, Zhiyuan Liu, Maosong Sun and Jie Zhou   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Long-form question answering (LFQA) aims at answering complex, open-ended questions with detailed, paragraph-length responses. The de facto paradigm of LFQA necessitates two procedures: information retrieval, which searches for relevant supporting facts, and information synthesis, which integrates these facts into a coherent answer. In this paper, we introduce WebCPM, the first Chinese LFQA dataset. One unique feature of WebCPM is that its information retrieval is based on interactive web search, which engages with a search engine in real time. Following WebGPT, we develop a web search interface. We recruit annotators to search for relevant information using our interface and then answer questions. Meanwhile, the web search behaviors of our annotators would be recorded. In total, we collect 5,500 high-quality question-answer pairs, together with 15,372 supporting facts and 125,954 web search actions. We fine-tune pre-trained language models to imitate human behaviors for web search and to generate answers based on the collected facts. Our LFQA pipeline, built on these fine-tuned models, generates answers that are no worse than human-written ones in 32.5% and 47.5% of the cases on our dataset and DuReader, respectively. The interface, dataset, and codes are publicly available at https://github.com/thunlp/WebCPM.

Should you marginalize over possible tokenizations?

Nadezhda Chirkova, Germán Kruszewski, Jos Rozen and Marc Dymetman   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Autoregressive language models (LMs) map token sequences to probabilities. The usual practice for computing the probability of any character string (e.g. English sentences) is to first transform it into a sequence of tokens that is scored by the model. However, there are exponentially many token sequences that represent any given string. To truly compute the probability of a string one should marginalize over all tokenizations, which is typically intractable. Here, we analyze whether the practice of ignoring the marginalization is justified. To this end, we devise an importance-sampling-based algorithm that allows us to compute estimates of the marginal probabilities and compare them to the default procedure in a range of state-of-the-art models and datasets. Our results show that the gap in log-likelihood is no larger than 0.5% in most cases, but that it becomes more pronounced for data with long complex words.

ALERT: Adapt Language Models to Reasoning Tasks

Ping Yu, Tianlu Wang, Olga Golovneva, Badr AlKhamissi, Siddharth Verma, Zhijing Jin, Gargi Ghosh, Mona Diab and Asli Celikyilmaz
11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Recent advancements in large language models have enabled them to perform well on complex tasks that require step-by-step reasoning with few-shot learning. However, it is unclear whether these models are applying reasoning skills they have learnt during pre-training , or if they are simply memorizing their training corpus at finer granularity and have learnt to better understand their context. To address this question, we introduce {pasted macro 'OUR'}model, a benchmark and suite of analyses for evaluating reasoning skills of language models. {pasted macro 'OUR'}model enables comparing pre-trained and finetuned models on complex tasks that require reasoning skills to solve. Our benchmark provides a test bed to asses any language model on fine-grained reasoning skills, which spans over 20 datasets and covers 10 different reasoning skills. By using {pasted macro 'OUR'}model we further investigate *the role of finetuning*. Our extensive empirical analysis shows that language models learn more reasoning skills such as textual entailment, abductive reasoning, and analogical reasoning during the finetuning stage compared to pretraining stage. However, we also find that when language models are finetuned they tend to overfit to the prompt template, which hurts the robustness of models causing generalization problems.

HINT: Hypernetwork Instruction Tuning for Efficient Zero- and Few-Shot Generalisation

Hamish Ivison, Akshita Bhagia, Yizhong Wang, Hannaneh Hajishirzi and Matthew Peters   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Recent NLP models have shown the remarkable ability to effectively generalise 'zero-shot' to new tasks using only natural language instructions as guidance. However, many of these approaches suffer from high computational costs due to their reliance on concatenating lengthy instructions with every input example, resulting in costly reprocessing of the instruction. To avoid this, we introduce Hypernetworks for INstruction Tuning (HINT), which convert task instructions and examples into parameter-efficient modules inserted into an underlying model using a pretrained text encoder, eliminating the need to include instructions in the model input. The hypernetwork in HINT also produces an encoded instruction, which we concatenate with encoded inputs during decoding to further improve performance. HINT models outperform strong state-of-the-art baselines by over 10% when controlling for compute (measured in FLOPs). By converting instructions into modules, HINT models can effectively disregard the length of instructions and few-shot example inputs in terms of compute usage. As a result, HINT can enhance its performance by up to 25% by incorporating additional few-shot data, while utilizing only up to 5% more compute. This combines the strengths of parameter-efficient fine-tuning and in-context learning.

Token-wise Decomposition of Autoregressive Language Model Hidden States for Analyzing Model Predictions

Byung-Doh Oh and William Schuler   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

While there is much recent interest in studying why Transformer-based large language models make predictions the way they do, the complex computations performed within each layer have made their behavior somewhat opaque. To mitigate this opacity, this work presents a linear decomposition of final hidden states from autoregressive language models based on each initial input token, which is exact for virtually all contemporary Transformer architectures. This decomposition allows the definition of probability distributions that ablate the contribution of specific input tokens, which can be used to analyze their influence on model probabilities over a sequence of upcoming words with only one forward pass from the model. Using the change in next-word probability as a measure of importance, this work first examines which context words make the biggest contribution to language model predictions. Regression experiments suggest that Transformer-based language models rely primarily on collocational associations, followed by linguistic factors such as syntactic dependencies and coreference relationships in making next-word predictions. Additionally, analyses using these measures to predict syntactic dependencies and coreferent mention spans show that collocational association and repetitions of the same token largely explain the language models' predictions on these tasks.

Sequence Parallelism: Long Sequence Training from System Perspective

Shenggui Li, Fuzhao Xue, Chaitanya Baranwal, Yongbin Li and Yang You   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Transformer achieves promising results on various tasks. However, self-attention suffers from quadratic memory requirements with respect to the sequence length. Existing work focuses on reducing time and space complexity from an algorithm perspective. In this work, we propose sequence parallelism, a memory-efficient parallelism to solve this issue from system perspective instead. Our approach is compatible with most existing parallelisms (e.g., data, pipeline, and tensor parallelism), which means our sequence parallelism makes 4D parallelism possible. More importantly, we no longer require a single device to hold the whole sequence. Besides, using efficient attention with linear complexity, our sequence parallelism enables us to train transformer with infinite long sequence. Specifically, we split the input sequence into multiple chunks and feed each chunk into its corresponding device (i.e., GPU). To compute the attention output, we integrated ring-style communication with self-attention calculation and proposed Ring Self-Attention (RSA). Experiments show that sequence parallelism performs well when scaling with batch size and sequence length. Compared with tensor parallelism, our approach achieved $13.7\times$ and $3.0\times$ maximum batch size and sequence length respectively when scaling up to 64 NVIDIA P100 GPUs. With efficient attention, sequence can handle sequence with over 114K tokens, which is over $27\times$ longer than existing efficient attention works holding the whole sequence on a single device.

Soft Alignment Objectives for Robust Adaptation of Language Generation

Michal Štefánik, Marek Kadlčík and Petr Sojka   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Domain adaptation allows generative language models to address specific flaws caused by the domain shift of their application. However, the traditional adaptation by further training on in-domain data rapidly weakens the model's ability to generalize to other domains, making the open-ended deployments of the adapted models prone to errors.

This work introduces novel training objectives built upon a semantic similarity of the predicted tokens to the reference. Our results show that (1) avoiding the common assumption of a single correct prediction by constructing the training target from tokens' semantic similarity can largely mitigate catastrophic forgetting of adaptation, while (2) preserving the adaptation in-domain quality, (3) with negligible additions to compute costs.

In the broader context, the objectives grounded in a continuous token similarity pioneer the exploration of the middle ground between the efficient but naive exact-match token-level objectives and expressive but computationally- and resource-intensive sequential objectives.

In-sample Curriculum Learning by Sequence Completion for Natural Language Generation

Qi Jia, Yizhu Liu, Haifeng Tang and Kenny Q. Zhu   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Curriculum learning has shown promising improvements in multiple domains by training machine learning models from easy samples to hard ones. Previous works which either design rules or train models for scoring the difficulty highly rely on task-specific expertise, and cannot generalize. Inspired by the "easy-to-hard" intuition, we propose to do in-sample curriculum learning for natural language generation tasks. Our learning strategy starts training the model to generate the last few words, i.e., do sequence completion, and gradually extends to generate the whole output sequence. Comprehensive experiments show that it generalizes well to different tasks and achieves significant improvements over strong baselines.

EEL: Efficiently Encoding Lattices for Reranking

Prasann Singhal, Jiacheng Xu, Xi Ye and Greg Durrett   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Standard decoding approaches for conditional text generation tasks typically search for an output hypothesis with high model probability,

but this may not yield the best hypothesis according to human judgments of quality. Reranking to optimize for "downstream" metrics can more closely optimize for quality, but many metrics of interest are computed with pre-trained language models, which are slow to apply to large numbers of hypotheses. We explore an approach for reranking hypotheses by using Transformers to efficiently encode lattices of generated outputs, a method we call EEL. With a single Transformer pass over the entire lattice, we can approximately compute a contextualized representation of each token as if it were only part of a single hypothesis in isolation. We combine this approach with a new class of token-factored rerankers (TFRs) that allow for efficient extraction of high reranker-scoring hypotheses from the lattice. Empirically, our approach incurs minimal degradation error compared to the exponentially slower approach of encoding each hypothesis individually. When applying EEL with TFRs across three text generation tasks, our results show both substantial speedup compared to naive reranking and often better performance on downstream metrics than comparable approaches.

Element-aware Summarization with Large Language Models: Expert-aligned Evaluation and Chain-of-Thought Method

Yiming Wang, Zhuosheng Zhang and Rui Wang   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Automatic summarization generates concise summaries that contain key ideas of source documents. As the most mainstream datasets for the news sub-domain, CNN/DailyMail and BBC XSum have been widely used for performance benchmarking. However, the reference summaries of those datasets turn out to be noisy, mainly in terms of factual hallucination and information redundancy. To address this challenge, we first annotate new expert-writing Element-aware test sets following the "Lasswell Communication Model" proposed by Lasswell, allowing reference summaries to focus on more fine-grained news elements objectively and comprehensively. Utilizing the new test sets, we observe the surprising zero-shot summary ability of LLMs, which addresses the issue of the inconsistent results between human preference and automatic evaluation metrics of LLMs' zero-shot summaries in prior work. Further, we propose a Summary Chain-of-Thought (SumCoT) technique to elicit LLMs to generate summaries step by step, which helps them integrate more fine-grained details of source documents into the final summaries that correlate with the human writing mindset. Experimental results show our method outperforms state-of-the-art fine-tuned PLMs and zero-shot LLMs by +4.33/+4.77 in ROUGE-L on the two datasets, respectively. Dataset and code are publicly available at https://github.com/Alsace08/SumCoT.

Improving Factuality of Abstractive Summarization without Sacrificing Summary Quality

Tanay Dixit, Fei Wang and Muhao Chen   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Improving factual consistency of abstractive summarization has been a widely studied topic. However, most of the prior works on training factuality-aware models have ignored the negative effect it has on summary quality. We propose {pasted macro 'MODEL'}name (i.e. Effective Factual Summarization), a candidate summary generation and ranking technique to improve summary factuality without sacrificing quality. We show that using a contrastive learning framework with our refined candidate summaries leads to significant gains on both factuality and similarity-based metrics. Specifically, we propose a ranking strategy in which we effectively combine two metrics, thereby preventing any conflict during training. Models trained using our approach show up to 6 points of absolute improvement over the base model with respect to FactCC on XSUM and 11 points on CNN/DM, without negatively affecting either similarity-based metrics or abstractiveness.

Tokenization and the Noiseless Channel

Vilém Zouhar, Clara Meister, Juan Luis Gastaldi, Li Du, Mrinmaya Sachan and Ryan Cotterell   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Subword tokenization is a key part of most NLP pipelines. However, little is known about why some tokenizer and hyperparameter combinations lead to improved downstream model performance over others. We propose that good tokenizers lead to efficient channel usage, where the channel is the means by which some input is conveyed to the model and efficiency can be quantified in information-theoretic terms as the ratio of the Shannon entropy to the maximum entropy of the subword distribution. Nevertheless, an optimal encoding according to Shannon entropy assigns extremely long codes to low-frequency subwords and very short codes to high-frequency subwords. Defining efficiency in terms of Rényi entropy, on the other hand, penalizes distributions with either very high or very low-frequency subwords. We posit that (1) extremely high-frequency subwords are problematic because their meaning is not distinct and (2) that low-frequency subwords may not appear frequently enough for their meaning to be learned properly; encodings that induce unigram distributions with either can harm model performance. In machine translation, we find that across multiple tokenizers, the Rényi entropy has a very strong correlation with BLEU: 0.82 in comparison to just -0.30 for compressed length.

Text Style Transfer Back-Translation

Daimeng Wei, Zhanglin Wu, Hengchao Shang, Zongyao Li, Minghan Wang, Jiaxin GUO, Xiaoyu Chen, Zhengzhe YU and Hao Yang   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Back Translation (BT) is widely used in the field of machine translation, as it has been proved effective for enhancing translation quality. However, BT mainly improves the translation of inputs that share a similar style (to be more specific, translation-liked inputs), since the source side of BT data is machine-translated. For natural inputs, BT brings only slight improvements and sometimes even adverse effects. To address this issue, we propose Text Style Transfer Back Translation (TST BT), which uses a style transfer to modify the source side of BT data. By making the style of source-side text more natural, we aim to improve the translation of natural inputs. Our experiments on various language pairs, including both high-resource and low-resource ones, demonstrate that TST BT significantly improves translation performance against popular BT benchmarks. In addition, TST BT is proved to be effective in domain adaptation so this strategy can be regarded as a generalized data augmentation method. Our training code and text style transfer model are open-sourced.

Exploring Better Text Image Translation with Multimodal Codebook

Zhibin Lan, Jiawei Yu, Xiang Li, Wen Zhang, Jian Luan, Bin Wang, Degen Huang and Jinsong Su   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Text image translation (TIT) aims to translate the source texts embedded in the image to target translations, which has a wide range of applications and thus has important research value. However, current studies on TIT are confronted with two main bottlenecks: 1) this task lacks a publicly available TIT dataset, 2) dominant models are constructed in a cascaded manner, which tends to suffer from the error propagation of optical character recognition (OCR). In this work, we first annotate a Chinese-English TIT dataset named OCRMT30K, providing convenience for subsequent studies. Then, we propose a TIT model with a multimodal codebook, which is able to associate the image with relevant texts, providing useful supplementary information for translation. Moreover, we present a multi-stage training framework involving text machine translation, image-text alignment, and TIT tasks, which fully exploits additional bilingual texts, OCR dataset and our OCRMT30K dataset to train our model. Extensive experiments and in-depth analyses strongly demonstrate the effectiveness of our proposed model and training framework.

WACO: Word-Aligned Contrastive Learning for Speech Translation

Siqi Ouyang, Rong Ye and Lei Li   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

End-to-end Speech Translation (E2E ST) aims to directly translate source speech into target text. Existing ST methods perform poorly when only extremely small speech-text data are available for training. We observe that an ST model's performance closely correlates with its embedding similarity between speech and source transcript. In this paper, we propose Word-Aligned COntrastive learning (WACO), a simple and effective method for extremely low-resource speech-to-text translation. Our key idea is bridging word-level representations for both speech

and text modalities via contrastive learning. We evaluate WACO and other methods on the MuST-C dataset, a widely used ST benchmark, and on a low-resource direction Maltese-English from IWSLT 2023. Our experiments demonstrate that WACO outperforms the best baseline by 9+ BLEU points with only 1-hour parallel ST data. Code is available at https://github.com/owaski/WACO.

RAMP: Retrieval and Attribute-Marking Enhanced Prompting for Attribute-Controlled Translation
Gabriele Sarti, Phu Mon Htut, Xing Niu, Benjamin Hsu, Anna Currey, Georgiana Dinu and Maria Nadejde   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Attribute-controlled translation (ACT) is a subtask of machine translation that involves controlling stylistic or linguistic attributes (like formality and gender) of translation outputs. While ACT has garnered attention in recent years due to its usefulness in real-world applications, progress in the task is currently limited by dataset availability, since most prior approaches rely on supervised methods. To address this limitation, we propose Retrieval and Attribute-Marking enhanced Prompting (RAMP), which leverages large multilingual language models to perform ACT in few-shot and zero-shot settings. RAMP improves generation accuracy over the standard prompting approach by (1) incorporating a semantic similarity retrieval component for selecting similar in-context examples, and (2) marking in-context examples with attribute annotations. Our comprehensive experiments show that RAMP is a viable approach in both zero-shot and few-shot settings.

Scene Graph as Pivoting: Inference-time Image-free Unsupervised Multimodal Machine Translation with Visual Scene Hallucination
Hao Fei, Qian Liu, Meishan Zhang, Min Zhang and Tat-Seng Chua   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
In this work, we investigate a more realistic unsupervised multimodal machine translation (UMMT) setup, inference-time image-free UMMT, where the model is trained with source-text image pairs, and tested with only source-text inputs. First, we represent the input images and texts with the visual and language scene graphs (SG), where such fine-grained vision-language features ensure a holistic understanding of the semantics. To enable pure-text input during inference, we devise a visual scene hallucination mechanism that dynamically generates pseudo visual SG from the given textual SG. Several SG-pivoting based learning objectives are introduced for unsupervised translation training. On the benchmark Multi30K data, our SG-based method outperforms the best-performing baseline by significant BLEU scores on the task and setup, helping yield translations with better completeness, relevance and fluency without relying on paired images. Further in-depth analyses reveal how our model advances in the task setting.

mOKB6: A Multilingual Open Knowledge Base Completion Benchmark
Shubham Mittal, Keshav Kolluru, Soumen Chakrabarti and Mausam -   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Automated completion of open knowledge bases (Open KBs), which are constructed from triples of the form (subject phrase, relation phrase, object phrase), obtained via open information extraction (Open IE) system, are useful for discovering novel facts that may not be directly present in the text. However, research in Open KB completion (Open KBC) has so far been limited to resource-rich languages like English. Using the latest advances in multilingual Open IE, we construct the first multilingual Open KBC dataset, called mOKB6, containing facts from Wikipedia in six languages (including English). Improving the previous Open KB construction pipeline by doing multilingual coreference resolution and keeping only entity-linked triples, we create a dense Open KB. We experiment with several models for the task and observe a consistent benefit of combining languages with the help of shared embedding space as well as translations of facts. We also observe that current multilingual models struggle to remember facts seen in languages of different scripts.

WSPAlign: Word Alignment Pre-training via Large-Scale Weakly Supervised Span Prediction
Qiyu Wu, Masaaki Nagata and Yoshimasa Tsuruoka   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Most existing word alignment methods rely on manual alignment datasets or parallel corpora, which limits their usefulness. Here, to mitigate the dependence on manual data, we broaden the source of supervision by relaxing the requirement for correct, fully-aligned, and parallel sentences. Specifically, we make noisy, partially aligned, and non-parallel paragraphs in this paper. We then use such a large-scale weakly-supervised dataset for word alignment pre-training via span prediction. Extensive experiments with various settings empirically demonstrate that our approach, which is named WSPAlign, is an effective and scalable way to pre-train word aligners without manual data. When fine-tuned on standard benchmarks, WSPAlign has set a new state of the art by improving upon the best supervised baseline by 3.3 6.1 points in F1 and 1.5 6.1 points in AER. Furthermore, WSPAlign also achieves competitive performance compared with the corresponding baselines in few-shot, zero-shot and cross-lingual tests, which demonstrates that WSPAlign is potentially more practical for low-resource languages than existing methods.

Revisiting Relation Extraction in the era of Large Language Models
Somin Wadhwa, Silvio Amir and Byron C. Wallace   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Relation extraction (RE) is the core NLP task of inferring semantic relationships between entities from text. Standard supervised RE techniques entail training modules to tag tokens comprising entity spans and then predict the relationship between them. Recent work has instead treated the problem as a sequence-to-sequence task, linearizing relations between entities as target strings to be generated conditioned on the input. Here we push the limits of this approach, using larger language models (GPT-3 and Flan-T5 large) than considered in prior work and evaluating their performance on standard RE tasks under varying levels of supervision. We address issues inherent to evaluating generative approaches to RE by doing human evaluations, in lieu of relying on exact matching. Under this refined evaluation, we find that: (1) Few-shot prompting with GPT-3 achieves near SOTA performance, i.e., roughly equivalent to existing fully supervised models; (2) Flan-T5 is not as capable in the few-shot setting, but supervising and fine-tuning with Chain-of-Thought (CoT) style explanations (generated via GPT-3) yields SOTA results. We release this model as a new baseline for RE tasks.

Actively Supervised Clustering for Open Relation Extraction
Jun Zhao, Yongxin Zhang, Qi Zhang, Tao Gui, Zhongyu Wei, Minlong Peng and Mingming Sun   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Current clustering-based Open Relation Extraction (OpenRE) methods usually adopt a two-stage pipeline, which simultaneously learns relation representations and assignments in the first stage, then manually labels relation for each cluster. However, unsupervised objectives struggle to explicitly optimize clusters to align with relational semantics, and the number of clusters K has to be supplied in advance. In this paper, we present a novel setting, named actively supervised clustering for OpenRE. Our insight lies in that clustering learning and relation labeling can be performed simultaneously, which provides the necessary guidance for clustering without a significant increase in human effort. Along with this setting, we propose an active labeling strategy tailored for clustering. Instead of only focusing on improving the clustering of relations that have been discovered, our strategy is encouraged to discover new relations through diversity regularization. This is particularly beneficial for long-tail relations in the real world. Experimental results show that our method is able to discover almost all relational clusters in the data and improve the SOTA methods by 13.8% and 10.6%, on two datasets respectively.

Peeking inside the black box: A Commonsense-aware Generative Framework for Explainable Complaint Detection
Apoorva Singh, Raghav Jain, Prince Jha and Sriparna Saha   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Complaining is an illocutionary act in which the speaker communicates his/her dissatisfaction with a set of circumstances and holds the hearer (the complainee) answerable, directly or indirectly. Considering breakthroughs in machine learning approaches, the complaint detection task has piqued the interest of the natural language processing (NLP) community. Most of the earlier studies failed to justify their findings, neces-

sitating the adoption of interpretable models that can explain the model's output in real time. We introduce an explainable complaint dataset, X-CI, the first benchmark dataset for explainable complaint detection. Each instance in the X-CI dataset is annotated with five labels: complaint label, emotion label, polarity label, complaint severity level, and rationale (explainability), i.e., the causal span explaining the reason for the complaint/non-complaint label. We address the task of explainable complaint detection and propose a commonsense-aware unified generative framework by reframing the multitask problem as a text-to-text generation task. Our framework can predict the complaint cause, severity level, emotion, and polarity of the text in addition to detecting whether it is a complaint or not. We further establish the advantages of our proposed model on various evaluation metrics over the state-of-the-art models and other baselines when applied to the X-CI dataset in both full and few-shot settings.

Tackling Ambiguity with Images: Improved Multimodal Machine Translation and Contrastive Evaluation
Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot and Rachel Bawden   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
One of the major challenges of machine translation (MT) is ambiguity, which can in some cases be resolved by accompanying context such as images. However, recent work in multimodal MT (MMT) has shown that obtaining improvements from images is challenging, limited not only by the difficulty of building effective cross-modal representations, but also by the lack of specific evaluation and training data. We present a new MMT approach based on a strong text-only MT model, which uses neural adapters, a novel guided self-attention mechanism and which is jointly trained on both visually-conditioned masking and MMT. We also introduce CoMMuTE, a Contrastive Multilingual Multimodal Translation Evaluation set of ambiguous sentences and their possible translations, accompanied by disambiguating images corresponding to each translation. Our approach obtains competitive results compared to strong text-only models on standard English→French, English→German and English→Czech benchmarks and outperforms baselines and state-of-the-art MMT systems by a large margin on our contrastive test set. Our code and CoMMuTE are freely available.

World-to-Words: Grounded Open Vocabulary Acquisition through Fast Mapping in Vision-Language Models
Ziqiao Ma, Jiayi Pan and Joyce Chai   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The ability to connect language units to their referents in the physical world, referred to as grounding, is crucial to learning and understanding grounded meanings of words. While humans demonstrate fast mapping in new word learning, it remains unclear whether modern vision-language models can truly represent language with their grounded meanings, and how grounding may further bootstrap new word learning. To this end, we introduce Grounded Open Vocabulary Acquisition (GOVA) to examine grounding and bootstrapping in open-world language learning. As an initial attempt, we propose World-to-Words (W2W), a novel visually-grounded language model by pre-training on image-text pairs highlighting grounding as an objective. Through extensive experiments and analysis, we demonstrate that W2W is a more coherent and fast grounded word learner, and that the grounding ability acquired during pre-training helps the model to learn unseen words more rapidly and robustly.

Non-Sequential Graph Script Induction via Multimedia Grounding
Yu Zhou, Sha Li, Manling Li, Xudong Lin, Shih-Fu Chang, Mohit Bansal and Heng Ji   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Online resources such as WikiHow compile a wide range of scripts for performing everyday tasks, which can assist models in learning to reason about procedures. However, the scripts are always presented in a linear manner, which does not reflect the flexibility displayed by people executing tasks in real life. For example, in the CrossTask Dataset, 64.5% of consecutive step pairs are also observed in the reverse order, suggesting their ordering is not fixed. In addition, each step has an average of 2.56 frequent next steps, demonstrating "branching". In this paper, we propose the new challenging task of non-sequential graph script induction, aiming to capture optional and interchangeable steps in procedural planning. To automate the induction of such graph scripts for given tasks, we propose to take advantage of loosely aligned videos of people performing the tasks. In particular, we design a multimodal framework to ground procedural videos to WikiHow textual steps and thus transform each video into an observed step path on the latent ground truth graph script. This key transformation enables us to train a script knowledge model capable of both generating explicit graph scripts for learnt tasks and predicting future steps given a partial step sequence. Our best model outperforms the strongest pure text/vision baselines by 17.52% absolute gains on F1@3 for next step prediction and 13.8% absolute gains on Acc@1 for partial sequence completion. Human evaluation shows our model outperforming the WikiHow linear baseline by 48.76% absolute gains in capturing sequential and non-sequential step relationships.

Denoising Bottleneck with Mutual Information Maximization for Video Multimodal Fusion
Shaoxiang Wu, Damai Dai, Ziwei Qin, Tianyu Liu, Binghuai Lin, Yunbo Cao and Zhifang Sui 11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Video multimodal fusion aims to integrate multimodal signals in videos, such as visual, audio and text, to make a complementary prediction with multiple modalities contents. However, unlike other image-text multimodal tasks, video has longer multimodal sequences with more redundancy and noise in both visual and audio modalities. Prior denoising methods like forget gate are coarse in the granularity of noise filtering. They often suppress the redundant and noisy information at the risk of losing critical information. Therefore, we propose a denoising bottleneck fusion (DBF) model for fine-grained video multimodal fusion. On the one hand, we employ a bottleneck mechanism to filter out noise and redundancy with a restrained receptive field. On the other hand, we use a mutual information maximization module to regulate the filter-out module to preserve key information within different modalities. Our DBF model achieves significant improvement over current state-of-the-art baselines on multiple benchmarks covering multimodal sentiment analysis and multimodal summarization tasks. It proves that our model can effectively capture salient features from noisy and redundant video, audio, and text inputs. The code for this paper will be publicly available at https://github.com/WSXRHFG/DBF

SpeechMatrix: A Large-Scale Mined Corpus of Multilingual Speech-to-Speech Translations
Paul-Ambroise Augustin Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswami, Changhan Wang, Juan Pino, Benoît Sagot and Holger Schwenk   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
We present SpeechMatrix, a large-scale multilingual corpus of speech-to-speech translations mined from real speech of European Parliament recordings. It contains speech alignments in 136 language pairs with a total of 418 thousand hours of speech. To evaluate the quality of this parallel speech, we train bilingual speech-to-speech translation models on mined data only and establish extensive baseline results on EuroParl-ST, VoxPopuli and FLEURS test sets. Enabled by the multilinguality of SpeechMatrix, we also explore multilingual speech-to-speech translation, a topic which was addressed by few other works. We also demonstrate that model pre-training and sparse scaling using Mixture-of-Experts bring large gains to translation performance. The mined data and models will be publicly released

Compositional Generalization without Trees using Multiset Tagging and Latent Permutations
Matthias Lindemann, Alexander Koller and Ivan Titov   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Seq2seq models have been shown to struggle with compositional generalization in semantic parsing, i.e. generalizing to unseen compositions of phenomena that the model handles correctly in isolation.

We phrase semantic parsing as a two-step process: we first tag each input token with a multiset of output tokens. Then we arrange the tokens into an output sequence using a new way of parameterizing and predicting permutations. We formulate predicting a permutation as solving a regularized linear program and we backpropagate through the solver. In contrast to prior work, our approach does not place a priori restrictions on possible permutations, making it very expressive.

Our model outperforms pretrained seq2seq models and prior work on realistic semantic parsing tasks that require generalization to longer examples. We also outperform non-tree-based models on structural generalization on the COGS benchmark. For the first time, we show that a model without an inductive bias provided by trees achieves high accuracy on generalization to deeper recursion depth.

Advancing Multi-Criteria Chinese Word Segmentation Through Criterion Classification and Denoising
Tzu Hsuan Chou, Chun-Yi Lin and Hung-Yu Kao   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Recent research on multi-criteria Chinese word segmentation (MCCWS) mainly focuses on building complex private structures, adding more handcrafted features, or introducing complex optimization processes. In this work, we show that through a simple yet elegant input-hint-based MCCWS model, we can achieve state-of-the-art (SoTA) performances on several datasets simultaneously. We further propose a novel criterion-denoising objective that hurts slightly on F1 score but achieves SoTA recall on out-of-vocabulary words. Our result establishes a simple yet strong baseline for future MCCWS research. Source code is available at https://github.com/IKMLab/MCCWS.

Generic Temporal Reasoning with Differential Analysis and Explanation
Yu Feng, Ben Zhou, Haoyu Wang, Helen Jin and Dan Roth   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Temporal reasoning is the task of predicting temporal relations of event pairs. While temporal reasoning models can perform reasonably well on in-domain benchmarks, we have little idea of these systems' generalizability due to existing datasets' limitations. In this work, we introduce a novel task named TODAY that bridges this gap with temporal differential analysis, which as the name suggests, evaluates whether systems can correctly understand the effect of incremental changes. Specifically, TODAY introduces slight contextual changes for given event pairs, and systems are asked to tell how this subtle contextual change would affect relevant temporal relation distributions. To facilitate learning, TODAY also annotates human explanations. We show that existing models, including GPT-3.5, drop to random guessing on TODAY, suggesting that they heavily rely on spurious information rather than proper reasoning for temporal predictions. On the other hand, we show that TODAY's supervision style and explanation annotations can be used in joint learning, encouraging models to use more appropriate signals during training and thus outperform across several benchmarks. TODAY can also be used to train models to solicit incidental supervision from noisy sources such as GPT-3.5, thus moving us more toward the goal of generic temporal reasoning systems.

Unbalanced Optimal Transport for Unbalanced Word Alignment
Yuki Arase, Han Bao and Sho Yokoi   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Monolingual word alignment is crucial to model semantic interactions between sentences. In particular, null alignment, a phenomenon in which words have no corresponding counterparts, is pervasive and critical in handling semantically divergent sentences. Identification of null alignment is useful on its own to reason about the semantic similarity of sentences by indicating there exists information inequality. To achieve unbalanced word alignment that values both alignment and null alignment, this study shows that the family of optimal transport (OT), i.e., balanced, partial, and unbalanced OT, are natural and powerful approaches even without tailor-made techniques. Our extensive experiments covering unsupervised and supervised settings indicate that our generic OT-based alignment methods are competitive against the state-of-the-arts specially designed for word alignment, remarkably on challenging datasets with high null alignment frequencies.

Conjunct Resolution in the Face of Verbal Omissions
Royi Rassin, Yoav Goldberg and Reut Tsarfaty   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Verbal omissions are complex syntactic phenomena in VP coordination structures. They occur when verbs and (some of) their arguments are omitted from subsequent clauses after being explicitly stated in an initial clause. Recovering these omitted elements is necessary for accurate interpretation of the sentence, and while humans easily and intuitively fill in the missing information, state-of-the-art models continue to struggle with this task. Previous work is limited to small-scale datasets, synthetic data creation methods, and to resolution methods in the dependency-graph level. In this work we propose a *conjunct resolution* task that operates directly on the text and makes use of a *split-and-rephrase* paradigm in order to recover the missing elements in the coordination structure. To this end, we first formulate a pragmatic framework of verbal omissions which describes the different types of omissions, and develop an automatic scalable collection method. Based on this method, we curate a large dataset, containing over 10K examples of naturally-occurring verbal omissions with crowd-sourced annotations of the resolved conjuncts. We train various neural baselines for this task, and show that while our best method obtains decent performance, it leaves ample space for improvement. We propose our dataset, metrics and models as a starting point for future research on this topic.

Annotation-Inspired Implicit Discourse Relation Classification with Auxiliary Discourse Connective Generation
Wei Liu and Michael Strube   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Implicit discourse relation classification is a challenging task due to the absence of discourse connectives. To overcome this issue, we design an end-to-end neural model to explicitly generate discourse connectives for the task, inspired by the annotation process of PDTB. Specifically, our model jointly learns to generate discourse connectives between arguments and predict discourse relations based on the arguments and the generated connectives. To prevent our relation classifier from being misled by poor connectives generated at the early stage of training while alleviating the discrepancy between training and inference, we adopt Scheduled Sampling to the joint learning. We evaluate our method on three benchmarks, PDTB 2.0, PDTB 3.0, and PCC. Results show that our joint model significantly outperforms various baselines on three datasets, demonstrating its superiority for the task.

GanLM: Encoder-Decoder Pre-training with an Auxiliary Discriminator
Jian Yang, Shuming Ma, Li Dong, Shaohan Huang, Haoyang Huang, Yuwei Yin, Dongdong Zhang, Liqun Yang, Furu Wei and Zhoujun Li
11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Pre-trained models have achieved remarkable success in natural language processing (NLP). However, existing pre-training methods underutilize the benefits of language understanding for generation. Inspired by the idea of Generative Adversarial Networks (GANs), we propose a GAN-style model for encoder-decoder pre-training by introducing an auxiliary discriminator, unifying the ability of language understanding and generation in a single model. Our model, named as GanLM, is trained with two pre-training objectives: replaced token detection and replaced token denoising. Specifically, given masked source sentences, the generator outputs the target distribution and the discriminator predicts whether the target sampled tokens from distribution are incorrect. The target sentence is replaced with misclassified tokens to construct noisy previous context, which is used to generate the gold sentence. In general, both tasks improve the ability of language understanding and generation by selectively using the denoising data. Extensive experiments in language generation benchmarks show that GanLM with the powerful language understanding capability outperforms various strong pre-trained language models (PLMs) and achieves state-of-the-art performance.

Stop Pre-Training: Adapt Visual-Language Models to Unseen Languages
Yasmine Karoui, Rémi Lebret, Negar Foroutan Eghlidi and Karl Aberer   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Vision-Language Pre-training (VLP) has advanced the performance of many vision-language tasks, such as image-text retrieval, visual entailment, and visual reasoning. The pre-training mostly utilizes lexical databases and image queries in English. Previous work has demonstrated that the pre-training in English does not transfer well to other languages in a zero-shot setting. However, multilingual pre-trained language models (MPLM) have excelled at a variety of single-modal language tasks. In this paper, we propose a simple yet efficient approach to adapt

VLP to unseen languages using MPLM. We utilize a cross-lingual contextualised token embeddings alignment approach to train text encoders for non-English languages. Our approach does not require image input and primarily uses machine translation, eliminating the need for target language data. Our evaluation across three distinct tasks (image-text retrieval, visual entailment, and natural language visual reasoning) demonstrates that this approach outperforms the state-of-the-art multilingual vision-language models without requiring large parallel corpora. Our code is available at https://github.com/Yasminekaroui/CliCoTea.

Cross-lingual Continual Learning

Meryem M'hamdi, Xiang Ren and Jonathan May   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

The longstanding goal of multi-lingual learning has been to develop a universal cross-lingual model that can withstand the changes in multi-lingual data distributions. There has been a large amount of work to adapt such multi-lingual models to unseen target languages. However, the majority of work in this direction focuses on the standard one-hop transfer learning pipeline from source to target languages, whereas in realistic scenarios, new languages can be incorporated at any time in a sequential manner. In this paper, we present a principled Cross-lingual Continual Learning (CCL) evaluation paradigm, where we analyze different categories of approaches used to continually adapt to emerging data from different languages. We provide insights into what makes multilingual sequential learning particularly challenging. To surmount such challenges, we benchmark a representative set of cross-lingual continual learning algorithms and analyze their knowledge preservation, accumulation, and generalization capabilities compared to baselines on carefully curated datastreams. The implications of this analysis include a recipe for how to measure and balance different cross-lingual continual learning desiderata, which go beyond conventional transfer learning.

Large-Scale Correlation Analysis of Automated Metrics for Topic Models

Jia Peng Lim and Hady Lauw   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Automated coherence metrics constitute an important and popular way to evaluate topic models. Previous works present a mixed picture of their presumed correlation with human judgement. In this paper, we conduct a large-scale correlation analysis of coherence metrics. We propose a novel sampling approach to mine topics for the purpose of metric evaluation, and conduct the analysis via three large corpora showing that certain automated coherence metrics are correlated. Moreover, we extend the analysis to measure topical differences between corpora. Lastly, we examine the reliability of human judgement by conducting an extensive user study, which is designed as an amalgamation of different proxy tasks to derive a finer insight into the human decision-making processes. Our findings reveal some correlation between automated coherence metrics and human judgement, especially for generic corpora.

An Ordinal Latent Variable Model of Conflict Intensity

Niklas Stoehr, Lucas Torroba Hennigen, Josef Valvoda, Robert West, Ryan Cotterell and Aaron Schein 11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Measuring the intensity of events is crucial for monitoring and tracking armed conflict. Advances in automated event extraction have yielded massive data sets of "who did what to whom" micro-records that enable data-driven approaches to monitoring conflict. The Goldstein scale is a widely-used expert-based measure that scores events on a conflictual–cooperative scale. It is based only on the action category ("what") and disregards the subject ("who") and object ("to whom") of an event, as well as contextual information, like associated casualty count, that should contribute to the perception of an event's "intensity". This paper takes a latent variable-based approach to measuring conflict intensity. We introduce a probabilistic generative model that assumes each observed event is associated with a latent intensity class. A novel aspect of this model is that it imposes an ordering on the classes, such that higher-valued classes denote higher levels of intensity. The ordinal nature of the latent variable is induced from naturally ordered aspects of the data (e.g., casualty counts) where higher values naturally indicate higher intensity. We evaluate the proposed model both intrinsically and extrinsically, showing that it obtains comparatively good held-out predictive performance.

Finding the SWEET Spot: Analysis and Improvement of Adaptive Inference in Low Resource Settings

Daniel Rotem, Michael Hassid, Jonathan Mamou and Roy Schwartz   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Adaptive inference is a simple method for reducing inference costs. The method works by maintaining multiple classifiers of different capacities, and allocating resources to each test instance according to its difficulty. In this work, we compare the two main approaches for adaptive inference, Early-Exit and Multi-Model, when training data is limited. First, we observe that for models with the same architecture and size, individual Multi-Model classifiers outperform their Early-Exit counterparts by an average of 2.3%. We show that this gap is caused by Early-Exit classifiers sharing model parameters during training, resulting in conflicting gradient updates of model weights. We find that despite this gap, Early-Exit still provides a better speed-accuracy trade-off due to the overhead of the Multi-Model approach. To address these issues, we propose SWEET (Separating Weights for Early-Exit Transformers) an Early-Exit fine-tuning method that assigns each classifier its own set of unique model weights, not updated by other classifiers. We compare SWEET's speed-accuracy curve to standard Early-Exit and Multi-Model baselines and find that it outperforms both methods at fast speeds while maintaining comparable scores to Early-Exit at slow speeds. Moreover, SWEET individual classifiers outperform Early-Exit ones by 1.1% on average. SWEET enjoys the benefits of both methods, paving the way for further reduction of inference costs in NLP.

Language model acceptability judgements are not always robust to context

Koustuv Sinha, Jon Gauthier, Aaron Mueller, Kanishka Misra, Keren Fuentes, Roger Levy and Adina Williams   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Targeted syntactic evaluations of language models ask whether models show stable preferences for syntactically acceptable content over minimal-pair unacceptable inputs. Our best syntactic evaluation datasets, however, provide substantially less linguistic context than models receive during pretraining. This mismatch raises an important question: how robust are models' syntactic judgements across different contexts? In this paper, we vary the input contexts based on: length, the types of syntactic phenomena it contains, and whether or not there are grammatical violations. We find that model judgements are generally robust when placed in randomly sampled linguistic contexts, but are unstable when contexts match the test stimuli in syntactic structure. Among all tested models (GPT-2 and five variants of OPT), we find that model performance is affected when we provided contexts with matching syntactic structure: performance significantly improves when contexts are acceptable, and it significantly declines when they are unacceptable. This effect is amplified by the length of the context, except for unrelated inputs. We show that these changes in model performance are not explainable by acceptability-preserving syntactic perturbations. This sensitivity to highly specific syntactic features of the context can only be explained by the models' implicit in-context learning abilities.

BITE: Textual Backdoor Attacks with Iterative Trigger Injection

Jun Yan, Vansh Gupta and Xiang Ren   11:00-12:30 (Frontenac Ballroom and Queen's Quay)

Backdoor attacks have become an emerging threat to NLP systems. By providing poisoned training data, the adversary can embed a "backdoor" into the victim model, which allows input instances satisfying certain textual patterns (e.g., containing a keyword) to be predicted as a target label of the adversary's choice. In this paper, we demonstrate that it is possible to design a backdoor attack that is both stealthy (i.e., hard to notice) and effective (i.e., has a high attack success rate). We propose BITE, a backdoor attack that poisons the training data to establish strong correlations between the target label and a set of "trigger words". These trigger words are iteratively identified and injected into the target-label instances through natural word-level perturbations. The poisoned training data instruct the victim model to predict the

target label on inputs containing trigger words, forming the backdoor. Experiments on four text classification datasets show that our proposed attack is significantly more effective than baseline methods while maintaining decent stealthiness, raising alarm on the usage of untrusted training data. We further propose a defense method named DeBITE based on potential trigger word removal, which outperforms existing methods in defending against BITE and generalizes well to handling other backdoor attacks.

Contrastive Error Attribution for Finetuned Language Models
Faisal Ladhak, Esin Durmus and Tatsunori Hashimoto   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Recent work has identified noisy and misannotated data as a core cause of hallucinations and unfaithful outputs in Natural Language Generation (NLG) tasks. Consequently, identifying and removing these examples is a key open challenge in creating reliable NLG systems. In this work, we introduce a framework to identify and remove low-quality training instances that lead to undesirable outputs, such as faithfulness errors in text summarization. We show that existing approaches for error tracing, such as gradient-based influence measures, do not perform reliably for detecting faithfulness errors in NLG datasets. We overcome the drawbacks of existing error tracing methods through a new, contrast-based estimate that compares undesired generations to human-corrected outputs. Our proposed method can achieve a mean average precision of 0.93 at detecting known data errors across synthetic tasks with known ground truth, substantially outperforming existing approaches. Using this approach and re-training models on cleaned data leads to a 70% reduction in entity hallucinations on the NYT dataset and a 55% reduction in semantic errors on the E2E dataset.

Subjective Crowd Disagreements for Subjective Data: Uncovering Meaningful CrowdOpinion with Population-level Learning
Tharindu Cyril Weerasooriya, Sarah Luger, Saloni Poddar, Ashiqur KhudaBukhsh and Christopher Homan   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Human-annotated data plays a critical role in the fairness of AI systems, including those that deal with life-altering decisions or moderating human-created web/social media content. Conventionally, annotator disagreements are resolved before any learning takes place. However, researchers are increasingly identifying annotator disagreement as pervasive and meaningful. They also question the performance of a system when annotators disagree. Particularly when minority views are disregarded, especially among groups that may already be underrepresented in the annotator population. In this paper, we introduce CrowdOpinion, an unsupervised learning based approach that uses language features and label distributions to pool similar items into larger samples of label distributions. We experiment with four generative and one density-based clustering method, applied to five linear combinations of label distributions and features. We use five publicly available benchmark datasets (with varying levels of annotator disagreements) from social media (Twitter, Gab, and Reddit). We also experiment in the wild using a dataset from Facebook, where annotations come from the platform itself by users reacting to posts. We evaluate CrowdOpinion as a label distribution prediction task using KL-divergence and a single-label problem using accuracy measures.

Exploiting Biased Models to De-bias Text: A Gender-Fair Rewriting Model
Chantal Amrhein, Florian Schottmann, Rico Sennrich and Samuel Läubli   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Natural language generation models reproduce and often amplify the biases present in their training data. Previous work explored using sequence-to-sequence rewriting models to transform biased model outputs (or original texts) into more gender-fair language by creating pseudo training data through linguistic rules. However, this approach is not practical for languages with more complex morphology than English. We hypothesise that creating training data in the reverse direction, i.e. starting from gender-fair text, is easier for morphologically complex languages and show that it matches the performance of state-of-the-art rewriting models for English. To eliminate the rule-based nature of data creation, we instead propose using machine translation models to create gender-biased text from real gender-fair text via round-trip translation. Our approach allows us to train a rewriting model for German without the need for elaborate handcrafted rules. The outputs of this model increased gender-fairness as shown in a human evaluation study.

BLIND: Bias Removal With No Demographics
Hadas Orgad and Yonatan Belinkov   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
Models trained on real-world data tend to imitate and amplify social biases. Common methods to mitigate biases require prior information on the types of biases that should be mitigated (e.g., gender or racial bias) and the social groups associated with each data sample. In this work, we introduce BLIND, a method for bias removal with no prior knowledge of the demographics in the dataset. While training a model on a downstream task, BLIND detects biased samples using an auxiliary model that predicts the main model's success, and down-weights those samples during the training process. Experiments with racial and gender biases in sentiment classification and occupation classification tasks demonstrate that BLIND mitigates social biases without relying on a costly demographic annotation process. Our method is competitive with other methods that require demographic information and sometimes even surpasses them.

Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models
Myra Cheng, Esin Durmus and Dan Jurafsky   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
To recognize and mitigate harms from large language models (LLMs), we need to understand the prevalence and nuances of stereotypes in LLM outputs. Toward this end, we present Marked Personas, a prompt-based method to measure stereotypes in LLMs for intersectional demographic groups without any lexicon or data labeling. Grounded in the sociolinguistic concept of markedness (which characterizes explicitly linguistically marked categories versus unmarked defaults), our proposed method is twofold: 1) prompting an LLM to generate personas, i.e., natural language descriptions, of the target demographic group alongside personas of unmarked, default groups; 2) identifying the words that significantly distinguish personas of the target group from corresponding unmarked ones. We find that the portrayals generated by GPT-3.5 and GPT-4 contain higher rates of racial stereotypes than human-written portrayals using the same prompts. The words distinguishing personas of marked (non-white, non-male) groups reflect patterns of othering and exoticizing these demographics. An intersectional lens further reveals tropes that dominate portrayals of marginalized groups, such as tropicalism and the hypersexualization of minoritized women. These representational harms have concerning implications for downstream applications like story generation.

Do CoNLL-2003 Named Entity Taggers Still Work Well in 2023?
Shuheng Liu and Alan Ritter   11:00-12:30 (Frontenac Ballroom and Queen's Quay)
The CoNLL-2003 English named entity recognition (NER) dataset has been widely used to train and evaluate NER models for almost 20 years. However, it is unclear how well models that are trained on this 20-year-old data and developed over a period of decades using the same test set will perform when applied on modern data. In this paper, we evaluate the generalization of over 20 different models trained on CoNLL-2003, and show that NER models have very different generalization. Surprisingly, we find no evidence of performance degradation in pre-trained Transformers, such as RoBERTa and T5, even when fine-tuned using decades-old data. We investigate why some models generalize well to new data while others do not, and attempt to disentangle the effects of temporal drift and overfitting due to test reuse. Our analysis suggests that most deterioration is due to temporal mismatch between the pre-training corpora and the downstream test sets. We found that four factors are important for good generalization: model architecture, number of parameters, time period of the pre-training corpus, in addition to the amount of fine-tuning data. We suggest current evaluation methods have, in some sense, underestimated progress on NER over the past 20 years, as NER models have not only improved on the original CoNLL-2003 test set, but improved even more on modern data. Our datasets can be found at https://github.com/ShuhengL/acl2023_conllpp.

Discourse and Pragmatics

11:00-12:30 (Pier 2&3)

[TACL] Multilingual Coreference Resolution in Multiparty Dialogue
Boyuan Zheng, Patrick Xia, Mahsa Yarmohammadi and Benjamin Van Durme   11:00-11:15 (Pier 2&3)
Existing multiparty dialogue datasets for coreference resolution are nascent, and many challenges are still unaddressed. We create a large-scale dataset, Multilingual Multiparty Coref (MMC), for this task based on TV transcripts. Due to the availability of gold-quality subtitles in multiple languages, we propose reusing the annotations to create silver coreference data in other languages (Chinese and Farsi) via annotation projection. On the gold (English) data, off-the-shelf models perform relatively poorly on MMC, suggesting that MMC has broader coverage of multiparty coreference than prior datasets. On the silver data, we find success both using it for data augmentation and training from scratch, which effectively simulates the zero-shot cross-lingual setting.

[TACL] Coreference Resolution through a seq2seq Transition-Based System
Bernd Bohnet, Chris Alberti and Michael Collins   11:15-11:30 (Pier 2&3)
Most recent coreference resolution systems use search algorithms over possible spans to identify mentions and resolve coreference. We instead present a coreference resolution system that uses a text-to-text (seq2seq) paradigm to predict mentions and links jointly. We implement the coreference system as a transition system and use multilingual T5 as an underlying language model. We obtain state-of-the-art accuracy on the CoNLL-2012 datasets with 83.3 F1-score for English (a 2.3 higher F1-score than previous work) using only CoNLL data for training, 68.5 F1-score for Arabic (+4.1 higher than previous work) and 74.3 F1-score for Chinese (+5.3). In addition we use the SemEval-2010 data sets for experiments in the zero-shot setting, a few-shot setting, and supervised setting using all available training data. We get substantially higher zero-shot F1-scores for 3 out of 4 languages than previous approaches and significantly exceed previous supervised state-of-the-art results for all five tested languages.

PairSpanBERT: An Enhanced Language Model for Bridging Resolution
Hideo Kobayashi, Yufang Hou and Vincent Ng   11:30-11:45 (Pier 2&3)
We present PairSpanBERT, a SpanBERT-based pre-trained model specialized for bridging resolution. To this end, we design a novel pre-training objective that aims to learn the contexts in which two mentions are implicitly linked to each other from a large amount of data automatically generated either heuristically or via distance supervision with a knowledge graph. Despite the noise inherent in the automatically generated data, we achieve the best results reported to date on three evaluation datasets for bridging resolution when replacing SpanBERT with PairSpanBERT in a state-of-the-art resolver that jointly performs entity coreference resolution and bridging resolution.

Annotating Mentions Alone Enables Efficient Domain Adaptation for Coreference Resolution
Nupoor Gandhi, Anjalie Field and Emma Strubell   11:45-12:00 (Pier 2&3)
Although recent neural models for coreference resolution have led to substantial improvements on benchmark datasets, it remains a challenge to successfully transfer these models to new target domains containing many out-of-vocabulary spans and requiring differing annotation schemes. Typical approaches involve continued training on annotated target-domain data, but obtaining annotations is costly and time-consuming. In this work, we show that adapting mention detection is the key component to successful domain adaptation of coreference models, rather than antecedent linking. We also show annotating mentions alone is nearly twice as fast as annotating full coreference chains. Based on these insights, we propose a method for efficiently adapting coreference models, which includes a high-precision mention detection objective and requires only mention annotations in the target domain. Extensive evaluation across three English coreference datasets: CoNLL-2012 (news/conversation), i2b2/VA (medical notes), and child welfare notes, reveals that our approach facilitates annotation-efficient transfer and results in a 7-14% improvement in average F1 without increasing annotator time.

Dual Cache for Long Document Neural Coreference Resolution
Qipeng Guo, Xiangkun Hu, Yue Zhang, Xipeng Qiu and Zheng Zhang   12:00-12:15 (Pier 2&3)
Recent works show the effectiveness of cache-based neural coreference resolution models on long documents. These models incrementally process a long document from left to right and extract relations between mentions and entities in a cache, resulting in much lower memory and computation cost compared to computing all mentions in parallel. However, they do not handle cache misses when high-quality entities are purged from the cache, which causes wrong assignments and leads to prediction errors. We propose a new hybrid cache that integrates two eviction policies to capture global and local entities separately, and effectively reduces the aggregated cache misses up to half as before, while improving F1 score of coreference by 0.7 5.7pt. As such, the hybrid policy can accelerate existing cache-based models and offer a new long document coreference resolution solution. Results show that our method outperforms existing methods on four benchmarks while saving up to 83% of inference time against non-cache-based models. Further, we achieve a new state-of-the-art on a long document coreference benchmark, LitBank.

Factual or Contextual? Disentangling Error Types in Entity Description Generation
Navita Goyal, Ani Nenkova and Hal Daumé III   12:15-12:30 (Pier 2&3)
In the task of entity description generation, given a context and a specified entity, a model must describe that entity correctly and in a contextually-relevant way. In this task, as well as broader language generation tasks, the generation of a nonfactual description (factual error) versus an incongruous description (contextual error) is fundamentally different, yet often conflated. We develop an evaluation paradigm that enables us to disentangle these two types of errors in naturally occurring textual contexts. We find that factuality and congruity are often at odds, and that models specifically struggle with accurate descriptions of entities that are less familiar to people. This shortcoming of language models raises concerns around the trustworthiness of such models, since factual errors on less well-known entities are exactly those that a human reader will not recognize.

Speech and Multimodality

11:00-12:30 (Pier 4&5)

Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation
Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky and Martijn Wieling   11:00-11:15 (Pier 4&5)
The performance of automatic speech recognition (ASR) systems has advanced substantially in recent years, particularly for languages for which a large amount of transcribed speech is available. Unfortunately, for low-resource languages, such as minority languages, regional languages or dialects, ASR performance generally remains much lower. In this study, we investigate whether data augmentation techniques

could help improve low-resource ASR performance, focusing on four typologically diverse minority languages or language variants (West Germanic: Gronings, West-Frisian; Malayo-Polynesian: Besemah, Nasal). For all four languages, we examine the use of self-training, where an ASR system trained with the available human-transcribed data is used to generate transcriptions, which are then combined with the original data to train a new ASR system. For Gronings, for which there was a pre-existing text-to-speech (TTS) system available, we also examined the use of TTS to generate ASR training data from text-only sources. We find that using a self-training approach consistently yields improved performance (a relative WER reduction up to 20.5% compared to using an ASR system trained on 24 minutes of manually transcribed speech). The performance gain from TTS augmentation for Gronings was even stronger (up to 25.5% relative reduction in WER compared to a system based on 24 minutes of manually transcribed speech). In sum, our results show the benefit of using self-training or (if possible) TTS-generated data as an efficient solution to overcome the limitations of data availability for resource-scarce languages in order to improve ASR performance.

ManagerTower: Aggregating the Insights of Uni-Modal Experts for Vision-Language Representation Learning
Xiao Xu, Bei Li, Chenfei Wu, Shao-Yen Tseng, Anahita Bhiwandiwalla, Shachar Rosenman, Vasudev Lal, Wanxiang Che and Nan Duan
11:15-11:30 (Pier 4&5)
Two-Tower Vision-Language (VL) models have shown promising improvements on various downstream VL tasks. Although the most advanced work improves performance by building bridges between encoders, it suffers from ineffective layer-by-layer utilization of uni-modal representations and cannot flexibly exploit different levels of uni-modal semantic knowledge. In this work, we propose ManagerTower, a novel VL model architecture that gathers and combines the insights of pre-trained uni-modal experts at different levels. The managers introduced in each cross-modal layer can adaptively aggregate uni-modal semantic knowledge to facilitate more comprehensive cross-modal alignment and fusion. ManagerTower outperforms previous strong baselines both with and without Vision-Language Pre-training (VLP). With only 4M VLP data, ManagerTower achieves superior performances on various downstream VL tasks, especially 79.15% accuracy on VQAv2 Test-Std, 86.56% IR@1 and 95.64% TR@1 on Flickr30K. Code and checkpoints are available at https://github.com/LooperXX/ManagerTower.

OpenSR: Open-Modality Speech Recognition via Maintaining Multi-Modality Alignment
Xize Cheng, Tao Jin, Linjun Li, Wang Lin, Xinyu Duan and Zhou Zhao
11:30-11:45 (Pier 4&5)
Speech Recognition builds a bridge between the multimedia streaming (audio-only, visual-only or audio-visual) and the corresponding text transcription. However, when training the specific model of new domain, it often gets stuck in the lack of new-domain utterances, especially the labeled visual utterances. To break through this restriction, we attempt to achieve zero-shot modality transfer by maintaining the multi-modality alignment in phoneme space learned with unlabeled multimedia utterances in the high resource domain during the pre-training, and propose a training system Open-modality Speech Recognition (**OpenSR**) that enables the models trained on a single modality (e.g., audio-only) applicable to more modalities (e.g., visual-only and audio-visual). Furthermore, we employ a cluster-based prompt tuning strategy to handle the domain shift for the scenarios with only common words in the new domain utterances. We demonstrate that OpenSR enables modality transfer from one to any in three different settings (zero-, few- and full-shot), and achieves highly competitive zero-shot performance compared to the existing few-shot and full-shot lip-reading methods. To the best of our knowledge, OpenSR achieves the state-of-the-art performance of word error rate in LRS2 on audio-visual speech recognition and lip-reading with 2.7% and 25.0%, respectively.

Hearing Lips in Noise: Universal Viseme-Phoneme Mapping and Transfer for Robust Audio-Visual Speech Recognition
Yuchen Hu, Ruizhe Li, Chen Chen, Chengwei Qin, Qiu-Shi Zhu and Eng Siong Chng
11:45-12:00 (Pier 4&5)
Audio-visual speech recognition (AVSR) provides a promising solution to ameliorate the noise-robustness of audio-only speech recognition with visual information. However, most existing efforts still focus on audio modality to improve robustness considering its dominance in AVSR task, with noise adaptation techniques such as front-end denoise processing. Though effective, these methods are usually faced with two practical challenges: 1) lack of sufficient labeled noisy audio-visual training data in some real-world scenarios and 2) less optimal model generality to unseen testing noises. In this work, we investigate the noise-invariant visual modality to strengthen robustness of AVSR, which can adapt to any testing noises while without dependence on noisy training data, a.k.a., unsupervised noise adaptation. Inspired by human perception mechanism, we propose a universal viseme-phoneme mapping (UniVPM) approach to implement modality transfer, which can restore clean audio from visual signals to enable speech recognition under any noisy conditions. Extensive experiments on public benchmarks LRS3 and LRS2 show that our approach achieves the state-of-the-art under various noisy as well as clean conditions. In addition, we also outperform previous state-of-the-arts on visual speech recognition task.

Vision Language Pre-training with Contrastive Learning with Cross-Modal Similarity Regulation
Chaoya Jiang, Wei Ye, Haiyang Xu, Songfang Huang, Fei Huang and Shikun Zhang
12:00-12:15 (Pier 4&5)
In this paper, we reconsider the problem of (partial) false negative samples from the Mutual Information (MI) Maximization perspective, the traditional contrastive loss (like InfoNCE loss) will equally push away the anchor of all positive samples and negative samples regardless of their possible semantic similarities. We theoretically show that InfoNCE loss will not only maximize the MI between the anchor and positive samples but minimize the MI between the anchor and false negative samples even though they share similar semantic which could provide a possible theoretical explanation for the observation of the existence of false negative samples in the cross-modal contrastive learning will decrease the downstream task performance of VLP models. Above analysis motivate us to propose the VLP model with a novel Semantic Awared Contrastive Learning framework named SACL where different negative samples are assigned with different contrastive weights according to the semantic similarity between them and the anchor.

MIR-GAN: Refining Frame-Level Modality-Invariant Representations with Adversarial Network for Audio-Visual Speech Recognition
Yuchen Hu, Chen Chen, Ruizhe Li, Heqing Zou and Eng Siong Chng
12:15-12:30 (Pier 4&5)
Audio-visual speech recognition (AVSR) attracts a surge of research interest recently by leveraging multimodal signals to understand human speech. Mainstream approaches addressing this task have developed sophisticated architectures and techniques for multi-modality fusion and representation learning. However, the natural heterogeneity of different modalities causes distribution gap between their representations, making it challenging to fuse them. In this paper, we aim to learn the shared representations across modalities to bridge their gap. Different from existing similar methods on other multimodal tasks like sentiment analysis, we focus on the temporal contextual dependencies considering the sequence-to-sequence task setting of AVSR. In particular, we propose an adversarial network to refine frame-level modality-invariant representations (MIR-GAN), which captures the commonality across modalities to ease the subsequent multimodal fusion process. Extensive experiments on public benchmarks LRS3 and LRS2 show that our approach outperforms the state-of-the-arts.

Virtual Poster

11:00-12:30 (Pier 7&8)

[CL] Certified Robustness to Text Adversarial Attacks by Randomized [MASK]
Jiehang Zeng, Jianhan Xu, Xiaoqing Zheng and Xuanjing Huang 11:00-12:30 (Pier 7&8)
Very recently, few certified defense methods have been developed to provably guarantee the robustness of a text classifier to adversarial syn-onym substitutions. However, all the existing certified defense methods assume that the defenders have been informed of how the adversaries generate synonyms, which is not a realistic scenario. In this study, we propose a certifiably robust defense method by randomly masking a certain proportion of the words in an input text, in which the above unrealistic assumption is no longer necessary. The proposed method can defend against not only word substitution-based attacks, but also character-level perturbations. We can certify the classifications of over 50

[TACL] MENLI: Robust Evaluation Metrics from Natural Language Inference
Yanran Chen and Steffen Eger 11:00-12:30 (Pier 7&8)
Recently proposed BERT-based evaluation metrics perform well on standard benchmarks but are vulnerable to adversarial attacks, e.g., re-lating to information correctness. We argue that this stems (in part) from the fact that they are models of semantic similarity. In contrast, we develop evaluation metrics based on *Natural Language Inference* (NLI), which we deem a more appropriate modeling. We design a preference-based adversarial attack framework and show that our NLI based metrics are much more robust to the attacks than the recent BERT-based metrics. On standard benchmarks, our NLI based metrics outperform existing summarization metrics, but perform below SOTA MT metrics. However, when combining existing metrics with our NLI metrics, we obtain both higher adversarial robustness (15%-30%) and higher quality metrics as measured on standard benchmarks (+5% to 30%).

[SRW] Semantic-Aware Dynamic Retrospective-Prospective Reasoning for Event-Level Video Question Answering
Chenyang Lyu, Tianbo Ji, Yvette Graham and Jennifer Foster 11:00-12:30 (Pier 7&8)
The paper is about video QA. It utilizes SRL partitioning to improve multi-step attention and reasoning of the models to attend to different frames for different parts of the question.

[SRW] I already said that! Degenerating redundant questions in open-domain dialogue systems.
Long Mai and Julie Carson-berndsen 11:00-12:30 (Pier 7&8)
This paper propose methods to reduce the number of redundant questions generated by open-domain dialogue systems.

[SRW] Data Selection for Fine-tuning Large Language Models Using Transferred Shapley Values
Stephanie Schoch, Yangfeng Ji and Ritwick Mishra 11:00-12:30 (Pier 7&8)
This paper proposes a sampling chain based method to make Shapley values computationally feasible for data valuation and selection for large language models.

[Demo] PersLEARN: Research Training through the Lens of Perspective Cultivation
Yixin Zhu, Wenjuan Han, Yuxi Ma, Lecheng Ruan, Jinhao Ji, Lin Qiu, Sijia Liu, Zhen Li, Yidong Lyu, Zijian Zhao, Xinyu Zhao, Xinyang Li, Bingru He, Shiyu Gu, Yifan Xu, Jiawen Liu, Qiao Xu, Xinyi Niu, Shiqian Li and Yu-Zhe Shi 11:00-12:30 (Pier 7&8)
Scientific research is inherently shaped by its authors' perspectives, influenced by various factors such as their personality, community, or society. Junior researchers often face challenges in identifying the perspectives reflected in the existing literature and struggle to develop their own viewpoints. In response to this issue, we introduce PersLEARN , a tool designed to facilitate the cultivation of scientific perspectives, starting from a basic seed idea and progressing to a well-articulated framework. By interacting with a prompt-based model, researchers can develop their perspectives explicitly. Our human study reveals that scientific perspectives developed by students using PersLEARN exhibit a superior level of logical coherence and depth compared to those that did not. Furthermore, our pipeline outperforms baseline approaches across multiple domains of literature from various perspectives. These results suggest that PersLEARN could help foster a greater apprecia-tion of diversity in scientific perspectives as an essential component of research training.

[Demo] SanskritShala: A Neural Sanskrit NLP Toolkit with Web-Based Interface for Pedagogical and Annotation Purposes
Pawan Goyal, Tushar Sandhan, Laxmidhar Behera, Anshul Agarwal and Jivnesh Sandhan 11:00-12:30 (Pier 7&8)
We present a neural Sanskrit Natural Language Processing (NLP) toolkit named SanskritShala (a school of Sanskrit) to facilitate compu-tational linguistic analyses for several tasks such as word segmentation, morphological tagging, dependency parsing, and compound type identification. Our systems currently report state-of-the-art performance on available benchmark datasets for all tasks. SanskritShala is de-ployed as a web-based application, which allows a user to get real-time analysis for the given input. It is built with easy-to-use interactive data annotation features that allow annotators to correct the system predictions when it makes mistakes. We publicly release the source codes of the 4 modules included in the toolkit, 7 word embedding models that have been trained on publicly available Sanskrit corpora and multiple annotated datasets such as word similarity, relatedness, categorization, analogy prediction to assess intrinsic properties of word embeddings. So far as we know, this is the first neural-based Sanskrit NLP toolkit that has a web-based interface and a number of NLP modules. We are sure that the people who are willing to work with Sanskrit will find it useful for pedagogical and annotative purposes. SanskritShala is available at: https://cnerg.iitkgp.ac.in/sanskritshala. The demo video of our platform can be accessed at: https://youtu.be/x0X31Y9k0mw4.

[Demo] Alfred: A System for Prompted Weak Supervision
Stephen Bach and Peilin Yu 11:00-12:30 (Pier 7&8)
Alfred is the first system for programmatic weak supervision (PWS) that creates training data for machine learning by prompting. In contrast to typical PWS systems where weak supervision sources are programs coded by experts, Alfred enables users to encode their subject matter expertise via natural language prompts for language and vision-language models. Alfred provides a simple Python interface for the key steps of this emerging paradigm, with a high-throughput backend for large-scale data labeling. Users can quickly create, evaluate, and refine their prompt-based weak supervision sources; map the results to weak labels; and resolve their disagreements with a label model. Alfred enables a seamless local development experience backed by models served from self-managed computing clusters. It automatically optimizes the execution of prompts with optimized batching mechanisms. We find that this optimization improves query throughput by 2.9x versus a naive approach. We present two example use cases demonstrating Alfred on YouTube comment spam detection and pet breeds classification. Alfred is open source, available at https://github.com/BatsResearch/alfred.

NusaCrowd: Open Source Initiative for Indonesian NLP Resources
Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Indra Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Halim Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Akbar Septiandri, James Jaya, Kaus-tubh Dhole, Arie Suryani, Rifki Afina Putri, Dan Su, Keith David Stevens, Made Nindyatama Nityasya, Muhammad Farid Adilazuarda, Ryan Ignatius Hadiwijaya, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Inastra Damapuspita, Haryo Akbarianto Wibowo, Cuk Tho, Ichwanul Muslim Karo Karo, Tirana Noor Fatyanosa, Ziwei Ji, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Pascale Fung, Herry Sujaini, Sakriani Sakti and Ayu Purwarianti 11:00-12:30 (Pier 7&8)
We present NusaCrowd, a collaborative initiative to collect and unify existing resources for Indonesian languages, including opening access to previously non-public resources. Through this initiative, we have brought together 137 datasets and 118 standardized data loaders. The quality of the datasets has been assessed manually and automatically, and their value is demonstrated through multiple experiments. NusaCrowd's

data collection enables the creation of the first zero-shot benchmarks for natural language understanding and generation in Indonesian and the local languages of Indonesia. Furthermore, NusaCrowd brings the creation of the first multilingual automatic speech recognition benchmark in Indonesian and the local languages of Indonesia. Our work strives to advance natural language processing (NLP) research for languages that are under-represented despite being widely spoken.

Can Language Models Make Fun? A Case Study in Chinese Comical Crosstalk

jianquan li, XiangBo Wu, Xiaokang Liu, Qianqian Xie, Prayag Tiwari and Benyou Wang   11:00-12:30 (Pier 7&8)

Language is the principal tool for human communication, in which humor is one of the most attractive parts. Producing natural language like humans using computers, a.k.a. Natural Language Generation (NLG), has been widely used for dialogue systems, chatbots, machine translation, as well as computer-aid creation e.g., idea generations, scriptwriting. However, the humor aspect of natural language is relatively under-investigated, especially in the age of pre-trained language models. In this work, we aim to preliminarily test *whether NLG can generate humor as humans do*. We build a largest dataset consisting of numerous **C**hinese **C**omical **C**rosstalk scripts (called **C**^{3} in short), which is for a popular Chinese performing art called 'Xiangsheng' since 1800s. We benchmark various generation approaches including training-from-scratch Seq2seq, fine-tuned middle-scale PLMs, and large-scale PLMs (with and without fine-tuning). Moreover, we also conduct a human assessment, showing that 1) *large-scale pretraining largely improves crosstalk generation quality*; and 2) *even the scripts generated from the best PLM is far from what we expect*. We conclude humor generation could be largely improved using large-scaled PLMs, but it is still in its infancy. The data and benchmarking code are publicly available in https://github.com/anonNo2/crosstalk-generation.

Revisiting Sample Size Determination in Natural Language Understanding

Ernie Chang, Muhammad Hassan Rashid, Pin-Jie Lin, Changsheng Zhao, Vera Demberg, Yangyang Shi and Vikas Chandra   11:00-12:30 (Pier 7&8)

Knowing exactly how many data points need to be labeled to achieve a certain model performance is a hugely beneficial step towards reducing the overall budgets for annotation. It pertains to both active learning and traditional data annotation, and is particularly beneficial for low resource scenarios. Nevertheless, it remains a largely under-explored area of research in NLP. We therefore explored various techniques for estimating the training sample size necessary to achieve a targeted performance value. We derived a simple yet effective approach to predict the maximum achievable model performance based on small amount of training samples – which serves as an early indicator during data annotation for data quality and sample size determination. We performed ablation studies on four language understanding tasks, and showed that the proposed approach allows us to forecast model performance within a small margin of mean absolute error (0.9%) with only 10% data.

MedNgage: A Dataset for Understanding Engagement in Patient-Nurse Conversations

Yan Wang, Heidi A.S. Donovan, Sabit Hassan and Malihe Alikhani   11:00-12:30 (Pier 7&8)

Patients who effectively manage their symptoms often demonstrate higher levels of engagement in conversations and interventions with healthcare practitioners. This engagement is multifaceted, encompassing cognitive and social dimensions. Consequently, it is crucial for AI systems to understand the engagement in natural conversations between patients and practitioners to better contribute toward patient care. In this paper, we present a novel dataset (MedNgage), which consists of patient-nurse conversations about cancer symptom management. We manually annotate the dataset with a novel framework of categories of patient engagement from two different angles, namely: i) socio-affective engagement (3.1K spans), and ii) cognitive engagement (1.8K spans). Through statistical analysis of the data that is annotated using our framework, we show a positive correlation between patient symptom management outcomes and their engagement in conversations. Additionally, we demonstrate that pre-trained transformer models fine-tuned on our dataset can reliably predict engagement categories in patient-nurse conversations. Lastly, we use LIME (Ribeiro et al., 2016) to analyze the underlying challenges of the tasks that state-of-the-art transformer models encounter. The de-identified data is available for research purposes upon request.

PLUE: Language Understanding Evaluation Benchmark for Privacy Policies in English

Jianfeng Chi, Wasi Uddin Ahmad, Yuan Tian and Kai-Wei Chang   11:00-12:30 (Pier 7&8)

Privacy policies provide individuals with information about their rights and how their personal information is handled. Natural language understanding (NLU) technologies can support individuals and practitioners to understand better privacy practices described in lengthy and complex documents. However, existing efforts that use NLU technologies are limited by processing the language in a way exclusive to a single task focusing on certain privacy practices. To this end, we introduce the Privacy Policy Language Understanding Evaluation (PLUE) benchmark, a multi-task benchmark for evaluating the privacy policy language understanding across various tasks. We also collect a large corpus of privacy policies to enable privacy policy domain-specific language model pre-training. We evaluate several generic pre-trained language models and continue pre-training them on the collected corpus. We demonstrate that domain-specific continual pre-training offers performance improvements across all tasks. The code and models are released at https://github.com/JFChi/PLUE.

The State of Profanity Obfuscation in Natural Language Processing Scientific Publications

Debora Nozza and Dirk Hovy   11:00-12:30 (Pier 7&8)

Work on hate speech has made considering rude and harmful examples in scientific publications inevitable. This situation raises various problems, such as whether or not to obscure profanities. While science must accurately disclose what it does, the unwarranted spread of hate speech can harm readers and increases its internet frequency. While maintaining publications' professional appearance, obfuscating profanities makes it challenging to evaluate the content, especially for non-native speakers. Surveying 150 ACL papers, we discovered that obfuscation is usually used for English but not other languages, and even then, quite unevenly. We discuss the problems with obfuscation and suggest a multilingual community resource called PrOf with a Python module to standardize profanity obfuscation processes. We believe PrOf can help scientific publication policies to make hate speech work accessible and comparable, irrespective of language.

A Diverse Set of Freely Available Linguistic Resources for Turkish

Duygu ALTINOK   11:00-12:30 (Pier 7&8)

This study presents a diverse set of freely available linguistic resources for Turkish natural language processing, including corpora, pretrained models and education material. Although Turkish is spoken by a sizeable population of over 80 million people, Turkish linguistic resources for natural language processing remain scarce. In this study, we provide corpora to allow practitioners to build their own applications and pretrained models that would assist industry researchers in creating quick prototypes. The provided corpora include named entity recognition datasets of diverse genres, including Wikipedia articles and supplement products customer reviews. In addition, crawling e-commerce and movie reviews websites, we compiled several sentiment analysis datasets of different genres. Our linguistic resources for Turkish also include pretrained spaCy language models. To the best of our knowledge, our models are the first spaCy models trained for the Turkish language. Finally, we provide various types of education material, such as video tutorials and code examples, that can support the interested audience on practicing Turkish NLP. The advantages of our linguistic resources are three-fold: they are freely available, they are first of their kind, and they are easy to use in a broad range of implementations. Along with a thorough description of the resource creation process, we also explain the position of our resources in the Turkish NLP world.

A Unified Evaluation Framework for Novelty Detection and Accommodation in NLP with an Instantiation in Authorship Attribution

Neeraj Varshney, Himanshu Gupta, Eric Robertson, Bing Liu and Chitta Baral   11:00-12:30 (Pier 7&8)

State-of-the-art natural language processing models have been shown to achieve remarkable performance in 'closed-world' settings where all the labels in the evaluation set are known at training time. However, in real-world settings, 'novel' instances that do not belong to any known class are often observed. This renders the ability to deal with novelties crucial. To initiate a systematic research in this important area of 'dealing with novelties', we introduce NoveltyTask, a multi-stage task to evaluate a system's performance on pipelined novelty 'detection' and 'accommodation' tasks. We provide mathematical formulation of NoveltyTask and instantiate it with the authorship attribution task that pertains to identifying the correct author of a given text. We use amazon reviews corpus and compile a large dataset (consisting of 250k instances across 200 authors/labels) for NoveltyTask. We conduct comprehensive experiments and explore several baseline methods for the task. Our results show that the methods achieve considerably low performance making the task challenging and leaving sufficient room for improvement. Finally, we believe our work will encourage research in this underexplored area of dealing with novelties, an important step en route to developing robust systems.

LEDA: a Large-Organization Email-Based Decision-Dialogue-Act Analysis Dataset

Mladen Karan, Prashant Khare, Ravi Shekhar, Stephen McQuistin, Ignacio Castro, Gareth Tyson, Colin Perkins, Patrick G.T. Healey and Matthew Purver   11:00-12:30 (Pier 7&8)

Collaboration increasingly happens online. This is especially true for large groups working on global tasks, with collaborators all around the globe. The size and distributed nature of such groups makes decision-making challenging. This paper proposes a set of dialog acts for the study of decision-making mechanisms in such groups, and provides a new annotated dataset based on real-world data from the public mail-archives of one such organisation – the Internet Engineering Task Force (IETF). We provide an initial data analysis showing that this dataset can be used to better understand decision-making in such organisations. Finally, we experiment with a preliminary transformer-based dialog act tagging model.

ScoNe: Benchmarking Negation Reasoning in Language Models With Fine-Tuning and In-Context Learning

Jingyuan S. She, Christopher Potts, Samuel R. Bowman and Atticus Geiger   11:00-12:30 (Pier 7&8)

A number of recent benchmarks seek to assess how well models handle natural language negation. However, these benchmarks lack the controlled example paradigms that would allow us to infer whether a model had truly learned how negation morphemes semantically scope. To fill these analytical gaps, we present the Scoped Negation NLI (ScoNe-NLI) benchmark, which contains contrast sets of six examples with up to two negations where either zero, one, or both negative morphemes affect the NLI label. We use ScoNe-NLI to assess fine-tuning and in-context learning strategies. We find that RoBERTa and DeBERTa models solve ScoNe-NLI after many shot fine-tuning. For in-context learning, we test the latest InstructGPT models and find that most prompt strategies are not successful, including those using step-by-step reasoning. To better understand this result, we extend ScoNe with ScoNe-NLG, a sentence completion test set that embeds negation reasoning in short narratives. Here, InstructGPT is successful, which reveals the model can correctly reason about negation, but struggles to do so on NLI examples outside of its core pretraining regime.

Take a Break in the Middle: Investigating Subgoals towards Hierarchical Script Generation

XINZE LI, Yixin Cao, Muhao Chen and Aixin Sun   11:00-12:30 (Pier 7&8)

Goal-oriented Script Generation is a new task of generating a list of steps that can fulfill the given goal. In this paper, we propose to extend the task from the perspective of cognitive theory. Instead of a simple flat structure, the steps are typically organized hierarchically — Human often decompose a complex task into subgoals, where each subgoal can be further decomposed into steps. To establish the benchmark, we contribute a new dataset, propose several baseline methods, and set up evaluation metrics. Both automatic and human evaluation verify the high-quality of dataset, as well as the effectiveness of incorporating subgoals into hierarchical script generation. Furthermore, We also design and evaluate the model to discover subgoal, and find that it is a bit more difficult to decompose the goals than summarizing from segmented steps.

InfoMetIC: An Informative Metric for Reference-free Image Caption Evaluation

Anwen Hu, Shizhe Chen, Liang Zhang and Qin Jin   11:00-12:30 (Pier 7&8)

Automatic image captioning evaluation is critical for benchmarking and promoting advances in image captioning research. Existing metrics only provide a single score to measure caption qualities, which are less explainable and informative. Instead, we humans can easily identify the problems of captions in details, e.g., which words are inaccurate and which salient objects are not described, and then rate the caption quality. To support such informative feedback, we propose an Informative Metric for Reference-free Image Caption evaluation (InfoMetIC). Given an image and a caption, InfoMetIC is able to report incorrect words and unmentioned image regions at fine-grained level, and also provide a text precision score, a vision recall score and an overall quality score at coarse-grained level. The coarse-grained score of InfoMetIC achieves significantly better correlation with human judgements than existing metrics on multiple benchmarks. We also construct a token-level evaluation dataset and demonstrate the effectiveness of InfoMetIC in fine-grained evaluation. Our code and datasets are publicly available at https://github.com/HAWLYQ/InfoMetIC.

An Inclusive Notion of Text

Ilia Kuznetsov and Iryna Gurevych   11:00-12:30 (Pier 7&8)

Natural language processing (NLP) researchers develop models of grammar, meaning and communication based on written text. Due to task and data differences, what is considered text can vary substantially across studies. A conceptual framework for systematically capturing these differences is lacking. We argue that clarity on the notion of text is crucial for reproducible and generalizable NLP. Towards that goal, we propose common terminology to discuss the production and transformation of textual data, and introduce a two-tier taxonomy of linguistic and non-linguistic elements that are available in textual sources and can be used in NLP modeling. We apply this taxonomy to survey existing work that extends the notion of text beyond the conservative language-centered view. We outline key desiderata and challenges of the emerging inclusive approach to text in NLP, and suggest community-level reporting as a crucial next step to consolidate the discussion.

Few-shot Adaptation Works with UnpredicTable Data

Jun Shern Chan, Michael Pieler, Jonathan Jao, Jérémy Scheurer and Ethan Perez   11:00-12:30 (Pier 7&8)

Prior work on language models (LMs) shows that training on a large number of diverse tasks improves few-shot learning (FSL) performance on new tasks. We take this to the extreme, automatically extracting 413,299 tasks from internet tables - orders of magnitude more than the next-largest public datasets. Finetuning on the resulting dataset leads to improved FSL performance on Natural Language Processing (NLP) tasks, but not proportionally to dataset scale. In fact, we find that narrow subsets of our dataset sometimes outperform more diverse datasets. For example, finetuning on software documentation from support.google.com raises FSL performance by a mean of +7.5% on 52 downstream tasks, which beats training on 40 human-curated NLP datasets (+6.7%). Finetuning on various narrow datasets leads to similar broad improvements across test tasks, suggesting that the gains are not from domain adaptation but adapting to FSL in general. We do not observe clear patterns between the datasets that lead to FSL gains, leaving open questions about why certain data helps with FSL.

NewsMet : A 'do it all' Dataset of Contemporary Metaphors in News Headlines

Rohan Joseph, Timothy Liu, Aik Beng Ng, Simon See and Sunny Rai   11:00-12:30 (Pier 7&8)

Metaphors are highly creative constructs of human language that grow old and eventually die. Popular datasets used for metaphor processing tasks were constructed from dated source texts. In this paper, we propose NewsMet, a large high-quality contemporary dataset of news headlines hand-annotated with metaphorical verbs. The dataset comprises headlines from various sources including political, satirical, reliable and fake. Our dataset serves the purpose of evaluation for the tasks of metaphor interpretation and generation. The experiments reveal several insights and limitations of using LLMs to automate metaphor processing tasks as frequently seen in the recent literature. The dataset is publicly available for research purposeshttps://github.com/AxleBlaze3/NewsMet_Metaphor_Dataset.

K-UniMorph: Korean Universal Morphology and its Feature Schema
Eunkyul Leah Jo, Kim Kyuwon, Xihan Wu, KyungTae Lim, Jungyeul Park and Chulwoo Park   11:00-12:30 (Pier 7&8)
We present in this work a new Universal Morphology dataset for Korean. Previously, the Korean language has been underrepresented in the field of morphological paradigms amongst hundreds of diverse world languages. Hence, we propose this Universal Morphological paradigms for the Korean language that preserve its distinct characteristics. For our K-UniMorph dataset, we outline each grammatical criterion in detail for the verbal endings, clarify how to extract inflected forms, and demonstrate how we generate the morphological schemata. This dataset adopts morphological feature schema from CITATION and CITATION for the Korean language as we extract inflected verb forms from the Sejong morphologically analyzed corpus that is one of the largest annotated corpora for Korean. During the data creation, our methodology also includes investigating the correctness of the conversion from the Sejong corpus. Furthermore, we carry out the inflection task using three different Korean word forms: letters, syllables and morphemes. Finally, we discuss and describe future perspectives on Korean morphological paradigms and the dataset.

A New Task and Dataset on Detecting Attacks on Human Rights Defenders
Shihao Ran, Di Lu, Aoife Cahill, Joel Tetreault and Alejandro Jaimes   11:00-12:30 (Pier 7&8)
The ability to conduct retrospective analyses of attacks on human rights defenders over time and by location is important for humanitarian organizations to better understand historical or ongoing human rights violations and thus better manage the global impact of such events. We hypothesize that NLP can support such efforts by quickly processing large collections of news articles to detect and summarize the characteristics of attacks on human rights defenders. To that end, we propose a new dataset for detecting Attacks on Human Rights Defenders (HRDsAttack) consisting of crowdsourced annotations on 500 online news articles. The annotations include fine-grained information about the type and location of the attacks, as well as information about the victim(s). We demonstrate the usefulness of the dataset by using it to train and evaluate baseline models on several sub-tasks to predict the annotated characteristics.

Comparative evaluation of boundary-relaxed annotation for Entity Linking performance
Gabriel Herman Bernardim Andrade, Shuntaro Yada and Eiji ARAMAKI   11:00-12:30 (Pier 7&8)
Entity Linking performance has a strong reliance on having a large quantity of high-quality annotated training data available. Yet, manual annotation of named entities, especially their boundaries, is ambiguous, error-prone, and raises many inconsistencies between annotators. While imprecise boundary annotation can degrade a model's performance, there are applications where accurate extraction of entities' surface form is not necessary. For those cases, a lenient annotation guideline could relieve the annotators' workload and speed up the process. This paper presents a case study designed to verify the feasibility of such annotation process and evaluate the impact of boundary-relaxed annotation in an Entity Linking pipeline. We first generate a set of noisy versions of the widely used AIDA CoNLL-YAGO dataset by expanding the boundaries subsets of annotated entity mentions and then train three Entity Linking models on this data and evaluate the relative impact of imprecise annotation on entity recognition and disambiguation performances. We demonstrate that the magnitude of effects caused by noise in the Named Entity Recognition phase is dependent on both model complexity and noise ratio, while Entity Disambiguation components are susceptible to entity boundary imprecision due to strong vocabulary dependency.

HaVQA: A Dataset for Visual Question Answering and Multimodal Research in Hausa Language
Shantipriya Parida, Idris Abdulmumin, Shamsuddeen Hassan Muhammad, Aneesh Bose, Guneet Singh Kohli, Ibrahim Said Ahmad, Ketan Kotwal, Sayan Deb Sarkar, Ondřej Bojar and Habeebah Adamu Kakudi   11:00-12:30 (Pier 7&8)
This paper presents "HaVQA", the first multimodal dataset for visual question answering (VQA) tasks in the Hausa language. The dataset was created by manually translating 6,022 English question-answer pairs, which are associated with 1,555 unique images from the Visual Genome dataset. As a result, the dataset provides 12,044 gold standard English-Hausa parallel sentences that were translated in a fashion that guarantees their semantic match with the corresponding visual information. We conducted several baseline experiments on the dataset, including visual question answering, visual question elicitation, text-only and multimodal machine translation.

Recurrent Attention Networks for Long-text Modeling
Xianming Li, Zongxi Li, Xiaotian Luo, Haoran Xie, Xing Lee, Yingbin Zhao, Fu Lee Wang and Qing Li   11:00-12:30 (Pier 7&8)
Self-attention-based models have achieved remarkable progress in short-text mining. However, the quadratic computational complexities restrict their application in long text processing. Prior works have adopted the chunking strategy to divide long documents into chunks and stack a self-attention backbone with the recurrent structure to extract semantic representation. Such an approach disables parallelization of the attention mechanism, significantly increasing the training cost and raising hardware requirements. Revisiting the self-attention mechanism and the recurrent structure, this paper proposes a novel long-document encoding model, Recurrent Attention Network (RAN), to enable the recurrent operation of self-attention. Combining the advantages from both sides, the well-designed RAN is capable of extracting global semantics in both token-level and document-level representations, making it inherently compatible with both sequential and classification tasks, respectively. Furthermore, RAN is computationally scalable as it supports parallelization on long document processing. Extensive experiments demonstrate the long-text encoding ability of the proposed RAN model on both classification and sequential tasks, showing its potential for a wide range of applications.

DynaMiTE: Discovering Explosive Topic Evolutions with User Guidance
Nishant Balepur, Shivam Agarwal, Karthik Venkat Ramanan, Susik Yoon, Diyi Yang and Jiawei Han   11:00-12:30 (Pier 7&8)
Dynamic topic models (DTMs) analyze text streams to capture the evolution of topics. Despite their popularity, existing DTMs are either fully supervised, requiring expensive human annotations, or fully unsupervised, producing topic evolutions that often do not cater to a user's needs. Further, the topic evolutions produced by DTMs tend to contain generic terms that are not indicative of their designated time steps. To address these issues, we propose the task of discriminative dynamic topic discovery. This task aims to discover topic evolutions from temporal corpora that distinctly align with a set of user-provided category names and uniquely capture topics at each time step. We solve this task by developing DynaMiTE, a framework that ensembles semantic similarity, category indicative, and time indicative scores to produce informative topic evolutions. Through experiments on three diverse datasets, including the use of a newly-designed human evaluation experiment, we demonstrate that DynaMiTE is a practical and efficient framework for helping users discover high-quality topic evolutions suited to their interests.

Large Language Models are Built-in Autoregressive Search Engines
Noah Ziems, Wenhao Yu, Zhihan Zhang and Meng Jiang   11:00-12:30 (Pier 7&8)
Document retrieval is a key stage of standard Web search engines. Existing dual-encoder dense retrievers obtain representations for questions

and documents independently, allowing for only shallow interactions between them. To overcome this limitation, recent autoregressive search engines replace the dual-encoder architecture by directly generating identifiers for relevant documents in the candidate pool. However, the training cost of such autoregressive search engines rises sharply as the number of candidate documents increases. In this paper, we find that large language models (LLMs) can follow human instructions to directly generate URLs for document retrieval.

Surprisingly, when providing a few Query-URL pairs as in-context demonstrations, LLMs can generate Web URLs where nearly 90% of the corresponding documents contain correct answers to open-domain questions. In this way, LLMs can be thought of as built-in search engines, since they have not been explicitly trained to map questions to document identifiers. Experiments demonstrate that our method can consistently achieve better retrieval performance than existing retrieval approaches by a significant margin on three open-domain question answering benchmarks, under both zero and few-shot settings. The code for this work can be found at https://github.com/Ziems/llm-url.

Nonparametric Decoding for Generative Retrieval
Hyunji Lee, JaeYoung Kim, Hoyeon Chang, hanseok Oh, Sohee Yang, Vladimir Karpukhin, Yi Lu and Minjoon Seo   11:00-12:30 (Pier 7&8)
The generative retrieval model depends solely on the information encoded in its model parameters without external memory, its information capacity is limited and fixed. To overcome the limitation, we propose Nonparametric Decoding (Np Decoding) which can be applied to existing generative retrieval models. Np Decoding uses nonparametric contextualized vocab embeddings (external memory) rather than vanilla vocab embeddings as decoder vocab embeddings. By leveraging the contextualized vocab embeddings, the generative retrieval model is able to utilize both the parametric and nonparametric space. Evaluation over 9 datasets (8 single-hop and 1 multi-hop) in the document retrieval task shows that applying Np Decoding to generative retrieval models significantly improves the performance. We also show that Np Decoding is data- and parameter-efficient, and shows high performance in the zero-shot setting.

SamToNe: Improving Contrastive Loss for Dual Encoder Retrieval Models with Same Tower Negatives
Fedor Moiseev, Gustavo Hernandez Abrego, Peter Dornbach, Imed Zitouni, Enrique Alfonseca and Zhe Dong   11:00-12:30 (Pier 7&8)
Dual encoders have been used for retrieval tasks and representation learning with good results. A standard way to train dual encoders is using a contrastive loss with in-batch negatives. In this work, we propose an improved contrastive learning objective by adding queries or documents from the same encoder towers to the negatives, for which we name it as "contrastive loss with SAMe TOwer NEgatives" (SamToNe). By evaluating on question answering retrieval benchmarks from MS MARCO and MultiReQA, and heterogenous zero-shot information retrieval benchmarks (BEIR), we demonstrate that SamToNe can effectively improve the retrieval quality for both symmetric and asymmetric dual encoders. By directly probing the embedding spaces of the two encoding towers via the t-SNE algorithm (van der Maaten and Hinton, 2008), we observe that SamToNe ensures the alignment between the embedding spaces from the two encoder towers. Based on the analysis of the embedding distance distributions of the top-1 retrieved results, we further explain the efficacy of the method from the perspective of regularisation.

SimLM: Pre-training with Representation Bottleneck for Dense Passage Retrieval
Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder and Furu Wei 11:00-12:30 (Pier 7&8)
In this paper, we propose SimLM (Similarity matching with Language Model pre-training), a simple yet effective pre-training method for dense passage retrieval. It employs a simple bottleneck architecture that learns to compress the passage information into a dense vector through self-supervised pre-training. We use a replaced language modeling objective, which is inspired by ELECTRA (Clark et al., 2020), to improve the sample efficiency and reduce the mismatch of the input distribution between pre-training and fine-tuning. SimLM only requires access to an unlabeled corpus and is more broadly applicable when there are no labeled data or queries. We conduct experiments on several large-scale passage retrieval datasets and show substantial improvements over strong baselines under various settings. Remarkably, SimLM even outperforms multi-vector approaches such as ColBERTv2 (Santhanam et al., 2021) which incurs significantly more storage cost. Our code and model check points are available at https://github.com/microsoft/unilm/tree/master/simlm .

Robust Representation Learning with Reliable Pseudo-labels Generation via Self-Adaptive Optimal Transport for Short Text Clustering
Xiaolin Zheng, Mengling Hu, Weiming Liu, Chaochao Chen and Xinting Liao   11:00-12:30 (Pier 7&8)
Short text clustering is challenging since it takes imbalanced and noisy data as inputs. Existing approaches cannot solve this problem well, since (1) they are prone to obtain degenerate solutions especially on heavy imbalanced datasets, and (2) they are vulnerable to noises. To tackle the above issues, we propose a Robust Short Text Clustering (RSTC) model to improve robustness against imbalanced and noisy data. RSTC includes two modules, i.e., pseudo-label generation module and robust representation learning module. The former generates pseudo-labels to provide supervision for the later, which contributes to more robust representations and correctly separated clusters. To provide robustness against the imbalance in data, we propose self-adaptive optimal transport in the pseudo-label generation module. To improve robustness against the noise in data, we further introduce both class-wise and instance-wise contrastive learning in the robust representation learning module. Our empirical studies on eight short text clustering datasets demonstrate that RSTC significantly outperforms the state-of-the-art models.

Multiview Identifiers Enhanced Generative Retrieval
Yongqi Li, Nan Yang, Liang Wang, Furu Wei and Wenjie Li   11:00-12:30 (Pier 7&8)
Instead of simply matching a query to pre-existing passages, generative retrieval generates identifier strings of passages as the retrieval target. At a cost, the identifier must be distinctive enough to represent a passage. Current approaches use either a numeric ID or a text piece (such as a title or substrings) as the identifier. However, these identifiers cannot cover a passage's content well. As such, we are motivated to propose a new type of identifier, synthetic identifiers, that are generated based on the content of a passage and could integrate contextualized information that text pieces lack. Furthermore, we simultaneously consider multiview identifiers, including synthetic identifiers, titles, and substrings. These views of identifiers complement each other and facilitate the holistic ranking of passages from multiple perspectives. We conduct a series of experiments on three public datasets, and the results indicate that our proposed approach performs the best in generative retrieval, demonstrating its effectiveness and robustness.

TOME: A Two-stage Approach for Model-based Retrieval
Ruiyang Ren, Wayne Xin Zhao, Jing Liu, Hua Wu, Ji-Rong Wen and Haifeng Wang   11:00-12:30 (Pier 7&8)
Recently, model-based retrieval has emerged as a new paradigm in text retrieval that discards the index in the traditional retrieval model and instead memorizes the candidate corpora using model parameters. This design employs a sequence-to-sequence paradigm to generate document identifiers, which enables the complete capture of the relevance between queries and documents and simplifies the classic index-retrieval-rerank pipeline. Despite its attractive qualities, there remain several major challenges in model-based retrieval, including the discrepancy between pre-training and fine-tuning, and the discrepancy between training and inference. To deal with the above challenges, we propose a novel two-stage model-based retrieval approach called TOME, which makes two major technical contributions, including the utilization of tokenized URLs as identifiers and the design of a two-stage generation architecture. We also propose a number of training strategies to deal with the training difficulty as the corpus size increases. Extensive experiments and analysis on MS MARCO and Natural Questions demonstrate the effectiveness of our proposed approach, and we investigate the scaling laws of TOME by examining various influencing factors.

MixPAVE: Mix-Prompt Tuning for Few-shot Product Attribute Value Extraction
Li Yang, Qifan Wang, Jingang Wang, Xiaojun Quan, Fuli Feng, Yu Chen, Madian Khabsa, Sinong Wang, Zenglin Xu and Dongfang Liu
11:00-12:30 (Pier 7&8)
The task of product attribute value extraction is to identify values of an attribute from product information. Product attributes are important features, which help improve online shopping experience of customers, such as product search, recommendation and comparison. Most existing works only focus on extracting values for a set of known attributes with sufficient training data. However, with the emerging nature of e-commerce, new products with their unique set of new attributes are constantly generated from different retailers and merchants. Collecting a large number of annotations for every new attribute is costly and time consuming. Therefore, it is an important research problem for product attribute value extraction with limited data. In this work, we propose a novel prompt tuning approach with **Mix**ed **P**rompts for few-shot **A**ttribute **V**alue **E**xtraction, namely MixPAVE. Specifically, MixPAVE introduces only a small amount (< 1%) of trainable parameters, i.e., a mixture of two learnable prompts, while keeping the existing extraction model frozen. In this way, MixPAVE not only benefits from parameter-efficient training, but also avoids model overfitting on limited training examples. Experimental results on two product benchmarks demonstrate the superior performance of the proposed approach over several state-of-the-art baselines. A comprehensive set of ablation studies validate the effectiveness of the prompt design, as well as the efficiency of our approach.

Disentangled Phonetic Representation for Chinese Spelling Correction
Zihong Liang, Xiaojun Quan and Qifan Wang
11:00-12:30 (Pier 7&8)
Chinese Spelling Correction (CSC) aims to detect and correct erroneous characters in Chinese texts. Although efforts have been made to introduce phonetic information (Hanyu Pinyin) in this task, they typically merge phonetic representations with character representations, which tends to weaken the representation effect of normal texts. In this work, we propose to disentangle the two types of features to allow for direct interaction between textual and phonetic information. To learn useful phonetic representations, we introduce a pinyin-to-character objective to ask the model to predict the correct characters based solely on phonetic information, where a separation mask is imposed to disable attention from phonetic input to text. To avoid overfitting the phonetics, we further design a self-distillation module to ensure that semantic information plays a major role in the prediction. Extensive experiments on three CSC benchmarks demonstrate the superiority of our method in using phonetic information.

Are You Copying My Model? Protecting the Copyright of Large Language Models for EaaS via Backdoor Watermark
Wenjun Peng, Jingwei Yi, Fangzhao Wu, Shangxi Wu, Bin Benjamin Bin Zhu, Lingjuan Lyu, Binxing Jiao, Tong Xu, Guangzhong Sun and Xing Xie
11:00-12:30 (Pier 7&8)
Large language models (LLMs) have demonstrated powerful capabilities in both text understanding and generation. Companies have begun to offer Embedding as a Service (EaaS) based on these LLMs, which can benefit various natural language processing (NLP) tasks for customers. However, previous studies have shown that EaaS is vulnerable to model extraction attacks, which can cause significant losses for the owners of LLMs, as training these models is extremely expensive. To protect the copyright of LLMs for EaaS, we propose an Embedding Watermark method called {pasted macro 'METHOD'} that implants backdoors on embeddings. Our method selects a group of moderate-frequency words from a general text corpus to form a trigger set, then selects a target embedding as the watermark, and inserts it into the embeddings of texts containing trigger words as the backdoor. The weight of insertion is proportional to the number of trigger words included in the text. This allows the watermark backdoor to be effectively transferred to EaaS-stealer's model for copyright verification while minimizing the adverse impact on the original embeddings' utility. Our extensive experiments on various datasets show that our method can effectively protect the copyright of EaaS models without compromising service quality. Our code is available at https://github.com/yjw1029/EmbMarker.

Better Language Models of Code through Self-Improvement
Hung Quoc To, Nghi D. Q. Bui, Jin L.C. Guo and Tien N Nguyen
11:00-12:30 (Pier 7&8)
Pre-trained language models for code (PLMCs) have gained attention in recent research. These models are pre-trained on large-scale datasets using multi-modal objectives. However, fine-tuning them requires extensive supervision and is limited by the size of the dataset provided. We aim to improve this issue by proposing a data augmentation framework using knowledge distillation. Our framework utilizes knowledge gained during the pre-training and fine-tuning stage to augment training data, which is then used for the next step. We incorporate this framework into the state-of-the-art language models, such as CodeT5, CodeBERT, and UnixCoder. The results show that our framework significantly improves PLMCs' performance in sequence-generation tasks, such as code summarization and code generation in the CodeXGLUE benchmark.

GEC-DePenD: Non-Autoregressive Grammatical Error Correction with Decoupled Permutation and Decoding
Konstantin Yakovlev, Alexander Podolskiy, Andrey Bout, Sergey I Nikolenko and Irina Piontkovskaya
11:00-12:30 (Pier 7&8)
Grammatical error correction (GEC) is an important NLP task that is currently usually solved with autoregressive sequence-to-sequence models. However, approaches of this class are inherently slow due to one-by-one token generation, so non-autoregressive alternatives are needed. In this work, we propose a novel non-autoregressive approach to GEC that decouples the architecture into a permutation network that outputs a self-attention weight matrix that can be used in beam search to find the best permutation of input tokens (with auxiliary <ins> tokens) and a decoder network based on a step-unrolled denoising autoencoder that fills in specific tokens. This allows us to find the token permutation after only one forward pass of the permutation network, avoiding autoregressive constructions. We show that the resulting network improves over previously known non-autoregressive methods for GEC and reaches the level of autoregressive methods that do not use language-specific synthetic data generation methods. Our results are supported by a comprehensive experimental validation on the ConLL-2014 and BEA datasets and an extensive ablation study that supports our architectural and algorithmic choices.

Zero-Shot Text Classification via Self-Supervised Tuning
Chaoqun Liu, Wenxuan Zhang, Guizhen Chen, Xiaobao Wu, Anh Tuan Luu, Chip Hong Chang and Lidong Bing
11:00-12:30 (Pier 7&8)
Existing solutions to zero-shot text classification either conduct prompting with pre-trained language models, which is sensitive to the choices of templates, or rely on large-scale annotated data of relevant tasks for meta-tuning. In this work, we propose a new paradigm based on self-supervised learning to solve zero-shot text classification tasks by tuning the language models with unlabeled data, called self-supervised tuning. By exploring the inherent structure of free texts, we propose a new learning objective called first sentence prediction to bridge the gap between unlabeled data and text classification tasks. After tuning the model to learn to predict the first sentence in a paragraph based on the rest, the model is able to conduct zero-shot inference on unseen tasks such as topic classification and sentiment analysis. Experimental results show that our model outperforms the state-of-the-art baselines on 7 out of 10 tasks. Moreover, the analysis reveals that our model is less sensitive to the prompt design. Our code and pre-trained models are publicly available at https://github.com/DAMO-NLP-SG/SSTuning

Causal Intervention and Counterfactual Reasoning for Multi-modal Fake News Detection
Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao and Liqiang Nie
11:00-12:30 (Pier 7&8)
Due to the rapid upgrade of social platforms, most of today's fake news is published and spread in a multi-modal form. Most existing multi-modal fake news detection methods neglect the fact that some label-specific features learned from the training set cannot generalize well to the testing set, thus inevitably suffering from the harm caused by the latent data bias. In this paper, we analyze and identify the psycholinguistic bias in the text and the bias of inferring news label based on only image features. We mitigate these biases from a causality perspective and

propose a Causal intervention and Counterfactual reasoning based Debiasing framework (CCD) for multi-modal fake news detection. To achieve our goal, we first utilize causal intervention to remove the psycholinguistic bias which introduces the spurious correlations between text features and news label. And then, we apply counterfactual reasoning by imagining a counterfactual world where each news has only image features for estimating the direct effect of the image. Therefore we can eliminate the image-only bias by deducting the direct effect of the image from the total effect on labels. Extensive experiments on two real-world benchmark datasets demonstrate the effectiveness of our framework for improving multi-modal fake news detection.

An Exploration of Encoder-Decoder Approaches to Multi-Label Classification for Legal and Biomedical Text
Yova Kementchedjhieva and Ilias Chalkidis   11:00-12:30 (Pier 7&8)
Standard methods for multi-label text classification largely rely on encoder-only pre-trained language models, whereas encoder-decoder models have proven more effective in other classification tasks. In this study, we compare four methods for multi-label classification, two based on an encoder only, and two based on an encoder-decoder. We carry out experiments on four datasets—two in the legal domain and two in the biomedical domain, each with two levels of label granularity— and always depart from the same pre-trained model, T5. Our results show that encoder-decoder methods outperform encoder-only methods, with a growing advantage on more complex datasets and labeling schemes of finer granularity. Using encoder-decoder models in a non-autoregressive fashion, in particular, yields the best performance overall, so we further study this approach through ablations to better understand its strengths.

Prompt- and Trait Relation-aware Cross-prompt Essay Trait Scoring
Heejin Do, Yunsu Kim and Gary Geunbae Lee   11:00-12:30 (Pier 7&8)
Automated essay scoring (AES) aims to score essays written for a given prompt, which defines the writing topic. Most existing AES systems assume to grade essays of the same prompt as used in training and assign only a holistic score. However, such settings conflict with real-education situations; pre-graded essays for a particular prompt are lacking, and detailed trait scores of sub-rubrics are required. Thus, predicting various trait scores of unseen-prompt essays (called cross-prompt essay trait scoring) is a remaining challenge of AES. In this paper, we propose a robust model: prompt- and trait relation-aware cross-prompt essay trait scorer. We encode prompt-aware essay representation by essay-prompt attention and utilizing the topic-coherence feature extracted by the topic-modeling mechanism without access to labeled data; therefore, our model considers the prompt adherence of an essay, even in a cross-prompt setting. To facilitate multi-trait scoring, we design trait-similarity loss that encapsulates the correlations of traits. Experiments prove the efficacy of our model, showing state-of-the-art results for all prompts and traits. Significant improvements in low-resource-prompt and inferior traits further indicate our model's strength.

Towards Identifying Fine-Grained Depression Symptoms from Memes
Shweta Yadav, Cornelia Caragea, Chenye Zhao, Naincy Kumari, Marvin A Solberg and Tanmay Sharma   11:00-12:30 (Pier 7&8)
The past decade has observed significant attention toward developing computational methods for classifying social media data based on the presence or absence of mental health conditions. In the context of mental health, for clinicians to make an accurate diagnosis or provide personalized intervention, it is crucial to identify fine-grained mental health symptoms. To this end, we conduct a focused study on depression disorder and introduce a new task of identifying fine-grained depressive symptoms from memes. Toward this, we create a high-quality dataset (RESTORE) annotated with 8 fine-grained depression symptoms based on the clinically adopted PHQ-9 questionnaire. We benchmark RESTORE on 20 strong monomodal and multimodal methods. Additionally, we show how imposing orthogonal constraints on textual and visual feature representations in a multimodal setting can enforce the model to learn non-redundant and de-correlated features leading to a better prediction of fine-grained depression symptoms. Further, we conduct an extensive human analysis and elaborate on the limitations of existing multimodal models that often overlook the implicit connection between visual and textual elements of a meme.

Are Pre-trained Language Models Useful for Model Ensemble in Chinese Grammatical Error Correction?
Chenming Tang, Xiuyu Wu and Yunfang Wu   11:00-12:30 (Pier 7&8)
Model ensemble has been in widespread use for Grammatical Error Correction (GEC), boosting model performance. We hypothesize that model ensemble based on the perplexity (PPL) computed by pre-trained language models (PLMs) should benefit the GEC system. To this end, we explore several ensemble strategies based on strong PLMs with four sophisticated single models. However, the performance does not improve but even gets worse after the PLM-based ensemble. This surprising result sets us doing a detailed analysis on the data and coming up with some insights on GEC. The human references of correct sentences is far from sufficient in the test data, and the gap between a correct sentence and an idiomatic one is worth our attention. Moreover, the PLM-based ensemble strategies provide an effective way to extend and improve GEC benchmark data. Our source code is available at https://github.com/JamyDon/PLM-based-CGEC-Model-Ensemble.

Improving Grammatical Error Correction with Multimodal Feature Integration
Tao Fang, jinpeng hu, Derek F. Wong, Xiang Wan, Lidia S. Chao and Tsung-Hui Chang   11:00-12:30 (Pier 7&8)
Grammatical error correction (GEC) is a promising task aimed at correcting errors in a text. Many methods have been proposed to facilitate this task with remarkable results. However, most of them only focus on enhancing textual feature extraction without exploring the usage of other modalities' information (e.g., speech), which can also provide valuable knowledge to help the model detect grammatical errors. To shore up this deficiency, we propose a novel framework that integrates both speech and text features to enhance GEC. In detail, we create new multimodal GEC datasets for English and German by generating audio from text using the advanced text-to-speech models. Subsequently, we extract acoustic and textual representations by a multimodal encoder that consists of a speech and a text encoder. A mixture-of-experts (MoE) layer is employed to selectively align representations from the two modalities, and then a dot attention mechanism is used to fuse them as final multimodal representations. Experimental results on CoNLL14, BEA19 English, and Falko-MERLIN German show that our multimodal GEC models achieve significant improvements over strong baselines and achieve a new state-of-the-art result on the Falko-MERLIN test set.

Contrastive Training Improves Zero-Shot Classification of Semi-structured Documents
Muhammad Khalifa, Yogarshi Vyas, Shuai Wang, Graham Horwood, Sunil Mallya and Miguel Ballesteros   11:00-12:30 (Pier 7&8)
We investigate semi-structured document classification in a zero-shot setting. Classification of semi-structured documents is more challenging than that of standard unstructured documents, as positional, layout, and style information play a vital role in interpreting such documents. The standard classification setting where categories are fixed during both training and testing falls short in dynamic environments where new classification categories could potentially emerge. We focus exclusively on the zero-shot learning setting where inference is done on new unseen classes. To address this task, we propose a matching-based approach that relies on a pairwise contrastive objective for both pretraining and fine-tuning. Our results show a significant boost in Macro F1 from the proposed pretraining step and comparable performance of the contrastive fine-tuning to a standard prediction objective in both supervised and unsupervised zero-shot settings.

Explainable Recommendation with Personalized Review Retrieval and Aspect Learning
Hao Cheng, Shuo Wang, Wensheng Lu, Wei Zhang, Mingyang Zhou, Kezhong Lu and Hao Liao   11:00-12:30 (Pier 7&8)
Explainable recommendation is a technique that combines prediction and generation tasks to produce more persuasive results. Among these tasks, textual generation demands large amounts of data to achieve satisfactory accuracy. However, historical user reviews of items are often insufficient, making it challenging to ensure the precision of generated explanation text. To address this issue, we propose a novel model, ERRA (Explainable Recommendation by personalized Review retrieval and Aspect learning). With retrieval enhancement, ERRA can obtain

additional information from the training sets. With this additional information, we can generate more accurate and informative explanations. Furthermore, to better capture users' preferences, we incorporate an aspect enhancement component into our model. By selecting the top-n aspects that users are most concerned about for different items, we can model user representation with more relevant details, making the explanation more persuasive. To verify the effectiveness of our model, extensive experiments on three datasets show that our model outperforms state-of-the-art baselines (for example, 3.4% improvement in prediction and 15.8% improvement in explanation for TripAdvisor).

Distractor Generation based on Text2Text Language Models with Pseudo Kullback-Leibler Divergence Regulation
Hui-Juan Wang, Kai-Yu Hsieh, Han-Cheng Yu, Jui-Ching Tsou, Yu An Shih, Chen-Hua Huang and Yao-Chung Fan   11:00-12:30 (Pier 7&8)
In this paper, we address the task of cloze-style multiple choice question (MCQs) distractor generation. Our study is featured by the following designs. First, we propose to formulate the cloze distractor generation as a Text2Text task. Second, we propose pseudo Kullback-Leibler Divergence for regulating the generation to consider the item discrimination index in education evaluation. Third, we explore the candidate augmentation strategy and multi-tasking training with cloze-related tasks to further boost the generation performance. Through experiments with benchmarking datasets, our best performing model advances the state-of-the-art result from 10.81 to 22.00 (p@1 score).

MolXPT: Wrapping Molecules with Text for Generative Pre-training
Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang and Tie-Yan Liu   11:00-12:30 (Pier 7&8)
Generative pre-trained Transformer (GPT) has demonstrates its great success in natural language processing and related techniques have been adapted into molecular modeling. Considering that text is the most important record for scientific discovery, in this paper, we propose MolXPT, a unified language model of text and molecules pre-trained on SMILES (a sequence representation of molecules) wrapped by text. Briefly, we detect the molecule names in each sequence and replace them to the corresponding SMILES. In this way, the SMILES could leverage the information from surrounding text, and vice versa. The above wrapped sequences, text sequences from PubMed and SMILES sequences from PubChem are all fed into a language model for pre-training. Experimental results demonstrate that MolXPT outperforms strong baselines of molecular property prediction on MoleculeNet, performs comparably to the best model in text-molecule translation while using less than half of its parameters, and enables zero-shot molecular generation without finetuning.

Scientific Fact-Checking: A Survey of Resources and Approaches
Juraj Vladika and Florian Matthes   11:00-12:30 (Pier 7&8)
The task of fact-checking deals with assessing the veracity of factual claims based on credible evidence and background knowledge. In particular, scientific fact-checking is the variation of the task concerned with verifying claims rooted in scientific knowledge. This task has received significant attention due to the growing importance of scientific and health discussions on online platforms. Automated scientific fact-checking methods based on NLP can help combat the spread of misinformation, assist researchers in knowledge discovery, and help individuals understand new scientific breakthroughs. In this paper, we present a comprehensive survey of existing research in this emerging field and its related tasks. We provide a task description, discuss the construction process of existing datasets, and analyze proposed models and approaches. Based on our findings, we identify intriguing challenges and outline potential future directions to advance the field.

HermEs: Interactive Spreadsheet Formula Prediction via Hierarchical Formulet Expansion
Wanrong He, Haoyu Dong, Yihuai Gao, zhichao fan, Xingzhuo Guo, Zhitao Hou, Xiao Lv, Ran Jia, Shi Han and Dongmei Zhang 11:00-12:30 (Pier 7&8)
We propose HermEs, the first approach for spreadsheet formula prediction via HiEraRchical forMulet ExpanSion, where hierarchical expansion means generating formulas following the underlying parse tree structure, and Formulet refers to commonly-used multi-level patterns mined from real formula parse trees. HermEs improves the formula prediction accuracy by (1) guaranteeing correct grammar by hierarchical generation rather than left-to-right generation and (2) significantly streamlining the token-level decoding with high-level Formulet. Notably, instead of generating formulas in a pre-defined fixed order, we propose a novel sampling strategy to systematically exploit a variety of hierarchical and multi-level expansion orders and provided solid mathematical proof, with the aim of meeting diverse human needs of the formula writing order in real applications. We further develop an interactive formula completion interface based on HERMES, which shows a new user experience in https://github.com/formulet/HERMES.

Exploring and Verbalizing Academic Ideas by Concept Co-occurrence
Yi Xu, Shuqian Sheng, Bo Xue, Luoyi Fu, Xinbing Wang and Chenghu Zhou   11:00-12:30 (Pier 7&8)
Researchers usually come up with new ideas only after thoroughly comprehending vast quantities of literature. The difficulty of this procedure is exacerbated by the fact that the number of academic publications is growing exponentially. In this study, we devise a framework based on concept co-occurrence for academic idea inspiration, which has been integrated into a research assistant system. From our perspective, the emergence of a new idea can be regarded as the fusion of two concepts that co-occur in an academic paper. We construct evolving concept graphs according to the co-occurrence relationship of concepts from 20 disciplines or topics. Then we design a temporal link prediction method based on masked language model to explore potential connections between different concepts. To verbalize the newly discovered connections, we also utilize the pretrained language model to generate a description of an idea based on a new data structure called co-occurrence citation quintuple. We evaluate our proposed system using both automatic metrics and human assessment. The results demonstrate that our system has broad prospects and can assist researchers in expediting the process of discovering new ideas.

GVdoc - Graph-based Visual DOcument Classification
Fnu Mohbat, Mohammed J Zaki, Catherine Finegan-Dollak and Ashish Verma   11:00-12:30 (Pier 7&8)
The robustness of a model for real-world deployment is decided by how well it performs on unseen data and distinguishes between in-domain and out-of-domain samples. Visual document classifiers have shown impressive performance on in-distribution test sets. However, they tend to have a hard time correctly classifying and differentiating out-of-distribution examples. Image-based classifiers lack the text component, whereas multi-modality transformer-based models face the token serialization problem in visual documents due to their diverse layouts. They also require a lot of computing power during inference, making them impractical for many real-world applications. We propose, GVdoc, a graph-based document classification model that addresses both of these challenges. Our approach generates a document graph based on its layout, and then trains a graph neural network to learn node and graph embeddings. Through experiments, we show that our model, even with fewer parameters, outperforms state-of-the-art models on out-of-distribution data while retaining comparable performance on the in-distribution test set.

AMR-TST: Abstract Meaning Representation-based Text Style Transfer
Kaize Shi, Xueyao Sun, LI HE, Dingxian Wang, Qing Li and Guandong Xu   11:00-12:30 (Pier 7&8)
Abstract Meaning Representation (AMR) is a semantic representation that can enhance natural language generation (NLG) by providing a logical semantic input. In this paper, we propose the AMR-TST, an AMR-based text style transfer (TST) technique. The AMR-TST converts the source text to an AMR graph and generates the transferred text based on the AMR graph modified by a TST policy named style rewriting. Our method combines both the explainability and diversity of explicit and implicit TST methods. The experiments show that the proposed method achieves state-of-the-art results compared with other baseline models in automatic and human evaluations. The generated transferred text in qualitative evaluation proves the AMR-TST have significant advantages in keeping semantic features and reducing hallucinations. To

the best of our knowledge, this work is the first to apply the AMR method focusing on node-level features to the TST task.

Multi-Grained Knowledge Retrieval for End-to-End Task-Oriented Dialog
Fanqi Wan, Weizhou Shen, Ke Yang, Xiaojun Quan and Wei Bi   11:00-12:30 (Pier 7&8)
Retrieving proper domain knowledge from an external database lies at the heart of end-to-end task-oriented dialog systems to generate informative responses. Most existing systems blend knowledge retrieval with response generation and optimize them with direct supervision from reference responses, leading to suboptimal retrieval performance when the knowledge base becomes large-scale. To address this, we propose to decouple knowledge retrieval from response generation and introduce a multi-grained knowledge retriever (MAKER) that includes an entity selector to search for relevant entities and an attribute selector to filter out irrelevant attributes. To train the retriever, we propose a novel distillation objective that derives supervision signals from the response generator. Experiments conducted on three standard benchmarks with both small and large-scale knowledge bases demonstrate that our retriever performs knowledge retrieval more effectively than existing methods. Our code has been made publicly available at https://github.com/18907305772/MAKER.

Model-Based Simulation for Optimising Smart Reply
Benjamin Towle and Ke Zhou   11:00-12:30 (Pier 7&8)
Smart Reply (SR) systems present a user with a set of replies, of which one can be selected in place of having to type out a response. To perform well at this task, a system should be able to effectively present the user with a diverse set of options, to maximise the chance that at least one of them conveys the user's desired response. This is a significant challenge, due to the lack of datasets containing sets of responses to learn from. Resultantly, previous work has focused largely on post-hoc diversification, rather than explicitly learning to predict sets of responses. Motivated by this problem, we present a novel method SimSR, that employs model-based simulation to discover high-value response sets, through simulating possible user responses with a learned world model. Unlike previous approaches, this allows our method to directly optimise the end-goal of SR–maximising the relevance of at least one of the predicted replies. Empirically on two public datasets, when compared to SoTA baselines, our method achieves up to 21% and 18% improvement in ROUGE score and Self-ROUGE score respectively.

Decoupling Pseudo Label Disambiguation and Representation Learning for Generalized Intent Discovery
Yutao Mou, Xiaoshuai Song, Keqing He, Chen Zeng, Pei Wang, Jingang Wang, Yunsen Xian and Weiran Xu   11:00-12:30 (Pier 7&8)
Generalized intent discovery aims to extend a closed-set in-domain intent classifier to an open-world intent set including in-domain and out-of-domain intents. The key challenges lie in pseudo label disambiguation and representation learning. Previous methods suffer from a coupling of pseudo label disambiguation and representation learning, that is, the reliability of pseudo labels relies on representation learning, and representation learning is restricted by pseudo labels in turn. In this paper, we propose a decoupled prototype learning framework (DPL) to decouple pseudo label disambiguation and representation learning. Specifically, we firstly introduce prototypical contrastive representation learning (PCL) to get discriminative representations. And then we adopt a prototype-based label disambiguation method (PLD) to obtain pseudo labels. We theoretically prove that PCL and PLD work in a collaborative fashion and facilitate pseudo label disambiguation. Experiments and analysis on three benchmark datasets show the effectiveness of our method.

Enhancing Personalized Dialogue Generation with Contrastive Latent Variables: Combining Sparse and Dense Persona
Yihong Tang, Bo Wang, Miao Fang, Dongming Zhao, Kun Huang, Ruifang He and Yuexian Hou   11:00-12:30 (Pier 7&8)
The personalized dialogue explores the consistent relationship between dialogue generation and personality. Existing personalized dialogue agents model persona profiles from three resources: sparse or dense persona descriptions and dialogue histories. However, sparse structured persona attributes are explicit but uninformative, dense persona texts contain rich persona descriptions with much noise, and dialogue history query is both noisy and uninformative for persona modeling. In this work, we combine the advantages of the three resources to obtain a richer and more accurate persona. We design a Contrastive Latent Variable-based model (CLV) that clusters the dense persona descriptions into sparse categories, which are combined with the history query to generate personalized responses. Experimental results on Chinese and English datasets demonstrate our model's superiority in personalization.

Two Birds One Stone: Dynamic Ensemble for OOD Intent Classification
Yunhua Zhou, Jianqiang Yang, Pengyu Wang and Xipeng Qiu   11:00-12:30 (Pier 7&8)
Out-of-domain (OOD) intent classification is an active field of natural language understanding, which is of great practical significance for intelligent devices such as the Task-Oriented Dialogue System. It mainly contains two challenges: it requires the model to know what it knows and what it does not know.This paper investigates "overthinking" in the open-world scenario and its impact on OOD intent classification. Inspired by this, we propose a two-birds-one-stone method, which allows the model to decide whether to make a decision on OOD classification early during inference and can ensure accuracy and accelerate inference. At the same time, to adapt to the behavior of dynamic inference, we also propose a training method based on ensemble methods. In addition to bringing certain theoretical insights, we also conduct detailed experiments on three real-world intent datasets. Compared with the previous baselines, our method can not only improve inference speed, but also achieve significant performance improvements.

Knowledge-enhanced Mixed-initiative Dialogue System for Emotional Support Conversations
Yang Deng, Wenxuan Zhang, Yifei Yuan and Wai Lam   11:00-12:30 (Pier 7&8)
Unlike empathetic dialogues, the system in emotional support conversations (ESC) is expected to not only convey empathy for comforting the help-seeker, but also proactively assist in exploring and addressing their problems during the conversation. In this work, we study the problem of mixed-initiative ESC where the user and system can both take the initiative in leading the conversation. Specifically, we conduct a novel analysis on mixed-initiative ESC systems with a tailor-designed schema that divides utterances into different types with speaker roles and initiative types. Four emotional support metrics are proposed to evaluate the mixed-initiative interactions. The analysis reveals the necessity and challenges of building mixed-initiative ESC systems. In the light of this, we propose a knowledge-enhanced mixed-initiative framework (KEMI) for ESC, which retrieves actual case knowledge from a large-scale mental health knowledge graph for generating mixed-initiative responses. Experimental results on two ESC datasets show the superiority of KEMI in both content-preserving evaluation and mixed initiative related analyses.

Multi-Domain Dialogue State Tracking with Disentangled Domain-Slot Attention
Longfei Yang, Jiyi Li, Sheng Li and Takahiro Shinozaki   11:00-12:30 (Pier 7&8)
As the core of task-oriented dialogue systems, dialogue state tracking (DST) is designed to track the dialogue state through the conversation between users and systems. Multi-domain DST has been an important challenge in which the dialogue states across multiple domains need to consider. In recent mainstream approaches, each domain and slot are aggregated and regarded as a single query feeding into attention with the dialogue history to obtain domain-slot specific representations. In this work, we propose disentangled domain-slot attention for multi-domain dialogue state tracking. The proposed approach disentangles the domain-slot specific information extraction in a flexible and context-dependent manner by separating the query about domains and slots in the attention component. Through a series of experiments on MultiWOZ 2.0 and MultiWOZ 2.4 datasets, we demonstrate that our proposed approach outperforms the standard multi-head attention with aggregated domain-slot query.

How Well Apply Simple MLP to Incomplete Utterance Rewriting?
Jiang Li, Xiangdong Su, Xinlan Ma and Guanglai Gao   11:00-12:30 (Pier 7&8)

Incomplete utterance rewriting (IUR) aims to restore the incomplete utterance with sufficient context information for comprehension. This paper introduces a simple yet efficient IUR method. Different from prior studies, we first employ only one-layer **MLP** architecture to mine latent semantic information between joint utterances for **IUR** task (**MIUR**). After that, we conduct a joint feature matrix to predict the token type and thus restore the incomplete utterance. The well-designed network and simple architecture make our method significantly superior to existing methods in terms of quality and inference speedOur code is available at https://github.com/IMU-MachineLearningSXD/MIUR.

How About Kind of Generating Hedges using End-to-End Neural Models?
Alafate Abulimiti, Chloé Clavel and Justine Cassell   11:00-12:30 (Pier 7&8)

Hedging is a strategy for softening the impact of a statement in conversation. In reducing the strength of an expression, it may help to avoid embarrassment (more technically, "face threat") to one's listener. For this reason, it is often found in contexts of instruction, such as tutoring. In this work, we develop a model of hedge generation based on i) fine-tuning state-of-the-art language models trained on human-human tutoring data, followed by ii) reranking to select the candidate that best matches the expected hedging strategy within a candidate pool using a hedge classifier. We apply this method to a natural peer-tutoring corpus containing a significant number of disfluencies, repetitions, and repairs. The results show that generation in this noisy environment is feasible with reranking. By conducting an error analysis for both approaches, we reveal the challenges faced by systems attempting to accomplish both social and task-oriented goals in conversation.

DualGATs: Dual Graph Attention Networks for Emotion Recognition in Conversations
Duzhen Zhang, Feilong Chen and Xiuyi Chen   11:00-12:30 (Pier 7&8)

Capturing complex contextual dependencies plays a vital role in Emotion Recognition in Conversations (ERC). Previous studies have predominantly focused on speaker-aware context modeling, overlooking the discourse structure of the conversation. In this paper, we introduce Dual Graph ATtention networks (DualGATs) to concurrently consider the complementary aspects of discourse structure and speaker-aware context, aiming for more precise ERC. Specifically, we devise a Discourse-aware GAT (DisGAT) module to incorporate discourse structural information by analyzing the discourse dependencies between utterances. Additionally, we develop a Speaker-aware GAT (SpkGAT) module to incorporate speaker-aware contextual information by considering the speaker dependencies between utterances. Furthermore, we design an interaction module that facilitates the integration of the DisGAT and SpkGAT modules, enabling the effective interchange of relevant information between the two modules. We extensively evaluate our method on four datasets, and experimental results demonstrate that our proposed DualGATs surpass state-of-the-art baselines on the majority of the datasets.

The Whole Truth and Nothing But the Truth: Faithful and Controllable Dialogue Response Generation with Dataflow Transduction and Constrained Decoding
Hao Fang, Anusha Balakrishnan, Harsh Jhamtani, John Bufe, Jean Crawford, Jayant Krishnamurthy, Adam Pauls, Jason Eisner, Jacob Andreas and Dan Klein   11:00-12:30 (Pier 7&8)

In a real-world dialogue system, generated text must be truthful and informative while remaining fluent and adhering to a prescribed style. Satisfying these constraints simultaneously is difficult for the two predominant paradigms in language generation: neural language modeling and rule-based generation. We describe a hybrid architecture for dialogue response generation that combines the strengths of both paradigms. The first component of this architecture is a rule-based content selection model defined using a new formal framework called dataflow transduction, which uses declarative rules to transduce a dialogue agent's actions and their results (represented as dataflow graphs) into context-free grammars representing the space of contextually acceptable responses. The second component is a constrained decoding procedure that uses these grammars to constrain the output of a neural language model, which selects fluent utterances. Our experiments show that this system outperforms both rule-based and learned approaches in human evaluations of fluency, relevance, and truthfulness.

Learning New Skills after Deployment: Improving open-domain internet-driven dialogue with human feedback
Jing Xu, Megan Ung, Mojtaba Komeili, Kushal Arora, Y-Lan Boureau and Jason Weston   11:00-12:30 (Pier 7&8)

Frozen models trained to mimic static datasets can never improve their performance. Models that can employ internet-retrieval for up-to-date information and obtain feedback from humans during deployment provide the promise of both adapting to new information, and improving their performance. In this work we study how to improve internet-driven conversational skills in such a learning framework. We collect deployment data, which we make publicly available, of human interactions, and collect various types of human feedback – including binary quality measurements, free-form text feedback, and fine-grained reasons for failure. We then study various algorithms for improving from such feedback, including standard supervised learning, rejection sampling, model-guiding and reward-based learning, in order to make recommendations on which type of feed- back and algorithms work best. We find the recently introduced DIRECTOR model (Arora et al., 2022) shows significant improvements over other existing approaches.

Injecting Comparison Skills in Task-Oriented Dialogue Systems for Database Search Results Disambiguation
Yongil Kim, Yerin Hwang, Joongbo Shin, Hyunkyung Bae and Kyomin Jung   11:00-12:30 (Pier 7&8)

In task-oriented dialogue (TOD) systems designed to aid users accomplish specific goals in one or more domains, the agent retrieves entities that satisfy user constraints from the database. However, when multiple database search results exist, an ambiguity occurs regarding which results to select and present to the user. Existing TOD systems handle this ambiguity by randomly selecting one or few results and presenting their names to the user. However, in a real scenario, users do not always accept a randomly recommended entity, and users should have access to more comprehensive information about the search results. To address this limitation, we propose a novel task called Comparison-Based database search Ambiguity handling (CBA), that handles ambiguity in database search results by comparing the properties of multiple entities to enable users to choose according to their preferences. Accordingly, we introduce a new framework for automatically collecting high-quality dialogue data along with the Disambiguating Schema-guided Dialogue (DSD) dataset, an augmented version of the SGD dataset. Experimental studies on the DSD dataset demonstrate that training baseline models with the dataset effectively address the CBA task. Our dataset and code will be publicized.

The CRINGE Loss: Learning what language not to model
Leonard Adolphs, Tianyu Gao, Jing Xu, Kurt Shuster, Sainbayar Sukhbaatar and Jason Weston   11:00-12:30 (Pier 7&8)

Standard language model training employs gold human documents or human-human interaction data, and treats all training data as positive examples. Growing evidence shows that even with very large amounts of positive training data, issues remain that can be alleviated with relatively small amounts of negative data – examples of what the model should not do. In this work, we propose a novel procedure to train with such data called the "CRINGE" loss (ContRastive Iterative Negative GEneration). We show the effectiveness of this approach across three different experiments on the tasks of safe generation, contradiction avoidance, and open-domain dialogue. Our models outperform multiple strong baselines and are conceptually simple, easy to train and implement.

Disfluency Generation for More Robust Dialogue Systems
Benjamin Marie   11:00-12:30 (Pier 7&8)

Disfluencies in user utterances can trigger a chain of errors impacting all the modules of a dialogue system: natural language understanding,

dialogue state tracking, and response generation. In this work, we first analyze existing dialogue datasets commonly used in research and show that they only contain a marginal number of disfluent utterances. Due to this relative absence of disfluencies in their training data, dialogue systems may then critically fail when exposed to disfluent utterances. Following this observation, we propose to augment existing datasets with disfluent user utterances by paraphrasing fluent utterances into disfluent ones. Relying on a pre-trained language model, our few-shot disfluent paraphraser guided by a disfluency classifier can generate useful disfluent utterances for training better dialogue systems. We report on improvements for both dialogue state tracking and response generation when the dialogue systems are trained on datasets augmented with our disfluent utterances.

RHO: Reducing Hallucination in Open-domain Dialogues with Knowledge Grounding
Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu, Bryan Wilie, Min Zeng and Pascale Fung   11:00-12:30 (Pier 7&8)
Dialogue systems can leverage large pre-trained language models and knowledge to generate fluent and informative responses. However, these models are still prone to produce hallucinated responses not supported by the input source, which greatly hinders their application. The heterogeneity between external knowledge and dialogue context challenges representation learning and source integration, which further contributes to unfaithfulness. To handle this challenge and generate more faithful responses, this paper presents RHO utilizing the representations of linked entities and relation predicates from a knowledge graph (KG). We propose (1) local knowledge grounding to combine textual embeddings with the corresponding KG embeddings; and (2) global knowledge grounding to equip RHO with multi-hop reasoning abilities via the attention mechanism. In addition, we devise a response re-ranking technique based on walks over KG sub-graphs for better conversational reasoning. Experimental results on OpenDialKG (Moon et al., 2019) show that our approach significantly outperforms state-of-the-art methods on both automatic and human evaluation by a large margin, especially in hallucination reduction (17.54% in FeQA (Durmus et al., 2020)).

Counterfactual Multihop QA: A Cause-Effect Approach for Reducing Disconnected Reasoning
Wangzhen Guo, Qinkang Gong, Yanghui Rao and Hanjiang Lai   11:00-12:30 (Pier 7&8)
Multi-hop QA requires reasoning over multiple supporting facts to answer the question. However, the existing QA models always rely on shortcuts, e.g., providing the true answer by only one fact, rather than multi-hop reasoning, which is referred as disconnected reasoning problem. To alleviate this issue, we propose a novel counterfactual multihop QA, a causal-effect approach that enables to reduce the disconnected reasoning. It builds upon explicitly modeling of causality: 1) the direct causal effects of disconnected reasoning and 2) the causal effect of true multi-hop reasoning from the total causal effect. With the causal graph, a counterfactual inference is proposed to disentangle the disconnected reasoning from the total causal effect, which provides us a new perspective and technology to learn a QA model that exploits the true multi-hop reasoning instead of shortcuts. Extensive experiments have been conducted on the benchmark HotpotQA dataset, which demonstrate that the proposed method can achieve notable improvement on reducing disconnected reasoning. For example, our method achieves 5.8% higher points of its Supps score on HotpotQA through true multihop reasoning. The code is available at https://github.com/guowzh/CFMQA.

Long-Tailed Question Answering in an Open World
Yi Dai, Hao Lang, Yinhe Zheng, Fei Huang and Yongbin Li   11:00-12:30 (Pier 7&8)
Real-world data often have an open long-tailed distribution, and building a unified QA model supporting various tasks is vital for practical QA applications. However, it is non-trivial to extend previous QA approaches since they either require access to seen tasks of adequate samples or do not explicitly model samples from unseen tasks. In this paper, we define Open Long-Tailed QA (OLTQA) as learning from long-tailed distributed data and optimizing performance over seen and unseen QA tasks. We propose an OLTQA model that encourages knowledge sharing between head, tail and unseen tasks, and explicitly mines knowledge from a large pre-trained language model (LM). Specifically, we organize our model through a pool of fine-grained components and dynamically combine these components for an input to facilitate knowledge sharing. A retrieve-then-rerank frame is further introduced to select in-context examples, which guild the LM to generate text that express knowledge for QA tasks. Moreover, a two-stage training approach is introduced to pre-train the framework by knowledge distillation (KD) from the LM and then jointly train the frame and a QA model through an adaptive mutual KD method. On a large-scale OLTQA dataset we curate from 43 existing QA datasets, our model consistently outperforms the state-of-the-art.

AttenWalker: Unsupervised Long-Document Question Answering via Attention-based Graph Walking
Yuxiang Nie, Heyan Huang, Wei Wei and Xian-Ling Mao   11:00-12:30 (Pier 7&8)
Annotating long-document question answering (long-document QA) pairs is time-consuming and expensive. To alleviate the problem, it might be possible to generate long-document QA pairs via unsupervised question answering (UQA) methods. However, existing UQA tasks are based on short documents, and can hardly incorporate long-range information. To tackle the problem, we propose a new task, named unsupervised long-document question answering (ULQA), aiming to generate high-quality long-document QA instances in an unsupervised manner. Besides, we propose AttenWalker, a novel unsupervised method to aggregate and generate answers with long-range dependency so as to construct long-document QA pairs. Specifically, AttenWalker is composed of three modules, i.e. span collector, span linker and answer aggregator. Firstly, the span collector takes advantage of constituent parsing and reconstruction loss to select informative candidate spans for constructing answers. Secondly, with the help of the attention graph of a pre-trained long-document model, potentially interrelated text spans (that might be far apart) could be linked together via an attention-walking algorithm. Thirdly, in the answer aggregator, linked spans are aggregated into the final answer via the mask-filling ability of a pre-trained model. Extensive experiments show that AttenWalker outperforms previous methods on NarrativeQA and Qasper. In addition, AttenWalker also shows strong performance in the few-shot learning setting.

TimelineQA: A Benchmark for Question Answering over Timelines
Wang-Chiew Tan, Jane Dwivedi-Yu, Yuliang Li, Lambert Mathias, Marzieh Saeidi, Jing Nathan Yan and Alon Y. Halevy   11:00-12:30 (Pier 7&8)
Lifelogs are descriptions of experiences that a person had during their life. Lifelogs are created by fusing data from the multitude of digital services, such as online photos, maps, shopping and content streaming services. Question answering over lifelogs can offer personal assistants a critical resource when they try to provide advice in context. However, obtaining answers to questions over lifelogs is beyond the current state of the art of question answering techniques for a variety of reasons, the most pronounced of which is that lifelogs combine free text with some degree of structure such as temporal and geographical information.

We create and publicly release TimelineQA, a benchmark for accelerating progress on querying lifelogs. TimelineQA generates lifelogs of imaginary people. The episodes in the lifelog range from major life episodes such as high school graduation to those that occur on a daily basis such as going for a run. We describe a set of experiments on TimelineQA with several state-of-the-art QA models. Our experiments reveal that for atomic queries, an extractive QA system significantly out-performs a state-of-the-art retrieval-augmented QA system. For multi-hop queries involving aggregates, we show that the best result is obtained with a state-of-the-art table QA technique, assuming the ground truth set of episodes for deriving the answer is available.

MVP-Tuning: Multi-View Knowledge Retrieval with Prompt Tuning for Commonsense Reasoning
Yongfeng Huang, Yanyang Li, Yichong Xu, Lin Zhang, ruyi gan, Jiaxing Zhang and Liwei Wang   11:00-12:30 (Pier 7&8)
Recent advances in pre-trained language models (PLMs) have facilitated the development of commonsense reasoning tasks. However, existing methods rely on multi-hop knowledge retrieval and thus suffer low accuracy due to embedded noise in the acquired knowledge. In addition, these methods often attain high computational costs and nontrivial knowledge loss because they encode the knowledge independently of the

PLM, making it less relevant to the task and thus resulting in a poor local optimum. In this work, we propose MultiView Knowledge Retrieval with Prompt Tuning (MVP-Tuning). MVP-Tuning leverages similar question-answer pairs in the training set to improve knowledge retrieval and employs a single prompt-tuned PLM to model knowledge and input text jointly. We conduct our experiments on five commonsense reasoning QA benchmarks to show that MVP-Tuning outperforms all other baselines in 4 out of 5 datasets with less than 2% trainable parameters. MVPTuning even gets a new state-of-the-art result on OpenBookQA and is number one on the leaderboard.

Solving Math Word Problems via Cooperative Reasoning induced Language Models
Xinyu Zhu, Junjie Wang, Lin Zhang, Yuxiang Zhang, Yongfeng Huang, ruyi gan, Jiaxing Zhang and Yujiu Yang   11:00-12:30 (Pier 7&8)
Large-scale pre-trained language models (PLMs) bring new opportunities to challenging problems, especially those that need high-level intelligence, such as the math word problem (MWPs). However, directly applying existing PLMs to MWPs can fail as the generation process lacks sufficient supervision and thus lacks fast adaptivity as humans. We notice that human reasoning has a dual reasoning framework that consists of an immediate reaction system (system 1) and a delicate reasoning system (system 2), where the entire reasoning is determined by their interaction. This inspires us to develop a cooperative reasoning-induced PLM for solving MWPs, called Cooperative Reasoning (CoRe), resulting in a human-like reasoning architecture with system 1 as the generator and system 2 as the verifier. In our approach, the generator is responsible for generating reasoning paths, and the verifiers are used to supervise the evaluation in order to obtain reliable feedback for the generator. We evaluate our CoRe framework on several mathematical reasoning datasets and achieve decent improvement over state-of-the-art methods, up to 9.6% increase over best baselines.

SConE: Simplified Cone Embeddings with Symbolic Operators for Complex Logical Queries
Chau Duc Minh Nguyen, Tim N French, Wei Liu and Michael Stewart   11:00-12:30 (Pier 7&8)
Geometric representation of query embeddings (using points, particles, rectangles and cones) can effectively achieve the task of answering complex logical queries expressed in first-order logic (FOL) form over knowledge graphs, allowing intuitive encodings. However, current geometric-based methods depend on the neural approach to model FOL operators (conjunction, disjunction and negation), which are not easily explainable with considerable computation cost. We overcome this challenge by introducing a symbolic modeling approach for the FOL operators, emphasizing the direct calculation of the intersection between geometric shapes, particularly sector-cones in the embedding space, to model the conjunction operator. This approach reduces the computation cost as a non-neural approach is involved in the core logic operators. Moreover, we propose to accelerate the learning in the relation projection operator using the neural approach to emphasize the essential role of this operator in all query structures. Although empirical evidence for explainability is challenging, our approach demonstrates a significant improvement in answering complex logical queries (both non-negative and negative FOL forms) over previous geometric-based models.

KoRC: Knowledge Oriented Reading Comprehension Benchmark for Deep Text Understanding
Zijun Yao, Yantao Liu, Xin Lv, Shulin Cao, Jifan Yu, Juanzi Li and Lei Hou   11:00-12:30 (Pier 7&8)
Deep text understanding, which requires the connections between a given document and prior knowledge beyond its text, has been highlighted by many benchmarks in recent years. However, these benchmarks have encountered two major limitations. On the one hand, most of them require human annotation of knowledge, which leads to limited knowledge coverage. On the other hand, they usually use choices or spans in the texts as the answers, which results in narrow answer space. To overcome these limitations, we build a new challenging benchmark named KoRC in this paper. Compared with previous benchmarks, KoRC has two advantages, i.e., broad knowledge coverage and flexible answer format. Specifically, we utilize massive knowledge bases to guide annotators or large language models (LLMs) to construct knowledgable questions. Moreover, we use labels in knowledge bases rather than spans or choices as the final answers. We test state-of-the-art models on KoRC and the experimental results show that the strongest baseline only achieves 68.3% and 30.0% F1 measure in the IID and OOD test set, respectively. These results indicate that deep text understanding is still an unsolved challenge. We will release our dataset and baseline methods upon acceptance.

Multi-granularity Temporal Question Answering over Knowledge Graphs
Ziyang Chen, Jinzhi Liao and Xiang Zhao   11:00-12:30 (Pier 7&8)
Recently, question answering over temporal knowledge graphs (i.e., TKGQA) has been introduced and investigated, in quest of reasoning about dynamic factual knowledge. To foster research on TKGQA, a few datasets have been curated (e.g., CronQuestions and ComplexCronQuestions), and various models have been proposed based on these datasets. Nevertheless, existing efforts overlook the fact that real-life applications of TKGQA also tend to be complex in temporal granularity, i.e., the questions may concern mixed temporal granularities (e.g., both day and month). To overcome the limitation, in this paper, we motivate the notion of multi-granularity temporal question answering over knowledge graphs and present a large scale dataset for multi-granularity TKGQA, namely MultiTQ. To the best of our knowledge, MultiTQis among the first of its kind, and compared with existing datasets on TKGQA, MultiTQfeatures at least two desirable aspects—ample relevant facts and multiple temporal granularities. It is expected to better reflect real-world challenges, and serve as a test bed for TKGQA models. In addition, we propose a competing baseline MultiQA over MultiTQ, which is experimentally demonstrated to be effective in dealing with TKGQA. The data and code are released at https://github.com/czy1999/MultiTQ.

IM-TQA: A Chinese Table Question Answering Dataset with Implicit and Multi-type Table Structures
Mingyu Zheng, Yang Hao, Wenbin Jiang, Zheng Lin, Yajuan Lyu, QiaoQiao She and Weiping Wang   11:00-12:30 (Pier 7&8)
Various datasets have been proposed to promote the development of Table Question Answering (TQA) technique. However, the problem setting of existing TQA benchmarks suffers from two limitations. First, they directly provide models with explicit table structures where row headers and column headers of the table are explicitly annotated and treated as model input during inference. Second, they only consider tables of limited types and ignore other tables especially complex tables with flexible header locations. Such simplified problem setting cannot cover practical scenarios where models need to process tables without header annotations in the inference phase or tables of different types. To address above issues, we construct a new TQA dataset with implicit and multi-type table structures, named IM-TQA, which not only requires the model to understand tables without directly available header annotations but also to handle multi-type tables including previously neglected complex tables. We investigate the performance of recent methods on our dataset and find that existing methods struggle in processing implicit and multi-type table structures. Correspondingly, we propose an RGCN-RCI framework outperforming recent baselines. We will release our dataset to facilitate future research.

Synthesize, Prompt and Transfer: Zero-shot Conversational Question Generation with Pre-trained Language Model
Hongwei Zeng, Bifan Wei, Jun Liu and Weiping Fu   11:00-12:30 (Pier 7&8)
Conversational question generation aims to generate questions that depend on both context and conversation history. Conventional works utilizing deep learning have shown promising results, but heavily rely on the availability of large-scale annotated conversations. In this paper, we introduce a more realistic and less explored setting, Zero-shot Conversational Question Generation (ZeroCQG), which requires no human-labeled conversations for training. To solve ZeroCQG, we propose a multi-stage knowledge transfer framework, Synthesize, Prompt, and trAnsfer with pRe-Trained lAnguage model (SPARTA) to effectively leverage knowledge from single-turn question generation instances. To validate the zero-shot performance of SPARTA, we conduct extensive experiments on three conversational datasets: CoQA, QuAC, and DoQA by transferring knowledge from three single-turn datasets: MS MARCO, NewsQA, and SQuAD. The experimental results demonstrate

the superior performance of our method. Specifically, SPARTA has achieved 14.81 BLEU-4 (88.2% absolute improvement compared to T5) in CoQA with knowledge transferred from SQuAD.

Hypothetical Training for Robust Machine Reading Comprehension of Tabular Context

Moxin Li, Wenjie Wang, Fuli Feng, Hanwang Zhang, Qifan Wang and Tat-Seng Chua 11:00-12:30 (Pier 7&8)

Machine Reading Comprehension (MRC) models easily learn spurious correlations from complex contexts such as tabular data. Counterfactual training—using the factual and counterfactual data by augmentation—has become a promising solution. However, it is costly to construct faithful counterfactual examples because it is tricky to maintain the consistency and dependency of the tabular data. In this paper, we take a more efficient fashion to ask **hypothetical questions** like *"in which year would the net profit be larger if the revenue in 2019 were $38,298?"*, whose effects on the answers are equivalent to those expensive counterfactual tables. We propose a hypothetical training framework that uses paired examples with different hypothetical questions to supervise the direction of model gradient towards the counterfactual answer change. The superior generalization results on tabular MRC datasets, including a newly constructed stress test and MultiHiertt, validate our effectiveness.

RobustQA: Benchmarking the Robustness of Domain Adaptation for Open-Domain Question Answering

Rujun Han, Peng Qi, Yuhao Zhang, Lan Liu, Juliette Burger, William Yang Wang, zhiheng huang, Bing Xiang and Dan Roth 11:00-12:30 (Pier 7&8)

Open-domain question answering (ODQA) is a crucial task in natural language processing. A typical ODQA system relies on a retriever module to select relevant contexts from a large corpus for a downstream reading comprehension model. Existing ODQA datasets consist mainly of Wikipedia corpus, and are insufficient to study models' generalizability across diverse domains as models are trained and evaluated on the same genre of data. We propose **RobustQA**, a novel benchmark consisting of datasets from 8 different domains, which facilitates the evaluation of ODQA's domain robustness. To build **RobustQA**, we annotate QA pairs in retrieval datasets with rigorous quality control. We further examine improving QA performances by incorporating unsupervised learning methods with target-domain corpus and adopting large generative language models. These methods can effectively improve model performances on **RobustQA**. However, experimental results demonstrate a significant gap from in-domain training, suggesting that **RobustQA** is a challenging benchmark to evaluate ODQA domain robustness.

Knowing-how & Knowing-that: A New Task for Machine Comprehension of User Manuals

Hongru Liang, Jia Liu, Weihong Du, Dingnan Jin, Wenqiang Lei, Zujie Wen and Jiancheng Lv 11:00-12:30 (Pier 7&8)

The machine reading comprehension (MRC) of user manuals has huge potential in customer service. However, current methods have trouble answering complex questions. Therefore, we introduce the knowing-how & knowing-that task that requires the model to answer factoid-style, procedure-style, and inconsistent questions about user manuals. We resolve this task by jointly representing the sTeps and fActs in a gRAh (TARA), which supports a unified inference of various questions. Towards a systematical benchmarking study, we design a heuristic method to automatically parse user manuals into TARAs and build an annotated dataset to test the model's ability in answering real-world questions. Empirical results demonstrate that representing user manuals as TARAs is a desired solution for the MRC of user manuals. An in-depth investigation of TARA further sheds light on the issues and broader impacts of future representations of user manuals. We hope our work can move the MRC of user manuals to a more complex and realistic stage.

USSA: A Unified Table Filling Scheme for Structured Sentiment Analysis

Zepeng Zhai, Hao Chen, Ruifan Li and Xiaojie WANG 11:00-12:30 (Pier 7&8)

Most previous studies on Structured Sentiment Analysis (SSA) have cast it as a problem of bi-lexical dependency parsing, which cannot address issues of overlap and discontinuity simultaneously. In this paper, we propose a niche-targeting and effective solution. Our approach involves creating a novel bi-lexical dependency parsing graph, which is then converted to a unified 2D table-filling scheme, namely USSA. The proposed scheme resolves the kernel bottleneck of previous SSA methods by utilizing 13 different types of relations. In addition, to closely collaborate with the USSA scheme, we have developed a model that includes a proposed bi-axial attention module to effectively capture the correlations among relations in the rows and columns of the table. Extensive experimental results on benchmark datasets demonstrate the effectiveness and robustness of our proposed framework, outperforming state-of-the-art methods consistently.

Don't Lose Yourself! Empathetic Response Generation via Explicit Self-Other Awareness

Weixiang Zhao, Yanyan Zhao, Xin Lu and Bing Qin 11:00-12:30 (Pier 7&8)

As a critical step to achieve human-like chatbots, empathetic response generation has attained increasing interests. Previous attempts are incomplete and not sufficient enough to elicit empathy because they only stay on the initial stage of empathy to automatically sense and simulate the feelings and thoughts of others via other-awareness. However, they ignore to include self-awareness to consider the own views of the self in their responses, which is a crucial process to achieve the empathy. To this end, we propose to generate Empathetic response with explicit Self-Other Awareness (EmpSOA). Specifically, three stages, self-other differentiation, self-other modulation and self-other generation, are devised to clearly maintain, regulate and inject the self-other aware information into the process of empathetic response generation. Both automatic and human evaluations on the benchmark dataset demonstrate the superiority of EmpSOA to generate more empathetic responses. Our source code will be publicly available.

Exploiting Rich Textual User-Product Context for Improving Personalized Sentiment Analysis

Chenyang Lyu, Linyi Yang, Yue Zhang, Yvette Graham and Jennifer Foster 11:00-12:30 (Pier 7&8)

User and product information associated with a review is useful for sentiment polarity prediction. Typical approaches incorporating such information focus on modeling users and products as implicitly learned representation vectors. Most do not exploit the potential of historical reviews, or those that currently do require unnecessary modifications to model architecture or do not make full use of user/product associations. The contribution of this work is twofold: i) a method to explicitly employ historical reviews belonging to the same user/product in initializing representations, and ii) efficient incorporation of textual associations between users and products via a user-product cross-context module. Experiments on the IMDb, Yelp-2013 and Yelp-2014 English benchmarks with BERT, SpanBERT and Longformer pretrained language models show that our approach substantially outperforms previous state-of-the-art.

DiaASQ: A Benchmark of Conversational Aspect-based Sentiment Quadruple Analysis

Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua and Donghong Ji 11:00-12:30 (Pier 7&8)

The rapid development of aspect-based sentiment analysis (ABSA) within recent decades shows great potential for real-world society. The current ABSA works, however, are mostly limited to the scenario of a single text piece, leaving the study in dialogue contexts unexplored. To bridge the gap between fine-grained sentiment analysis and conversational opinion mining, in this work, we introduce a novel task of conversational aspect-based sentiment quadruple analysis, namely DiaASQ, aiming to detect the quadruple of target-aspect-opinion-sentiment in a dialogue. We manually construct a large-scale high-quality DiaASQ dataset in both Chinese and English languages. We deliberately develop a neural model to benchmark the task, which advances in effectively performing end-to-end quadruple prediction, and manages to incorporate rich dialogue-specific and discourse feature representations for better cross-utterance quadruple extraction. We hope the new benchmark will

spur more advancements in the sentiment analysis community.

On Text-based Personality Computing: Challenges and Future Directions

Qixiang Fang, Anastasia Giachanou, Ayoub Bagheri, Laura Boeschoten, Erik-Jan van Kesteren, Mahdi Shafiee Kamalabad and Daniel Oberski 11:00-12:30 (Pier 7&8)

Text-based personality computing (TPC) has gained many research interests in NLP. In this paper, we describe 15 challenges that we consider deserving the attention of the NLP research community. These challenges are organized by the following topics: personality taxonomies, measurement quality, datasets, performance evaluation, modelling choices, as well as ethics and fairness. When addressing each challenge, not only do we combine perspectives from both NLP and social sciences, but also offer concrete suggestions. We hope to inspire more valid and reliable TPC research.

Two Heads Are Better Than One: Improving Fake News Video Detection by Correlating with Neighbors

Peng Qi, Yuyang Zhao, Yufeng Shen, Wei Ji, Juan Cao and Tat-Seng Chua 11:00-12:30 (Pier 7&8)

The prevalence of short video platforms has spawned a lot of fake news videos, which have stronger propagation ability than textual fake news. Thus, automatically detecting fake news videos has been an important countermeasure in practice. Previous works commonly verify each news video individually with multimodal information. Nevertheless, news videos from different perspectives regarding the same event are commonly posted together, which contain complementary or contradictory information and thus can be used to evaluate each other mutually. To this end, we introduce a new and practical paradigm, i.e., cross-sample fake news video detection, and propose a novel framework, Neighbor-Enhanced fakE news video Detection (NEED), which integrates the neighborhood relationship of new videos belonging to the same event. NEED can be readily combined with existing single-sample detectors and further enhance their performances with the proposed graph aggregation (GA) and debunking rectification (DR) modules. Specifically, given the feature representations obtained from single-sample detectors, GA aggregates the neighborhood information with the dynamic graph to enrich the features of independent samples. After that, DR explicitly leverages the relationship between debunking videos and fake news videos to refute the candidate videos via textual and visual consistency. Extensive experiments on the public benchmark demonstrate that NEED greatly improves the performance of both single-modal (up to 8.34% in accuracy) and multimodal (up to 4.97% in accuracy) base detectors.

Towards Open-Domain Twitter User Profile Inference

Haoyang Wen, Zhenxin Xiao, Eduard H Hovy and Alexander Hauptmann 11:00-12:30 (Pier 7&8)

Twitter user profile inference utilizes information from Twitter to predict user attributes (e.g., occupation, location), which is controversial because of its usefulness for downstream applications and its potential to reveal users' privacy. Therefore, it is important for researchers to determine the extent of profiling in a safe environment to facilitate proper use and make the public aware of the potential risks. Contrary to existing approaches on limited attributes, we explore open-domain Twitter user profile inference. We conduct a case study where we collect publicly available WikiData public figure profiles and use diverse WikiData predicates for profile inference. After removing sensitive attributes, our data contains over 150K public figure profiles from WikiData, over 50 different attribute predicates, and over 700K attribute values. We further propose a prompt-based generation method, which can infer values that are implicitly mentioned in the Twitter information. Experimental results show that the generation-based approach can infer more comprehensive user profiles than baseline extraction-based methods, but limitations still remain to be applied for real-world use. We also enclose a detailed ethical statement for our data, potential benefits and risks from this work, and our efforts to mitigate the risks.

Measuring Intersectional Biases in Historical Documents

Nadav Borenstein, Karolina Stanczak, Thea Rolskov, Natacha Klein Käfer, Natália da Silva Perez and Isabelle Augenstein 11:00-12:30 (Pier 7&8)

Data-driven analyses of biases in historical texts can help illuminate the origin and development of biases prevailing in modern society. However, digitised historical documents pose a challenge for NLP practitioners as these corpora suffer from errors introduced by optical character recognition (OCR) and are written in an archaic language. In this paper, we investigate the continuities and transformations of bias in historical newspapers published in the Caribbean during the colonial era (18th to 19th centuries). Our analyses are performed along the axes of gender, race, and their intersection. We examine these biases by conducting a temporal study in which we measure the development of lexical associations using distributional semantics models and word embeddings. Further, we evaluate the effectiveness of techniques designed to process OCR-generated data and assess their stability when trained on and applied to the noisy historical newspapers. We find that there is a trade-off between the stability of the word embeddings and their compatibility with the historical dataset. We provide evidence that gender and racial biases are interdependent, and their intersection triggers distinct effects. These findings align with the theory of intersectionality, which stresses that biases affecting people with multiple marginalised identities compound to more than the sum of their constituents.

It's not Sexually Suggestive; It's Educative | Separating Sex Education from Suggestive Content on TikTok videos

Enfa Rose George and Mihai Surdeanu 11:00-12:30 (Pier 7&8)

We introduce SexTok, a multi-modal dataset composed of TikTok videos labeled as sexually suggestive (from the annotator's point of view), sex-educational content, or neither. Such a dataset is necessary to address the challenge of distinguishing between sexually suggestive content and virtual sex education videos on TikTok. Children's exposure to sexually suggestive videos has been shown to have adversarial effects on their development (Collins et al. 2017). Meanwhile, virtual sex education, especially on subjects that are more relevant to the LGBTQIA+ community, is very valuable (Mitchell et al. 2014). The platform's current system removes/punishes some of both types of videos, even though they serve different purposes. Our dataset contains video URLs, and it is also audio transcribed. To validate its importance, we explore two transformer-based models for classifying the videos. Our preliminary results suggest that the task of distinguishing between these types of videos is learnable but challenging. These experiments suggest that this dataset is meaningful and invites further study on the subject.

Contrastive Learning of Sociopragmatic Meaning in Social Media

Chiyu Zhang, Muhammad Abdul-Mageed and Ganesh Jawahar 11:00-12:30 (Pier 7&8)

Recent progress in representation and contrastive learning in NLP has not widely considered the class of sociopragmatic meaning (i.e., meaning in interaction within different language communities). To bridge this gap, we propose a novel framework for learning task-agnostic representations transferable to a wide range of sociopragmatic tasks (e.g., emotion, hate speech, humor, sarcasm). Our framework outperforms other contrastive learning frameworks for both in-domain and out-of-domain data, across both the general and few-shot settings. For example, compared to two popular pre-trained language models, our model obtains an improvement of 11.66 average F1 on 16 datasets when fine-tuned on only 20 training samples per dataset. We also show that our framework improves uniformity and preserves the semantic structure of representations. Our code is available at: https://github.com/UBC-NLP/infodcl

Causal Matching with Text Embeddings: A Case Study in Estimating the Causal Effects of Peer Review Policies

Raymond Zhang, Neha Nayak Kennard, Daniel S Smith, Daniel A McFarland, Andrew McCallum and Katherine A Keith 11:00-12:30 (Pier 7&8)

A promising approach to estimate the causal effects of peer review policies is to analyze data from publication venues that shift policies from single-blind to double-blind from one year to the next. However, in these settings the content of the manuscript is a confounding variable—

each year has a different distribution of scientific content which may naturally affect the distribution of reviewer scores. To address this textual confounding, we extend variable ratio nearest neighbor matching to incorporate text embeddings. We compare this matching method to a widely-used causal method of stratified propensity score matching and a baseline of randomly selected matches. For our case study of the ICLR conference shifting from single- to double-blind review from 2017 to 2018, we find human judges prefer manuscript matches from our method in 70% of cases. While the unadjusted estimate of the average causal effect of reviewers' scores is -0.25, our method shifts the estimate to -0.17, a slightly smaller difference between the outcomes of single- and double-blind policies. We hope this case study enables exploration of additional text-based causal estimation methods and domains in the future.

Responsibility Perspective Transfer for Italian Femicide News
Gosse Minnema, Huiyuan Lai, Benedetta Muscato and Malvina Nissim 11:00-12:30 (Pier 7&8)
Different ways of linguistically expressing the same real-world event can lead to different perceptions of what happened. Previous work has shown that different descriptions of gender-based violence (GBV) influence the reader's perception of who is to blame for the violence, possibly reinforcing stereotypes which see the victim as partly responsible, too. As a contribution to raise awareness on perspective-based writing, and to facilitate access to alternative perspectives, we introduce the novel task of automatically rewriting GBV descriptions as a means to alter the perceived level of blame on the perpetrator. We present a quasi-parallel dataset of sentences with low and high perceived responsibility levels for the perpetrator, and experiment with unsupervised (mBART-based), zero-shot and few-shot (GPT3-based) methods for rewriting sentences. We evaluate our models using a questionnaire study and a suite of automatic metrics.

Supervised Adversarial Contrastive Learning for Emotion Recognition in Conversations
Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou and Songlin Hu 11:00-12:30 (Pier 7&8)
Extracting generalized and robust representations is a major challenge in emotion recognition in conversations (ERC). To address this, we propose a supervised adversarial contrastive learning (SACL) framework for learning class-spread structured representations in a supervised manner. SACL applies contrast-aware adversarial training to generate worst-case samples and uses joint class-spread contrastive learning to extract structured representations. It can effectively utilize label-level feature consistency and retain fine-grained intra-class features. To avoid the negative impact of adversarial perturbations on context-dependent data, we design a contextual adversarial training (CAT) strategy to learn more diverse features from context and enhance the model's context robustness. Under the framework with CAT, we develop a sequence-based SACL-LSTM to learn label-consistent and context-robust features for ERC. Experiments on three datasets show that SACL-LSTM achieves state-of-the-art performance on ERC. Extended experiments prove the effectiveness of SACL and CAT.

Ideology Prediction from Scarce and Biased Supervision: Learn to Disregard the "What" and Focus on the "How"!
Chen Chen, Dylan Walker and Venkatesh Saligrama 11:00-12:30 (Pier 7&8)
We propose a novel supervised learning approach for political ideology prediction (PIP) that is capable of predicting out-of-distribution inputs. This problem is motivated by the fact that manual data-labeling is expensive, while self-reported labels are often scarce and exhibit significant selection bias. We propose a novel statistical model that decomposes the document embeddings into a linear superposition of two vectors; a latent neutral *context* vector independent of ideology, and a latent *position* vector aligned with ideology. We train an end-to-end model that has intermediate contextual and positional vectors as outputs. At deployment time, our model predicts labels for input documents by exclusively leveraging the predicted positional vectors. On two benchmark datasets we show that our model is capable of outputting predictions even when trained with as little as 5% biased data, and is significantly more accurate than the state-of-the-art. Through crowd-sourcing we validate the neutrality of contextual vectors, and show that context filtering results in ideological concentration, allowing for prediction on out-of-distribution examples.

Geo-Seq2seq: Twitter User Geolocation on Noisy Data through Sequence to Sequence Learning
Jingyu Zhang, Alexandra DeLucia, Chenyu Zhang and Mark Dredze 11:00-12:30 (Pier 7&8)
Location information can support social media analyses by providing geographic context. Some of the most accurate and popular Twitter geolocation systems rely on rule-based methods that examine the user-provided profile location, which fail to handle informal or noisy location names. We propose Geo-Seq2seq, a sequence-to-sequence (seq2seq) model for Twitter user geolocation that rewrites noisy, multilingual user-provided location strings into structured English location names. We train our system on tens of millions of multilingual location string and geotagged-tweet pairs. Compared to leading methods, our model vastly increases coverage (i.e., the number of users we can geolocate) while achieving comparable or superior accuracy. Our error analysis reveals that constrained decoding helps the model produce valid locations according to a location database. Finally, we measure biases across language, country of origin, and time to evaluate fairness, and find that while our model can generalize well to unseen temporal data, performance does vary by language and country.

NormMark: A Weakly Supervised Markov Model for Socio-cultural Norm Discovery
Farhad Moghimifar, shilin qu, Tongtong Wu, Yuan-Fang Li and Gholamreza Haffari 11:00-12:30 (Pier 7&8)
Norms, which are culturally accepted guidelines for behaviours, can be integrated into conversational models to generate utterances that are appropriate for the socio-cultural context. Existing methods for norm recognition tend to focus only on surface-level features of dialogues and do not take into account the interactions within a conversation. To address this issue, we propose NormMark, a probabilistic generative Markov model to carry the latent features throughout a dialogue. These features are captured by discrete and continuous latent variables conditioned on the conversation history, and improve the model's ability in norm recognition. The model is trainable on weakly annotated data using the variational technique. On a dataset with limited norm annotations, we show that our approach achieves higher F1 score, outperforming current state-of-the-art methods, including GPT3.

Improving Gradient Trade-offs between Tasks in Multi-task Text Classification
Heyan Chai, Jinhao Cui, Ye Wang, Min Zhang, Binxing Fang and Qing Liao 11:00-12:30 (Pier 7&8)
Multi-task learning (MTL) has emerged as a promising approach for sharing inductive bias across multiple tasks to enable more efficient learning in text classification. However, training all tasks simultaneously often yields degraded performance of each task than learning them independently, since different tasks might conflict with each other. Existing MTL methods for alleviating this issue is to leverage heuristics or gradient-based algorithm to achieve an arbitrary Pareto optimal trade-off among different tasks. In this paper, we present a novel gradient trade-off approach to mitigate the task conflict problem, dubbed GetMTL, which can achieve a specific trade-off among different tasks nearby the main objective of multi-task text classification (MTC), so as to improve the performance of each task simultaneously. The results of extensive experiments on two benchmark datasets back up our theoretical analysis and validate the superiority of our proposed GetMTL.

Reinforced Active Learning for Low-Resource, Domain-Specific, Multi-Label Text Classification
Lukas Wertz, Jasmina Bogojeska, Katsiaryna Mirylenka and Jonas Kuhn 11:00-12:30 (Pier 7&8)
Text classification datasets from specialised or technical domains are in high demand, especially in industrial applications. However, due to the high cost of annotation such datasets are usually expensive to create. While Active Learning (AL) can reduce the labeling cost, required AL strategies are often only tested on general knowledge domains and tend to use information sources that are not consistent across tasks. We propose Reinforced Active Learning (RAL) to train a Reinforcement Learning policy that utilizes many different aspects of the data and the task in order to select the most informative unlabeled subset dynamically over the course of the AL procedure. We demonstrate the superior

performance of the proposed RAL framework compared to strong AL baselines across four intricate multi-class, multi-label text classification datasets taken from specialised domains. In addition, we experiment with a unique data augmentation approach to further reduce the number of samples RAL needs to annotate.

AD-KD: Attribution-Driven Knowledge Distillation for Language Model Compression
Siyue Wu, Hongzhan Chen, Xiaojun Quan, Qifan Wang and Rui Wang 11:00-12:30 (Pier 7&8)
Knowledge distillation has attracted a great deal of interest recently to compress large language models. However, existing knowledge distillation methods suffer from two limitations. First, the student model simply imitates the teacher's behavior while ignoring the reasoning behind it. Second, these methods usually focus on the transfer of sophisticated model-specific knowledge but overlook data-specific knowledge. In this paper, we present a novel attribution-driven knowledge distillation approach, which explores the token-level rationale behind the teacher model based on Integrated Gradients (IG) and transfers attribution knowledge to the student model. To enhance the knowledge transfer of model reasoning and generalization, we further explore multi-view attribution distillation on all potential decisions of the teacher. Comprehensive experiments are conducted with BERT on the GLUE benchmark. The experimental results demonstrate the superior performance of our approach to several state-of-the-art methods.

ECOLA: Enhancing Temporal Knowledge Embeddings with Contextualized Language Representations
Zhen Han, Ruotong Liao, Jindong Gu, Yao Zhang, Zifeng Ding, Yujia Gu, Heinz Koeppl, Hinrich Schütze and Volker Tresp 11:00-12:30 (Pier 7&8)
Since conventional knowledge embedding models cannot take full advantage of the abundant textual information, there have been extensive research efforts in enhancing knowledge embedding using texts. However, existing enhancement approaches cannot apply to *temporal knowledge graphs* (tKGs), which contain time-dependent event knowledge with complex temporal dynamics. Specifically, existing enhancement approaches often assume knowledge embedding is time-independent. In contrast, the entity embedding in tKG models usually evolves, which poses the challenge of aligning *temporally relevant* texts with entities. To this end, we propose to study enhancing temporal knowledge embedding with textual data in this paper. As an approach to this task, we propose Enhanced Temporal Knowledge Embeddings with Contextualized Language Representations (ECOLA), which takes the temporal aspect into account and injects textual information into temporal knowledge embedding. To evaluate ECOLA, we introduce three new datasets for training and evaluating ECOLA. Extensive experiments show that ECOLA significantly enhances temporal KG embedding models with up to 287% relative improvements regarding Hits@1 on the link prediction task. The code and models are publicly available on https://github.com/mayhugotong/ECOLA.

What Makes Pre-trained Language Models Better Zero-shot Learners?
Jinghui Lu, dongsheng zhu, weidong Han, Rui Zhao, Brian Mac Namee and Fei Tan 11:00-12:30 (Pier 7&8)
Current methods for prompt learning in zero-shot scenarios widely rely on a development set with sufficient human-annotated data to select the best-performing prompt template a posteriori. This is not ideal because in a real-world zero-shot scenario of practical relevance, no labelled data is available. Thus, we propose a simple yet effective method for screening reasonable prompt templates in zero-shot text classification: Perplexity Selection (Perplection). We hypothesize that language discrepancy can be used to measure the efficacy of prompt templates, and thereby develop a substantiated perplexity-based scheme allowing for forecasting the performance of prompt templates in advance. Experiments show that our method leads to improved prediction performance in a realistic zero-shot setting, eliminating the need for any labelled examples.

On the Expressivity Role of LayerNorm in Transformers' Attention
Shaked Brody, Uri Alon and Eran Yahav 11:00-12:30 (Pier 7&8)
Layer Normalization (LayerNorm) is an inherent component in all Transformer-based models. In this paper, we show that LayerNorm is crucial to the expressivity of the multi-head attention layer that follows it. This is in contrast to the common belief that LayerNorm's only role is to normalize the activations during the forward pass, and their gradients during the backward pass.

We consider a geometric interpretation of LayerNorm and show that it consists of two components: (a) projection of the input vectors to a d-1 space that is orthogonal to the [1,1,...,1] vector, and (b) scaling of all vectors to the same norm of root d. We show that each of these components is important for the attention layer that follows it in Transformers: (a) projection allows the attention mechanism to create an attention query that attends to all keys equally, offloading the need to learn this operation in the attention; and (b) scaling allows each key to potentially receive the highest attention, and prevents keys from being "un-select-able". We show empirically that Transformers do indeed benefit from these properties of LayeNorm in general language modeling and even in computing simple functions such as "majority".

Peer-Label Assisted Hierarchical Text Classification
Junru Song, Feifei Wang and Yang Yang 11:00-12:30 (Pier 7&8)
Hierarchical text classification (HTC) is a challenging task, in which the labels of texts can be organized into a category hierarchy. To deal with the HTC problem, many existing works focus on utilizing the parent-child relationships that are explicitly shown in the hierarchy. However, texts with a category hierarchy also have some latent relevancy among labels in the same level of the hierarchy. We refer to these labels as peer labels, from which the peer effects are originally utilized in our work to improve the classification performance. To fully explore the peer-label relationship, we develop a PeerHTC method. This method innovatively measures the latent relevancy of peer labels through several metrics and then encodes the relevancy with a Graph Convolutional Neural Network. We also propose a sample importance learning method to ameliorate the side effects raised by modelling the peer label relevancy. Our experiments on several standard datasets demonstrate the evidence of peer labels and the superiority of PeerHTC over other state-of-the-art HTC methods in terms of classification accuracy.

Unsupervised Open-domain Keyphrase Generation
Lam Thanh Do, Pritom Saha Akash and Kevin Chen-Chuan Chang 11:00-12:30 (Pier 7&8)
In this work, we study the problem of unsupervised open-domain keyphrase generation, where the objective is a keyphrase generation model that can be built without using human-labeled data and can perform consistently across domains. To solve this problem, we propose a seq2seq model that consists of two modules, namely phraseness and informativeness module, both of which can be built in an unsupervised and open-domain fashion. The phraseness module generates phrases, while the informativeness module guides the generation towards those that represent the core concepts of the text. We thoroughly evaluate our proposed method using eight benchmark datasets from different domains. Results on in-domain datasets show that our approach achieves state-of-the-art results compared with existing unsupervised models, and overall narrows the gap between supervised and unsupervised methods down to about 16%. Furthermore, we demonstrate that our model performs consistently across domains, as it surpasses the baselines on out-of-domain datasets.

CFL: Causally Fair Language Models Through Token-level Attribute Controlled Generation
Rahul Madhavan, Rishabh Garg, Kahini Wadhawan and Sameep Mehta 11:00-12:30 (Pier 7&8)
We propose a method to control the attributes of Language Models (LMs) for the text generation task using Causal Average Treatment Effect (ATE) scores and counterfactual augmentation. We explore this method, in the context of LM detoxification, and propose the Causally Fair Language (CFL) architecture for detoxifying pre-trained LMs in a plug-and-play manner. Our architecture is based on a Structural Causal Model (SCM) that is mathematically transparent and computationally efficient as compared with many existing detoxification techniques.

We also propose several new metrics that aim to better understand the behaviour of LMs in the context of toxic text generation. Further, we achieve state of the art performance for toxic degeneration, which are computed using Real Toxicity Prompts. Our experiments show that CFL achieves such a detoxification without much impact on the model perplexity. We also show that CFL mitigates the unintended bias problem through experiments on the BOLD dataset.

Exploring Robust Overfitting for Pre-trained Language Models
Bin Zhu and Yanghui Rao   11:00-12:30 (Pier 7&8)
We identify the robust overfitting issue for pre-trained language models by showing that the robust test loss increases as the epoch grows. Through comprehensive exploration of the robust loss on the training set, we attribute robust overfitting to the model's memorization of the adversarial training data. We attempt to mitigate robust overfitting by combining regularization methods with adversarial training. Following the philosophy that prevents the model from memorizing the adversarial data, we find that flooding, a regularization method with loss scaling, can mitigate robust overfitting for pre-trained language models. Eventually, we investigate the effect of flooding levels and evaluate the models' adversarial robustness under textual attacks. Extensive experiments demonstrate that our methods can mitigate robust overfitting upon three top adversarial training methods and further promote adversarial robustness.

Gradient-based Intra-attention Pruning on Pre-trained Language Models
Ziqing Yang, Yiming Cui, Xin Yao and Shijin Wang   11:00-12:30 (Pier 7&8)
Pre-trained language models achieve superior performance but are computationally expensive. Techniques such as pruning and knowledge distillation have been developed to reduce their sizes and latencies. In this work, we propose a structured pruning method GRAIN (gradient-based intra-attention pruning), which performs task-specific pruning with knowledge distillation and yields highly effective models. Different from common approaches that prune each attention head as a whole, GRAIN inspects and prunes intra-attention structures, which greatly expands the structure search space and enables more flexible models. We also propose a gradient separation strategy that reduces the interference of distillation on pruning for a better combination of the two approaches. Experiments on GLUE, SQuAD, and CoNLL 2003 show that GRAIN notably outperforms other methods, especially in the high sparsity regime, and achieves 6 7x speedups while maintaining 93% 99% performance. Under extreme compression where only 3% transformer weights remain, the pruned model is still competitive compared to larger models.

Real-World Compositional Generalization with Disentangled Sequence-to-Sequence Learning
Hao Zheng and Mirella Lapata   11:00-12:30 (Pier 7&8)
Compositional generalization is a basic mechanism in human language learning, which current neural networks struggle with. A recently proposed Disentangled sequence-to-sequence model (Dangle) shows promising generalization capability by learning specialized encodings for each decoding step. We introduce two key modifications to this model which encourage more disentangled representations and improve its compute and memory efficiency, allowing us to tackle compositional generalization in a more realistic setting. Specifically, instead of adaptively re-encoding source keys and values at each time step, we disentangle their representations and only re-encode keys periodically, at some interval. Our new architecture leads to better generalization performance across existing tasks and datasets, and a new machine translation benchmark which we create by detecting naturally occurring compositional patterns in relation to a training set. We show this methodology better emulates real-world requirements than artificial challenges.

Teaching Small Language Models to Reason
Lucie Charlotte Magister, Jonathan Mallinson, Jakub Dominik Adamek, Eric Malmi and Aliaksei Severyn   11:00-12:30 (Pier 7&8)
Chain of thought prompting successfully improves the reasoning capabilities of large language models, achieving state of the art results on a range of datasets. However, these reasoning capabilities only appear to emerge in models with at least tens of billions of parameters. In this paper, we explore the transfer of such reasoning capabilities to smaller models via knowledge distillation, also investigating model and dataset size trade-off. Specifically, we finetune a student model on the chain of thought outputs generated by a larger teacher model. Our experiments show that the proposed method improves task performance across arithmetic, commonsense and symbolic reasoning datasets. For example, the accuracy of T5 XXL on GSM8K improves from 8.11% to 21.99% and 18.42% when finetuned on PaLM 540B and GPT-3 175B generated chains of thought, respectively.

B2T Connection: Serving Stability and Performance in Deep Transformers
Sho Takase, Shun Kiyono, Sosuke Kobayashi and Jun Suzuki   11:00-12:30 (Pier 7&8)
In the perspective of a layer normalization (LN) position, the architecture of Transformers can be categorized into two types: Post-LN and Pre-LN. Recent Transformers prefer to select Pre-LN because the training in Post-LN with deep Transformers, e.g., ten or more layers, often becomes unstable, resulting in useless models. However, in contrast, Post-LN has also consistently achieved better performance than Pre-LN in relatively shallow Transformers, e.g., six or fewer layers. This study first investigates the reason for these discrepant observations empirically and theoretically and discovers 1, the LN in Post-LN is the source of the vanishing gradient problem that mainly leads the unstable training whereas Pre-LN prevents it, and 2, Post-LN tends to preserve larger gradient norms in higher layers during the back-propagation that may lead an effective training. Exploiting the new findings, we propose a method that can equip both higher stability and effective training by a simple modification from Post-LN. We conduct experiments on a wide range of text generation tasks and demonstrate that our method outperforms Pre-LN, and stable training regardless of the shallow or deep layer settings.

Dialog-Post: Multi-Level Self-Supervised Objectives and Hierarchical Model for Dialogue Post-Training
Zhenyu Zhang, Lei Shen, Yuming Zhao, Meng Chen and Xiaodong He   11:00-12:30 (Pier 7&8)
Dialogue representation and understanding aim to convert conversational inputs into embeddings and fulfill discriminative tasks. Compared with free-form text, dialogue has two important characteristics, hierarchical semantic structure and multi-facet attributes. Therefore, directly applying the pretrained language models (PLMs) might result in unsatisfactory performance. Recently, several work focused on the dialogue-adaptive post-training (DialPost) that further trains PLMs to fit dialogues. To model dialogues more comprehensively, we propose a DialPost method, Dialog-Post, with multi-level self-supervised objectives and a hierarchical model. These objectives leverage dialogue-specific attributes and use self-supervised signals to fully facilitate the representation and understanding of dialogues. The novel model is a hierarchical segment-wise self-attention network, which contains inner-segment and inter-segment self-attention sub-layers followed by an aggregation and updating module. To evaluate the effectiveness of our methods, we first apply two public datasets for the verification of representation ability. Then we conduct experiments on a newly-labelled dataset that is annotated with 4 dialogue understanding tasks. Experimental results show that our method outperforms existing SOTA models and achieves a 3.3% improvement on average.

Contrastive Bootstrapping for Label Refinement
Shudi Hou, Yu Xia, Muhao Chen and Sujian Li   11:00-12:30 (Pier 7&8)
Traditional text classification typically categorizes texts into pre-defined coarse-grained classes, from which the produced models cannot handle the real-world scenario where finer categories emerge periodically for accurate services. In this work, we investigate the setting where fine-grained classification is done only using the annotation of coarse-grained categories and the coarse-to-fine mapping. We propose a lightweight contrastive clustering-based bootstrapping method to iteratively refine the labels of passages. During clustering, it pulls away

negative passage-prototype pairs under the guidance of the mapping from both global and local perspectives. Experiments on NYT and 20News show that our method outperforms the state-of-the-art methods by a large margin.

Using Domain Knowledge to Guide Dialog Structure Induction via Neural Probabilistic Soft Logic
Connor F Pryor, Quan Yuan, Jeremiah Liu, Mehran Kazemi, Deepak Ramachandran, Tania Bedrax-Weiss and Lise Getoor 11:00-12:30 (Pier 7&8)
Dialog Structure Induction (DSI) is the task of inferring the latent dialog structure (i.e., a set of dialog states and their temporal transitions) of a given goal-oriented dialog. It is a critical component for modern dialog system design and discourse analysis. Existing DSI approaches are often purely data-driven, deploy models that infer latent states without access to domain knowledge, underperform when the training corpus is limited/noisy, or have difficulty when test dialogs exhibit distributional shifts from the training domain. This work explores a neural-symbolic approach as a potential solution to these problems. We introduce Neural Probabilistic Soft Logic Dialogue Structure Induction (NEUPSL DSI), a principled approach that injects symbolic knowledge into the latent space of a generative neural model. We conduct a thorough empirical investigation on the effect of NEUPSL DSI learning on hidden representation quality, few-shot learning, and out-of-domain generalization performance. Over three dialog structure induction datasets and across unsupervised and semi-supervised settings for standard and cross-domain generalization, the injection of symbolic knowledge using NEUPSL DSI provides a consistent boost in performance over the canonical baselines.

Decoder Tuning: Efficient Language Understanding as Decoding
Ganqu CUI, Wentao Li, Ning Ding, Longtao Huang, Zhiyuan Liu and Maosong Sun 11:00-12:30 (Pier 7&8)
With the evergrowing sizes of pre-trained models (PTMs), it has been an emerging practice to only provide the inference APIs for users, namely model-as-a-service (MaaS) setting. To adapt PTMs with model parameters frozen, most current approaches focus on the input side, seeking powerful prompts to stimulate models for correct answers. However, we argue that input-side adaptation could be arduous due to the lack of gradient signals and they usually require thousands of API queries, resulting in high computation and time costs. Specifically, DecT first extracts prompt-stimulated output scores for initial predictions. On top of that, we train an additional decoder network on the output representations to incorporate posterior data knowledge. By gradient-based optimization, DecT can be trained within several seconds and requires only one PTM query per sample. Empirically, we conduct extensive natural language understanding experiments and show that DecT significantly outperforms state-of-the-art algorithms with a 200x speed-up. Our code is available at https://github.com/thunlp/DecT.

Free Lunch for Efficient Textual Commonsense Integration in Language Models
Wanyun Cui and Xingran Chen 11:00-12:30 (Pier 7&8)
Recent years have witnessed the emergence of textual commonsense knowledge bases, aimed at providing more nuanced and context-rich knowledge. The integration of external commonsense into language models has been shown to be a key enabler in advancing the state-of-the-art for a wide range of NLP tasks. However, incorporating textual commonsense descriptions is computationally expensive, as compared to encoding conventional symbolic knowledge. In this paper, we propose a method to improve its efficiency without modifying the model. Our idea is to group training samples with similar commonsense descriptions into a single batch, thus reusing the encoded description across multiple samples. We theoretically investigate this problem and demonstrate that its upper bound can be reduced to the classic *graph k-cut problem*. Consequently, we propose a spectral clustering-based algorithm to solve this problem. Extensive experiments illustrate that the proposed batch partitioning approach effectively reduces the computational cost while preserving performance. The efficiency improvement is more pronounced on larger datasets and on devices with more memory capacity, attesting to its practical utility for large-scale applications.

Fourier Transformer: Fast Long Range Modeling by Removing Sequence Redundancy with FFT Operator
Ziwei He, Meng Yang, Minwei Feng, Jingcheng Yin, Xinbing Wang, Jingwen Leng and Zhouhan Lin 11:00-12:30 (Pier 7&8)
The transformer model is known to be computationally demanding, and prohibitively costly for long sequences, as the self-attention module uses a quadratic time and space complexity with respect to sequence length. Many researchers have focused on designing new forms of self-attention or introducing new parameters to overcome this limitation, however a large portion of them prohibits the model to inherit weights from large pretrained models. In this work, the transformer's inefficiency has been taken care of from another perspective. We propose Fourier Transformer, a simple yet effective approach by progressively removing redundancies in hidden sequence using the ready-made Fast Fourier Transform (FFT) operator to perform Discrete Cosine Transformation (DCT). Fourier Transformer is able to significantly reduce computational costs while retain the ability to inherit from various large pretrained models. Experiments show that our model achieves state-of-the-art performances among all transformer-based models on the long-range modeling benchmark LRA with significant improvement in both speed and space. For generative seq-to-seq tasks including CNN/DailyMail and ELI5, by inheriting the BART weights our model outperforms the standard BART and other efficient models. Our code will be publicly available at https://github.com/LUMIA-Group/FourierTransformer

Cost-effective Distillation of Large Language Models
Sayantan Dasgupta, Trevor Cohn and Timothy Baldwin 11:00-12:30 (Pier 7&8)
Knowledge distillation (KD) involves training a small "student" model to replicate the strong performance of a high-capacity "teacher" model, enabling efficient deployment in resource-constrained settings. Top-performing methods tend to be task- or architecture-specific and lack generalizability. Several existing approaches require pretraining of the teacher on task-specific datasets, which can be costly for large and unstable for small datasets. Here we propose an approach for improving KD through a novel distillation loss agnostic to the task and model architecture. We successfully apply our method to the distillation of the BERT-base and achieve highly competitive results from the distilled student across a range of GLUE tasks, especially for tasks with smaller datasets.

Efficient Document Embeddings via Self-Contrastive Bregman Divergence Learning
Daniel Saggau, Mina Rezaei, Bernd Bischl and Ilias Chalkidis 11:00-12:30 (Pier 7&8)
Learning quality document embeddings is a fundamental problem in natural language processing (NLP), information retrieval (IR), recommendation systems, and search engines. Despite recent advances in the development of transformer-based models that produce sentence embeddings with self-contrastive learning, the encoding of long documents (Ks of words) is still challenging with respect to both efficiency and quality considerations. Therefore, we train Longfomer-based document encoders using a state-of-the-art unsupervised contrastive learning method (SimCSE). Further on, we complement the baseline method -siamese neural network- with additional convex neural networks based on functional Bregman divergence aiming to enhance the quality of the output document representations. We show that overall the combination of a self-contrastive siamese network and our proposed neural Bregman network outperforms the baselines in two linear classification settings on three long document topic classification tasks from the legal and biomedical domains.

Semi-Supervised Domain Adaptation for Emotion-Related Tasks
Mahshid Hosseini and Cornelia Caragea 11:00-12:30 (Pier 7&8)
Semi-supervised domain adaptation (SSDA) adopts a model trained from a label-rich source domain to a new but related domain with a few labels of target data. It is shown that, in an SSDA setting, a simple combination of domain adaptation (DA) with semi-supervised learning (SSL) techniques often fails to effectively utilize the target supervision and cannot address distribution shifts across different domains due to the training data bias toward the source-labeled samples. In this paper, inspired by the co-learning of multiple classifiers for the computer

vision tasks, we propose to decompose the SSDA framework for emotion-related tasks into two subcomponents of unsupervised domain adaptation (UDA) from the source to the target domain and semi-supervised learning (SSL) in the target domain where the two models iteratively teach each other by interchanging their high confident predictions. We further propose a novel data cartography-based regularization technique for pseudo-label denoising that employs training dynamics to further hone our models' performance. We publicly release our code.

Multilingual Pre-training with Self-supervision from Global Co-occurrence Information
Xi Ai and Bin Fang   11:00-12:30 (Pier 7&8)
Global co-occurrence information is the primary source of structural information on multilingual corpora, and we find that analogical/parallel compound words across languages have similar co-occurrence counts/frequencies (normalized) giving weak but stable self-supervision for cross-lingual transfer. Following the observation, we aim at associating contextualized representations with relevant (contextualized) representations across languages with the help of co-occurrence counts. The result is MLM-GC (MLM with Global Co-occurrence) pre-training that the model learns local bidirectional information from MLM and global co-occurrence information from a log-bilinear regression. Experiments show that MLM-GC pre-training substantially outperforms MLM pre-training for 4 downstream cross-lingual tasks and 1 additional monolingual task, showing the advantages of forming isomorphic spaces across languages.

Learning to Initialize: Can Meta Learning Improve Cross-task Generalization in Prompt Tuning?
Chengwei Qin, Shafiq Joty, Qian Li and Ruochen Zhao   11:00-12:30 (Pier 7&8)
Prompt tuning (PT) which only tunes the embeddings of an additional sequence of tokens per task, keeping the pre-trained language model (PLM) frozen, has shown remarkable performance in few-shot learning. Despite this, PT has been shown to rely heavily on good initialization of the prompt embeddings. In this work, we study meta prompt tuning (MPT) to systematically explore how meta-learning can help improve (if it can) cross-task generalization in PT through learning to initialize the prompt embeddings from other relevant tasks. We empirically analyze a representative set of meta learning algorithms in a wide range of adaptation settings with different source/target task configurations on a large set of few-shot tasks. With extensive experiments and analysis, we demonstrate the effectiveness of MPT. We find the improvement to be significant particularly on classification tasks. For other kinds of tasks such as question answering, we observe that while MPT can outperform PT in most cases, it does not always outperform multi-task learning. We further provide an in-depth analysis from the perspective of task similarity.

From Characters to Words: Hierarchical Pre-trained Language Model for Open-vocabulary Language Understanding
Li Sun, Florian Luisier, Kayhan Batmanghelich, Dinei Florencio and Cha Zhang   11:00-12:30 (Pier 7&8)
Current state-of-the-art models for natural language understanding require a preprocessing step to convert raw text into discrete tokens. This process known as tokenization relies on a pre-built vocabulary of words or sub-word morphemes. This fixed vocabulary limits the model's robustness to spelling errors and its capacity to adapt to new domains. In this work, we introduce a novel open-vocabulary language model that adopts a hierarchical two-level approach: one at the word level and another at the sequence level. Concretely, we design an intra-word module that uses a shallow Transformer architecture to learn word representations from their characters, and a deep inter-word Transformer module that contextualizes each word representation by attending to the entire word sequence. Our model thus directly operates on character sequences with explicit awareness of word boundaries, but without biased sub-word or word-level vocabulary. Experiments on various downstream tasks show that our method outperforms strong baselines. We also demonstrate that our hierarchical model is robust to textual corruption and domain shift.

Beyond Positive Scaling: How Negation Impacts Scaling Trends of Language Models
Yuhui Zhang, Michihiro Yasunaga, Zhengping Zhou, Jeff Z. HaoChen, James Zou, Percy Liang and Serena Yeung   11:00-12:30 (Pier 7&8)
Language models have been shown to exhibit positive scaling, where performance improves as models are scaled up in terms of size, compute, or data. In this work, we introduce NeQA, a dataset consisting of questions with negation in which language models do not exhibit straightforward positive scaling. We show that this task can exhibit inverse scaling, U-shaped scaling, or positive scaling, and the three scaling trends shift in this order as we use more powerful prompting methods or model families. We hypothesize that solving NeQA depends on two subtasks: question answering (task 1) and negation understanding (task 2). We find that task 1 has linear scaling, while task 2 has sigmoid-shaped scaling with an emergent transition point, and composing these two scaling trends yields the final scaling trend of NeQA. Our work reveals and provides a way to analyze the complex scaling trends of language models.

How Do In-Context Examples Affect Compositional Generalization?
Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang LOU and Dongmei Zhang   11:00-12:30 (Pier 7&8)
Compositional generalization–understanding unseen combinations of seen primitives–is an essential reasoning capability in human intelligence. The AI community mainly studies this capability by fine-tuning neural networks on lots of training samples, while it is still unclear whether and how in-context learning–the prevailing few-shot paradigm based on large language models–exhibits compositional generalization. In this paper, we present CoFe, a test suite to investigate in-context compositional generalization. We find that the compositional generalization performance can be easily affected by the selection of in-context examples, thus raising the research question what the key factors are to make good in-context examples for compositional generalization. We study three potential factors: similarity, diversity and complexity. Our systematic experiments indicate that in-context examples should be structurally similar to the test case, diverse from each other, and individually simple. Furthermore, two strong limitations are observed: in-context compositional generalization on fictional words is much weaker than that on commonly used words; it is still critical that the in-context examples should cover required linguistic structures, even though the backbone model has been pre-trained on large corpus. We hope our analysis would facilitate the understanding and utilization of in-context learning paradigm.

Nonparametric Masked Language Modeling
Sewon Min, Weijia Shi, Mike Lewis, Xilun Chen, Wen-tau Yih, Hannaneh Hajishirzi and Luke Zettlemoyer   11:00-12:30 (Pier 7&8)
Existing language models (LMs) predict tokens with a softmax over a finite vocabulary, which can make it difficult to predict rare tokens or phrases. We introduce NPM, the first nonparametric masked language model that replaces this softmax with a nonparametric distribution over every phrase in a reference corpus. NPM fills in the [MASK] solely from retrieving a token from a text corpus. We show that NPM can be efficiently trained with a contrastive objective and an in-batch approximation to full corpus retrieval. Zero-shot evaluation on 16 tasks including classification, fact probing and question answering demonstrates that NPM outperforms significantly larger parametric models, either with or without a retrieve-and-generate approach. It is particularly better at dealing with rare patterns (word senses or facts) and predicting rare or nearly unseen words (e.g., non-Latin script). We release the model and code at github.com/facebookresearch/NPM.

Parameter-efficient Weight Ensembling Facilitates Task-level Knowledge Transfer
Xingtai Lv, Ning Ding, Yujia Qin, Zhiyuan Liu and Maosong Sun   11:00-12:30 (Pier 7&8)
Recent studies show that large-scale pre-trained language models could be efficaciously adapted to particular tasks in a parameter-efficient manner. The trained lightweight set of parameters, such as adapters, can be easily stored and shared as a capability equipped with the corresponding models. Owning many lightweight parameters, we focus on transferring them between tasks to acquire an improvement in performance of new tasks, the key point of which is to obtain the similarity between tasks. In this paper, we explore 5 parameter-efficient weight

ensembling methods to achieve such transferability and verify the effectiveness of them. These methods extract the information of datasets and trained lightweight parameters from different perspectives to obtain the similarity between tasks, and weight the existing lightweight parameters according to the comparability to acquire a suitable module for the initialization of new tasks. We apply them to three parameter-efficient tuning methods and test them on a wide set of downstream tasks. Experimental results show that our methods show an improvement of 5%~8% over baselines and could largely facilitate task-level knowledge transfer.

Mixture-of-Domain-Adapters: Decoupling and Injecting Domain Knowledge to Pre-trained Language Models' Memories
Shizhe Diao, Tianyang Xu, Ruijia Xu, Jiawei Wang and Tong Zhang   11:00-12:30 (Pier 7&8)
Pre-trained language models (PLMs) demonstrate excellent abilities to understand texts in the generic domain while struggling in a specific domain. Although continued pre-training on a large domain-specific corpus is effective, it is costly to tune all the parameters on the domain. In this paper, we investigate whether we can adapt PLMs both effectively and efficiently by only tuning a few parameters. Specifically, we decouple the feed-forward networks (FFNs) of the Transformer architecture into two parts: the original pre-trained FFNs to maintain the old-domain knowledge and our novel domain-specific adapters to inject domain-specific knowledge in parallel. Then we adopt a mixture-of-adapters gate to fuse the knowledge from different domain adapters dynamically. Our proposed Mixture-of-Domain-Adapters (MixDA) employs a two-stage adapter-tuning strategy that leverages both unlabeled data and labeled data to help the domain adaptation: i) domain-specific adapter on unlabeled data; followed by ii) the task-specific adapter on labeled data. MixDA can be seamlessly plugged into the pretraining-finetuning paradigm and our experiments demonstrate that MixDA achieves superior performance on in-domain tasks (GLUE), out-of-domain tasks (ChemProt, RCT, IMDB, Amazon), and knowledge-intensive tasks (KILT). Further analyses demonstrate the reliability, scalability, and efficiency of our method.

Controlling the Extraction of Memorized Data from Large Language Models via Prompt-Tuning
Mustafa Safa Ozdayi, Charith Peris, Jack FitzGerald, Christophe Dupuy, Jimit Majmudar, Haidar Khan, Rahil Parikh and Rahul Gupta
11:00-12:30 (Pier 7&8)
Large Language Models (LLMs) are known to memorize significant portions of their training data. Parts of this memorized content have been shown to be extractable by simply querying the model, which poses a privacy risk. We present a novel approach which uses prompt-tuning to control the extraction rates of memorized content in LLMs. We present two prompt training strategies to increase and decrease extraction rates, which correspond to an attack and a defense, respectively. We demonstrate the effectiveness of our techniques by using models from the GPT-Neo family on a public benchmark. For the 1.3B parameter GPT-Neo model, our attack yields a 9.3 percentage point increase in extraction rate compared to our baseline. Our defense can be tuned to achieve different privacy-utility trade-offs by a user-specified hyperparameter. We achieve an extraction rate reduction of up to 97.7% relative to our baseline, with a perplexity increase of 16.9%.

Z-Code++: A Pre-trained Language Model Optimized for Abstractive Summarization
Pengcheng He, Baolin Peng, Song Wang, Yang Liu, Ruochen Xu, Hany Hassan, Yu Shi, Chenguang Zhu, Wayne Xiong, Michael Zeng, Jianfeng Gao and Xuedong Huang   11:00-12:30 (Pier 7&8)
This paper presents Z-Code++, a new pre-trained language model optimized for abstractive text summarization. The model extends the state-of-the-art encoder-decoder model using three techniques. First, we use a two-phase pre-training to improve the model's performance on low-resource summarization tasks. The model is first pre-trained using text corpora for language understanding, then is continually pre-trained on summarization corpora for grounded text generation. Second, we replace self-attention layers in the encoder with disentangled attention layers, where each word is represented using two vectors that encode its content and position, respectively. Third, we use fusion-in-encoder, a simple yet effective method of encoding long sequences in a hierarchical manner. Z-Code++ creates a new state-of-the-art on 9 of 13 text summarization tasks across 5 languages. Our model is parameter-efficient in that it outperforms the 600x larger PaLM540B on XSum, and the finetuned 200x larger GPT3175B on SAMSum. In zero-shot and few-shot settings, our model substantially outperforms the competing models.

Recipes for Sequential Pre-training of Multilingual Encoder and Seq2Seq Models
Saleh Soltan, Andy Rosenbaum, Tobias Falke, Qin Lu, Anna Rumshisky and Wael Hamza   11:00-12:30 (Pier 7&8)
Pre-trained encoder-only and sequence-to-sequence (seq2seq) models each have advantages, however training both model types from scratch is computationally expensive. We explore recipes to improve pre-training efficiency by initializing one model from the other. (1) Extracting the encoder from a seq2seq model, we show it under-performs a Masked Language Modeling (MLM) encoder, particularly on sequence labeling tasks. Variations of masking during seq2seq training, reducing the decoder size, and continuing with a small amount of MLM training do not close the gap. (2) Conversely, using an encoder to warm-start seq2seq training, we show that by unfreezing the encoder partway through training, we can match task performance of a from-scratch seq2seq model. Overall, this two-stage approach is an efficient recipe to obtain both a multilingual encoder and a seq2seq model, matching the performance of training each model from scratch while reducing the total compute cost by 27%.

Residual Prompt Tuning: improving prompt tuning with residual reparameterization
Anastasiia Razdaibiedina, Yuning Mao, Madian Khabsa, Mike Lewis, Rui Hou, Jimmy Ba and Amjad Almahairi   11:00-12:30 (Pier 7&8)
Prompt tuning is one of the successful approaches for parameter-efficient tuning of pre-trained language models. Despite being arguably the most parameter-efficient (tuned soft prompts constitute <0.1% of total parameters), it typically performs worse than other efficient tuning methods and is quite sensitive to hyper-parameters. In this work, we introduce Residual Prompt Tuning - a simple and efficient method that significantly improves the performance and stability of prompt tuning. We propose to reparameterize soft prompt embeddings using a shallow network with a residual connection. Our experiments show that Residual Prompt Tuning significantly outperforms prompt tuning across T5-Large, T5-Base and BERT-Base models. Notably, our method reaches +7 points improvement over prompt tuning on SuperGLUE benchmark with T5-Base model and allows to reduce the prompt length by 10 times without hurting performance. In addition, we show that our approach is robust to the choice of learning rate and prompt initialization, and is effective in few-shot settings.

Better Zero-Shot Reasoning with Self-Adaptive Prompting
Xingchen Wan, Ruoxi SUN, Hanjun Dai, Sercan Arik and Tomas Pfister   11:00-12:30 (Pier 7&8)
Modern large language models (LLMs) have demonstrated impressive capabilities at sophisticated tasks, often through step-by-step reasoning similar to humans. This is made possible by their strong few- and zero-shot abilities – they can effectively learn from a handful of handcrafted, completed responses ("in-context examples"), or are prompted to reason spontaneously through specially designed triggers. Nonetheless, some limitations have been observed. First, performance in the few-shot setting is sensitive to the choice of the examples, whose design requires significant human effort. Moreover, given the diverse downstream tasks of LLMs, it may be difficult or laborious to handcraft per-task labels. Second, while the zero-shot setting does not require handcrafting, its performance is limited due to the lack of guidance to the LLMs. To address these limitations, we propose Consistency-based Self-adaptive Prompting (COSP), a novel prompt design method for LLMs. Requiring neither handcrafted responses nor ground-truth labels, COSP selects and builds the set of examples from the LLM zero-shot outputs via carefully designed criteria combining consistency, diversity and repetition. In the zero-shot setting for three different LLMs, we show that using only LLM predictions, COSP significantly improves performance up to 15% compared to zero-shot baselines and matches or exceeds few-shot baselines at a range of reasoning tasks.

Let Me Check the Examples: Enhancing Demonstration Learning via Explicit Imitation
Sirui Wang, Kaiwen Wei, Hongzhi Zhang, Yuntao Li and Wei Wu   11:00-12:30 (Pier 7&8)
Demonstration learning aims to guide the prompt prediction by providing answered demonstrations in the few shot settings. Despite achieving promising results, existing work only concatenates the answered examples as demonstrations to the prompt template (including the raw context) without any additional operation, neglecting the prompt-demonstration dependencies. Besides, prior research found that randomly replacing the labels of demonstrations marginally hurts performance, illustrating that the model could not properly learn the knowledge brought by the demonstrations. Inspired by the human learning process, in this paper, we introduce Imitation DEMOnstration learning (Imitation-Demo) to strengthen demonstration learning via explicitly imitating human review behaviour, which includes: (1) contrastive learning mechanism to concentrate on similar demonstrations.(2) demonstration-label re-prediction method to consolidate known knowledge. Experiment results show that our proposed method achieves state-of-the-art performance on 5 out of 14 classification corpus. Further studies also prove that Imitation-Demo strengthens the associations between the prompt and demonstrations, which could provide the basis for exploring how demonstration learning works.

Masked Latent Semantic Modeling: an Efficient Pre-training Alternative to Masked Language Modeling
Gábor Berend   11:00-12:30 (Pier 7&8)
In this paper, we propose an alternative to the classic masked language modeling (MLM) pre-training paradigm, where the objective is altered from the reconstruction of the exact identity of randomly selected masked subwords to the prediction of their latent semantic properties. We coin the proposed pre-training technique masked latent semantic modeling (MLSM for short). In order to make the contextualized determination of the latent semantic properties of the masked subwords possible, we rely on an unsupervised technique which uses sparse coding. Our experimental results reveal that the fine-tuned performance of those models that we pre-trained via MLSM is consistently and significantly better compared to the use of vanilla MLM pretraining and other strong baselines.

Knowledgeable Parameter Efficient Tuning Network for Commonsense Question Answering
Ziwang Zhao, Linmei Hu, Hanyu Zhao, Yingxia Shao and Yequan Wang   11:00-12:30 (Pier 7&8)
Commonsense question answering is important for making decisions about everyday matters. Although existing commonsense question answering works based on fully fine-tuned PLMs have achieved promising results, they suffer from prohibitive computation costs as well as poor interpretability. Some works improve the PLMs by incorporating knowledge to provide certain evidence, via elaborately designed GNN modules which require expertise. In this paper, we propose a simple knowledgeable parameter efficient tuning network to couple PLMs with external knowledge for commonsense question answering. Specifically, we design a trainable parameter-sharing adapter attached to a parameter-freezing PLM to incorporate knowledge at a small cost. The adapter is equipped with both entity- and query-related knowledge via two auxiliary knowledge-related tasks (i.e., span masking and relation discrimination). To make the adapter focus on the relevant knowledge, we design gating and attention mechanisms to respectively filter and fuse the query information from the PLM. Extensive experiments on two benchmark datasets show that KPE is parameter-efficient and can effectively incorporate knowledge for improving commonsense question answering.

Improved Logical Reasoning of Language Models via Differentiable Symbolic Programming
Hanlin Zhang, Jiani Huang, Ziyang Li, MAYUR NAIK and Eric Xing   11:00-12:30 (Pier 7&8)
Pre-trained large language models (LMs) struggle to perform logical reasoning reliably despite advances in scale and compositionality. In this work, we tackle this challenge through the lens of symbolic programming. We propose DSR-LM, a Differentiable Symbolic Reasoning framework where pre-trained LMs govern the perception of factual knowledge, and a symbolic module performs deductive reasoning. In contrast to works that rely on hand-crafted logic rules, our differentiable symbolic reasoning framework efficiently learns weighted rules and applies semantic loss to further improve LMs. DSR-LM is scalable, interpretable, and allows easy integration of prior knowledge, thereby supporting extensive symbolic programming to robustly derive a logical conclusion. The results of our experiments suggest that DSR-LM improves the logical reasoning abilities of pre-trained language models, resulting in a significant increase in accuracy of over 20% on deductive reasoning benchmarks. Furthermore, DSR-LM outperforms a variety of competitive baselines when faced with systematic changes in sequence length.

Reasoning in Large Language Models Through Symbolic Math Word Problems
Vedant Gaur and Nikunj Saunshi   11:00-12:30 (Pier 7&8)
Large language models (LLMs) have revolutionized NLP by solving downstream tasks with little to no labeled data. Despite their versatile abilities, the larger question of their ability to reason remains ill-understood. This paper addresses reasoning in math word problems (MWPs) by studying symbolic versions of the numeric problems, since a symbolic expression is a "concise explanation" of the numeric answer. We create and use a symbolic version of the SVAMP dataset and find that GPT-3's davinci-002 model also has good zero-shot accuracy on symbolic MWPs. To evaluate the faithfulness of the model's reasoning, we go beyond accuracy and additionally evaluate the alignment between the final answer and the outputted reasoning, which correspond to numeric and symbolic answers respectively for MWPs. We explore a self-prompting approach to encourage the symbolic reasoning to align with the numeric answer, thus equipping the LLM with the ability to provide a concise and verifiable reasoning and making it more interpretable. Surprisingly, self-prompting also improves the symbolic accuracy to be higher than both the numeric and symbolic accuracies, thus providing an ensembling effect. The SVAMP-Sym dataset will be released for future research on symbolic math problems.

Honey, I Shrunk the Language: Language Model Behavior at Reduced Scale.
Vijeta Deshpande, Dan Pechi, Shree Thatte, Vladislav Lialin and Anna Rumshisky   11:00-12:30 (Pier 7&8)
In recent years, language models have drastically grown in size, and the abilities of these models have been shown to improve with scale. The majority of recent scaling laws studies focused on high-compute high-parameter count settings, leaving the question of when these abilities begin to emerge largely unanswered. In this paper, we investigate whether the effects of pre-training can be observed when the problem size is reduced, modeling a smaller, reduced-vocabulary language. We show the benefits of pre-training with masked language modeling (MLM) objective in models as small as 1.25M parameters, and establish a strong correlation between pre-training perplexity and downstream performance (GLUE benchmark). We examine downscaling effects, extending scaling laws to models as small as 1M parameters. At this scale, we observe a break of the power law for compute-optimal models and show that the MLM loss does not scale smoothly with compute-cost (FLOPs) below 2.2×10^{15} FLOPs. We also find that adding layers does not always benefit downstream performance.Our filtered pre-training data, reduced English vocabulary, and code are available at https://github.com/text-machine-lab/mini_bert$github.com/text-machine-lab/mini\_bert$

Gradient Ascent Post-training Enhances Language Model Generalization
Dongkeun Yoon, Joel Jang, Sungdong Kim and Minjoon Seo   11:00-12:30 (Pier 7&8)
In this work, we empirically show that updating pretrained LMs (350M, 1.3B, 2.7B) with just a few steps of Gradient Ascent Post-training (GAP) on random, unlabeled text corpora enhances its zero-shot generalization capabilities across diverse NLP tasks. Specifically, we show that GAP can allow LMs to become comparable to 2-3x times larger LMs across 12 different NLP tasks. We also show that applying GAP on

out-of-distribution corpora leads to the most reliable performance improvements. Our findings indicate that GAP can be a promising method for improving the generalization capability of LMs without any task-specific fine-tuning.

MVP: Multi-task Supervised Pre-training for Natural Language Generation
Tianyi Tang, Junyi Li, Wayne Xin Zhao and Ji-Rong Wen   11:00-12:30 (Pier 7&8)
Pre-trained language models (PLMs) have achieved remarkable success in natural language generation (NLG) tasks. Up to now, most NLG-oriented PLMs are pre-trained in an unsupervised manner using the large-scale general corpus. In the meanwhile, an increasing number of models pre-trained with labeled data (i.e. "supervised pre-training") showcase superior performance compared to unsupervised pre-trained models. Motivated by the success of supervised pre-training, we propose Multi-task superVised Pre-training (MVP) for natural language generation. We collect a large-scale natural language generation corpus, MVPCorpus, from 77 datasets over 11 diverse NLG tasks. Then we unify these examples into a general text-to-text format to pre-train the text generation model MVP in a supervised manner. For each task, we further pre-train specific soft prompts to stimulate the model's capacity to perform a specific task. Our MVP model can be seen as a practice that utilizes recent instruction tuning on relatively small PLMs. Extensive experiments have demonstrated the effectiveness and generality of our MVP model in a number of NLG tasks, which achieves state-of-the-art performance on 13 out of 17 datasets, outperforming BART by 9.3% and Flan-T5 by 5.8%.

Teaching the Pre-trained Model to Generate Simple Texts for Text Simplification
Renliang Sun, Wei Xu and Xiaojun Wan   11:00-12:30 (Pier 7&8)
Randomly masking text spans in ordinary texts in the pre-training stage hardly allows models to acquire the ability to generate simple texts. It can hurt the performance of pre-trained models on text simplification tasks. In this paper, we propose a new continued pre-training strategy to teach the pre-trained model to generate simple texts. We continue pre-training BART, a representative model, to obtain SimpleBART. It consistently and significantly improves the results on lexical simplification, sentence simplification, and document-level simplification tasks over BART. At the end, we compare SimpleBART with several representative large language models (LLMs).

Contrastive Decoding: Open-ended Text Generation as Optimization
Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer and Mike Lewis 11:00-12:30 (Pier 7&8)
Given a language model (LM), maximum probability is a poor decoding objective for open-ended generation, because it produces short and repetitive text. On the other hand, sampling can often produce incoherent text that drifts from the original topics. We propose contrastive decoding (CD), a reliable decoding approach that optimizes a contrastive objective subject to a plausibility constraint. The contrastive objective returns the difference between the likelihood under a large LM (called the expert, e.g. OPT-13B) and a small LM (called the amateur, e.g. OPT-125M), and the constraint ensures that the outputs are plausible. CD is inspired by the fact that the failures of larger LMs (e.g., repetition, inco- herence) are even more prevalent in smaller LMs, and that this difference signals which texts should be preferred. CD requires zero additional training, and produces higher quality text than decoding from the larger LM alone. It also works across model scales (OPT-13B and GPT2-1.5B) and significantly outperforms four strong decoding algorithms (e.g., nucleus, top-k) in automatic and human evaluations across wikipedia, news and story domains.

Fantastic Expressions and Where to Find Them: Chinese Simile Generation with Multiple Constraints
Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Xiangpeng Wei, Zhengyuan Liu and Jun Xie   11:00-12:30 (Pier 7&8)
Similes occur in the creative context of describing a concept (i.e., tenor) by making a literally false yet figuratively meaningful comparison to another (i.e., vehicle). Previous efforts form simile generation as a context-free generation task, focusing on simile-style transfer or writing a simile from a given prefix. However, generated texts under such settings might be undesirable, such as hardly meeting the simile definition (e.g., missing vehicle) or difficult to address certain preferences of content as humans wish (e.g., describe the color of apples through the simile). We believe that a simile could be more qualified and user-oriented if incorporated with pre-specified constraints. To this end, we introduce controllable simile generation (CSG), a new task that requires the model to generate a simile with multiple simile elements, e.g., context and vehicle. To facilitate this task, we present GraCe, including 61.3k simile-element annotated Chinese similes. Based on it, we propose a CSG model Similor to benchmark this task, including a vehicle retrieval module Scorer to obtain the explicable comparison for a given tenor in the vehicle-unknown situation. Both statistical and experimental analyses show that GraCe is of high quality beyond all other Chinese simile datasets, in terms of the number (8 vs. 3) of annotation elements, Is-Simile accuracy (98.9% vs. 78.7%), and increasing model-performance gains for both uncontrollable and controllable simile generation. Meanwhile, Similor can serve as a strong baseline for CSG, especially with Scorer, which beats model-based retrieval methods without any re-training.

Nano: Nested Human-in-the-Loop Reward Learning for Few-shot Language Model Control
Xiang Fan, Yiwei Lyu, Paul Pu Liang, Ruslan Salakhutdinov and Louis-Philippe Morency   11:00-12:30 (Pier 7&8)
Pretrained language models have demonstrated extraordinary capabilities in language generation. However, real-world tasks often require controlling the distribution of generated text in order to mitigate bias, promote fairness, and achieve personalization. Existing techniques for controlling the distribution of generated text only work with quantified distributions, which require pre-defined categories, proportions of the distribution, or an existing corpus following the desired distributions. However, many important distributions, such as personal preferences, are unquantified. In this work, we tackle the problem of generating text following arbitrary distributions (quantified and unquantified) by proposing NANO, a few-shot human-in-the-loop training algorithm that continuously learns from human feedback. NANO achieves state-of-the-art results on single topic/attribute as well as quantified distribution control compared to previous works. We also show that NANO is able to learn unquantified distributions, achieves personalization, and captures differences between different individuals' personal preferences with high sample efficiency.

Differentiable Instruction Optimization for Cross-Task Generalization
Masaru Isonuma, Junichiro Mori and Ichiro Sakata   11:00-12:30 (Pier 7&8)
Instruction tuning has been attracting much attention to achieve generalization ability across a wide variety of tasks. Although various types of instructions have been manually created for instruction tuning, it is still unclear what kind of instruction is optimal to obtain cross-task generalization ability. This work presents instruction optimization, which optimizes training instructions with respect to generalization ability. Rather than manually tuning instructions, we introduce learnable instructions and optimize them with gradient descent by leveraging bilevel optimization. Experimental results show that the learned instruction enhances the diversity of instructions and improves the generalization ability compared to using only manually created instructions.

PREADD: Prefix-Adaptive Decoding for Controlled Text Generation
Jonathan Pei, Kevin Yang and Dan Klein   11:00-12:30 (Pier 7&8)
We propose Prefix-Adaptive Decoding (PREADD), a flexible method for controlled text generation. Unlike existing methods that use aux-iliary expert models to control for attributes, PREADD does not require an external model, instead relying on linearly combining output logits from multiple prompts. Specifically, PREADD contrasts the output logits generated using a raw prompt against those generated using a prefix-prepended prompt, enabling both positive and negative control with respect to any attribute encapsulated by the prefix. We evaluate

PREADD on three tasks—toxic output mitigation, gender bias reduction, and sentiment control—and find that PREADD outperforms not only prompting baselines, but also an auxiliary-expert control method, by 12% or more in relative gain on our main metrics for each task.

Efficient Out-of-Domain Detection for Sequence to Sequence Models
Artem Vazhentsev, Akim Tsvigun, Roman Konstantinovich Vashurin, Sergey Petrakov, Daniil Vasilev, Maxim Panov, Alexander Panchenko and Artem Shelmanov 11:00-12:30 (Pier 7&8)
Sequence-to-sequence (seq2seq) models based on the Transformer architecture have become a ubiquitous tool applicable not only to classical text generation tasks such as machine translation and summarization but also to any other task where an answer can be represented in a form of a finite text fragment (e.g., question answering). However, when deploying in practice, we need not only high performance but also an ability to determine cases where the model is not applicable. Uncertainty estimation (UE) techniques provide a tool for identifying out-of-domain (OOD) input where the model is susceptible to errors. State-of-the-art UE methods for seq2seq models rely on computationally heavyweight and impractical deep ensembles. In this work, we perform an empirical investigation of various novel UE methods for large pre-trained seq2seq models T5 and BART on three tasks: machine translation, text summarization, and question answering. We apply computationally lightweight density-based UE methods to seq2seq models and show that they often outperform heavyweight deep ensembles on the task of OOD detection.

Best-k Search Algorithm for Neural Text Generation
Jiacheng Xu, Caiming Xiong, silvio savarese and Yingbo Zhou 11:00-12:30 (Pier 7&8)
Modern natural language generation paradigms require a decoding strategy to obtain quality sequences out of the model. Beam search yields high-quality but low diversity outputs; stochastic approaches suffer from high variance and sometimes low quality. In this work, we propose a deterministic search algorithm balancing both quality and diversity. We first investigate the vanilla best-first search (BFS) algorithm and then propose the best-k search algorithm. Inspired by BFS, we greedily expand the top k nodes, instead of the first node, to boost efficiency and diversity. Upweighting recently discovered nodes accompanied by heap pruning ensures the completeness of the search procedure. Experiments on four NLG tasks show that best-k search yields more diverse and natural outputs compared to strong baselines, while our approach maintains high text quality. The proposed algorithm is parameter-free, lightweight, efficient, and easy-to-use.

Few-Shot Data-to-Text Generation via Unified Representation and Multi-Source Learning
Alexander Hanbo Li, Mingyue Shang, Evangelia Spiliopoulou, Jie Ma, Patrick Ng, Zhiguo Wang, Bonan Min, William Yang Wang, Kathleen McKeown, Vittorio Castelli, Dan Roth and Bing Xiang 11:00-12:30 (Pier 7&8)
In this paper, we present a novel approach for data-to-text generation that addresses the limitations of current methods that primarily focus on specific types of structured data. Our proposed method aims to improve performance in multi-task training, zero-shot and few-shot scenarios by providing a unified representation that can handle various forms of structured data such as tables, knowledge graph triples, and meaning representations. We demonstrate that our proposed approach can effectively adapt to new structured forms, and can improve performance in comparison to current methods. For example, our method resulted in a 66% improvement in zero-shot BLEU scores when transferring models trained on table inputs to a knowledge graph dataset. Our proposed method is an important step towards a more general data-to-text generation framework.

Dynamic and Efficient Inference for Text Generation via BERT Family
Xiaobo Liang, Juntao Li, Lijun Wu, Ziqiang Cao and Min Zhang 11:00-12:30 (Pier 7&8)
Despite the excellent performance of Pre-trained Language Models on many text generation tasks, they suffer from inefficient inference on computation and memory due to their large-scale parameters and the universal autoregressive decoding paradigm. In this work, we propose a novel fine-tuning method **DEER**, which can make a single pre-trained model support **D**ynamic and **E**fficient inf**ER**ence and achieve an adaptive trade-off between model performance and latency. In particular, our critical insight is to jointly utilize the non-autoregressive (NAR) generation and dynamic parameter pruning techniques, which can flexibly control the decoding iteration steps and model sizes according to memory and latency limitations. Besides, we also explore the effectiveness of the pre-trained MLMs (i.e., the BERT family) for text generation tasks since their bidirectional attention nature is more suitable for the NAR training objective. Extensive experiments on both monolingual and multilingual pre-trained MLMs demonstrate the effectiveness of our proposed DEER method by consistently achieving (1) higher BLEU scores than the strong autoregressive Transformer model on three neural machine translation tasks with $3 \rightarrow 12$ times speedup, (2) competitive performance (but with much faster inference speed) compared with the BART model on four GLGE benchmark tasks. Our code will be publicly available at GitHubhttps://github.com/dropreg/DEER.

Distilling Reasoning Capabilities into Smaller Language Models
Kumar Shridhar, Alessandro Stolfo and Mrinmaya Sachan 11:00-12:30 (Pier 7&8)
Step-by-step reasoning approaches like chain of thought (CoT) have proved to be very effective in inducing reasoning capabilities in large language models. However, the success of the CoT approach is fundamentally tied to the model size, and billion parameter-scale models are often needed to get CoT to work. In this paper, we propose a knowledge distillation approach that leverages the step-by-step CoT reasoning capabilities of larger models and distills these abilities into smaller models.

In this work, we propose an alternative reasoning scheme, Socratic CoT that learns a decomposition of the original problem into a sequence of subproblems and uses it to guide the intermediate reasoning steps. We use Socratic CoT to train a combination of two small distilled models: a problem decomposer and a subproblem solver. In practice, given a new problem, the two distilled models work in sync to decompose and solve complex problems. On multiple reasoning datasets (GSM8K, StrategyQA, and SVAMP), our proposed distillation strategies boosts the performance of smaller models over 70% compared to the baselines. Finally, we investigate when Socratic CoT is an effective alternative to CoT, demonstrating cases where a much smaller model (GPT-2 large) can outperform a 10X larger model (GPT-3 6B). Our code is available: https://github.com/kumar-shridhar/Distiiling-LM.

RISE: Leveraging Retrieval Techniques for Summarization Evaluation
David Uthus and Jianmo Ni 11:00-12:30 (Pier 7&8)
Evaluating automatically-generated text summaries is a challenging task. While there have been many interesting approaches, they still fall short of human evaluations. We present RISE, a new approach for evaluating summaries by leveraging techniques from information retrieval. RISE is first trained as a retrieval task using a dual-encoder retrieval setup, and can then be subsequently utilized for evaluating a generated summary given an input document, without gold reference summaries. RISE is especially well suited when working on new datasets where one may not have reference summaries available for evaluation. We conduct comprehensive experiments on the SummEval benchmark (Fabbri et al., 2021) and a long document summarization benchmark. The results show that RISE consistently achieves higher correlation with human evaluations compared to many past approaches to summarization evaluation. Furthermore, RISE also demonstrates data-efficiency and generalizability across languages.

Summary-Oriented Vision Modeling for Multimodal Abstractive Summarization
Yunlong Liang, Fandong Meng, Jinan Xu, Jiaan Wang, Yufeng Chen and Jie Zhou 11:00-12:30 (Pier 7&8)
The goal of multimodal abstractive summarization (MAS) is to produce a concise summary given the multimodal data (text and vision).

Existing studies on MAS mainly focus on how to effectively use the extracted visual features, having achieved impressive success on the high-resource English dataset. However, less attention has been paid to the quality of the visual features to the summary, which may limit the model performance, especially in the low- and zero-resource scenarios. In this paper, we propose to improve the summary quality through summary-oriented visual features. To this end, we devise two auxiliary tasks including vision to summary task and masked image modeling task. Together with the main summarization task, we optimize the MAS model via the training objectives of all these tasks. By these means, the MAS model can be enhanced by capturing the summary-oriented visual features, thereby yielding more accurate summaries. Experiments on 44 languages, covering mid-high-, low-, and zero-resource scenarios, verify the effectiveness and superiority of the proposed approach, which achieves state-of-the-art performance under all scenarios. Additionally, we will contribute a large-scale multilingual multimodal abstractive summarization (MM-Sum) dataset to the research community.

OpineSum: Entailment-based self-training for abstractive opinion summarization
Annie Louis and Joshua Maynez   11:00-12:30 (Pier 7&8)
A typical product or place often has hundreds of reviews, and summarization of these texts is an important and challenging problem. Recent progress on abstractive summarization in domains such as news has been driven by supervised systems trained on hundreds of thousands of news articles paired with human-written summaries. However for opinion texts, such large scale datasets are rarely available. Unsupervised methods, self-training, and few-shot learning approaches bridge that gap. In this work, we present a novel self-training approach, OpineSum for abstractive opinion summarization. The self-training summaries in this approach are built automatically using a novel application of textual entailment and capture the consensus of opinions across the various reviews for an item. This method can be used to obtain silver-standard summaries on a large scale and train both unsupervised and few-shot abstractive summarization systems. OpineSum outperforms strong peer systems in both settings.

MeetingBank: A Benchmark Dataset for Meeting Summarization
Yebowen Hu, Timothy Jeewun Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh and Fei Liu   11:00-12:30 (Pier 7&8)
As the number of recorded meetings increases, it becomes increasingly important to utilize summarization technology to create useful summaries of these recordings. However, there is a crucial lack of annotated meeting corpora for developing this technology, as it can be hard to collect meetings, especially when the topics discussed are confidential. Furthermore, meeting summaries written by experienced writers are scarce, making it hard for abstractive summarizers to produce sensible output without a reliable reference. This lack of annotated corpora has hindered the development of meeting summarization technology. In this paper, we present MeetingBank, a new benchmark dataset of city council meetings over the past decade. MeetingBank is unique among other meeting corpora due to its divide-and-conquer approach, which involves dividing professionally written meeting minutes into shorter passages and aligning them with specific segments of the meeting. This breaks down the process of summarizing a lengthy meeting into smaller, more manageable tasks. The dataset provides a new testbed of various meeting summarization systems and also allows the public to gain insight into how council decisions are made. We make the collection, including meeting video links, transcripts, reference summaries, agenda, and other metadata, publicly available to facilitate the development of better meeting summarization techniques.

An Investigation of Evaluation Methods in Automatic Medical Note Generation
Asma Ben Abacha, Wen-wai Yim, George Michalopoulos and Thomas Lin   11:00-12:30 (Pier 7&8)
Recent studies on automatic note generation have shown that doctors can save significant amounts of time when using automatic clinical note generation (Knoll et al., 2022). Summarization models have been used for this task to generate clinical notes as summaries of doctor-patient conversations (Krishna et al., 2021; Cai et al., 2022). However, assessing which model would best serve clinicians in their daily practice is still a challenging task due to the large set of possible correct summaries, and the potential limitations of automatic evaluation metrics. In this paper we study evaluation methods and metrics for the automatic generation of clinical notes from medical conversation. In particular, we propose new task-specific metrics and we compare them to SOTA evaluation metrics in text summarization and generation, including: (i) knowledge-graph embedding-based metrics, (ii) customized model-based metrics with domain-specific weights, (iii) domain-adapted/fine-tuned metrics, and (iv) ensemble metrics. To study the correlation between the automatic metrics and manual judgments, we evaluate automatic notes/summaries by comparing the system and reference facts and computing the factual correctness, and the hallucination and omission rates for critical medical facts. This study relied on seven datasets manually annotated by domain experts. Our experiments show that automatic evaluation metrics can have substantially different behaviors on different types of clinical notes datasets. However, the results highlight one stable subset of metrics as the most correlated with human judgments with a relevant aggregation of different evaluation criteria.

Towards Unifying Multi-Lingual and Cross-Lingual Summarization
Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu and Jie Zhou   11:00-12:30 (Pier 7&8)
To adapt text summarization to the multilingual world, previous work proposes multi-lingual summarization (MLS) and cross-lingual summarization (CLS). However, these two tasks have been studied separately due to the different definitions, which limits the compatible and systematic research on both of them. In this paper, we aim to unify MLS and CLS into a more general setting, i.e., many-to-many summarization (M2MS), where a single model could process documents in any language and generate their summaries also in any language. As the first step towards M2MS, we conduct preliminary studies to show that M2MS can better transfer task knowledge across different languages than MLS and CLS. Furthermore, we propose Pisces, a pre-trained M2MS model that learns language modeling, cross-lingual ability and summarization ability via three-stage pre-training. Experimental results indicate that our Pisces significantly outperforms the state-of-the-art baselines, especially in the zero-shot directions, where there is no training data from the source-language documents to the target-language summaries.

Generating User-Engaging News Headlines
Pengshan Cai, Kaiqiang Song, Sangwoo Cho, Hongwei Wang, Xiaoyang Wang, hong yu, Fei Liu and Dong Yu   11:00-12:30 (Pier 7&8)
The potential choices for news article headlines are enormous, and finding the right balance between conveying the essential message and capturing the reader's attention is key to effective headlining. However, presenting the same news headline to all readers is a suboptimal strategy, because it does not take into account the different preferences and interests of diverse readers, who may be confused about why a particular article has been recommended to them and do not see a clear connection between their interests and the recommended article. In this paper, we present a novel framework that addresses these challenges by incorporating user profiling to generate personalized headlines, and a combination of automated and human evaluation methods to determine user preference for personalized headlines. Our framework utilizes a learnable relevance function to assign personalized signature phrases to users based on their reading histories, which are then used to personalize headline generation. Through extensive evaluation, we demonstrate the effectiveness of our proposed framework in generating personalized headlines that meet the needs of a diverse audience. Our framework has the potential to improve the efficacy of news recommendations and facilitate creation of personalized content.

Fixing MoE Over-Fitting on Low-Resource Languages in Multilingual Machine Translation
Maha Elbayad, Anna Y Sun and Shruti Bhosale   11:00-12:30 (Pier 7&8)
Sparsely gated Mixture of Experts (MoE) models have been shown to be a compute-efficient method to scale model capacity for multilingual

machine translation. However, for low-resource tasks, MoE models severely over-fit. We show effective regularization strategies, namely dropout techniques for MoE layers in EOM and FOM, Conditional MoE Routing and Curriculum Learning methods that prevent over-fitting and improve the performance of MoE models on low-resource tasks without adversely affecting high-resource tasks. On a massively multilingual machine translation benchmark, our strategies result in about +1 chrF++ improvement in very low resource language pairs. We perform an extensive analysis of the learned MoE routing to better understand the impact of our regularization methods and how we can improve them.

In-context Examples Selection for Machine Translation

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer and Marjan Ghazvininejad   11:00-12:30 (Pier 7&8)

Large-scale generative models show an impressive ability to perform a wide range of Natural Language Processing (NLP) tasks using in-context learning, where a few examples are used to describe a task to the model. For Machine Translation (MT), these examples are typically randomly sampled from the development dataset with a similar distribution as the evaluation set. However, it is unclear how the choice of these in context examples and their ordering impacts the output translation quality. In this work, we aim to understand the properties of good in-context examples for MT in both in-domain and out-of-domain settings. We show that the translation quality and the domain of the in-context examples matter and that 1-shot noisy unrelated examples can have a catastrophic impact on output quality. While concatenating multiple random examples reduces the effect of noise, a single good prompt optimized to maximize translation quality on the development dataset can elicit learned information from the pre-trained language model. Adding similar examples based on an n-gram overlap with the test source significantly and consistently improves the translation quality of the outputs, outperforming a strong kNN-MT baseline in 2 out of 4 out-of-domain datasets.

Disambiguated Lexically Constrained Neural Machine Translation

Jinpeng Zhang, Nini Xiao, Ke Wang, Chuanqi Dong, Xiangyu Duan, Yuqi Zhang and Min Zhang   11:00-12:30 (Pier 7&8)

Lexically constrained neural machine translation (LCNMT), which controls the translation generation with pre-specified constraints, is important in many practical applications. Current approaches to LCNMT typically assume that the pre-specified lexicon constraints are contextually appropriate. This assumption limits their application to real-world scenarios where a source lexicon may have multiple target constraints, and disambiguation is needed to select the most suitable one. In this paper, we propose disambiguated LCNMT (D-LCNMT) to solve the problem. D-LCNMT is a robust and effective two-stage framework that disambiguates the constraints based on contexts at first, then integrates the disambiguated constraints into LCNMT. Experimental results show that our approach outperforms strong baselines including existing data argumentation based approaches on benchmark datasets, and comprehensive experiments in scenarios where a source lexicon corresponds to multiple target constraints demonstrate the constraint disambiguation superiority of our approach.

Learning Optimal Policy for Simultaneous Machine Translation via Binary Search

Shoutao Guo, Shaolei Zhang and Yang Feng   11:00-12:30 (Pier 7&8)

Simultaneous machine translation (SiMT) starts to output translation while reading the source sentence and needs a precise policy to decide when to output the generated translation. Therefore, the policy determines the number of source tokens read during the translation of each target token. However, it is difficult to learn a precise translation policy to achieve good latency-quality trade-offs, because there is no golden policy corresponding to parallel sentences as explicit supervision. In this paper, we present a new method for constructing the optimal policy online via binary search. By employing explicit supervision, our approach enables the SiMT model to learn the optimal policy, which can guide the model in completing the translation during inference. Experiments on four translation tasks show that our method can exceed strong baselines across all latency scenarios.

Easy Guided Decoding in Providing Suggestions for Interactive Machine Translation

Ke Wang, Xin Ge, Jiayi Wang, Yuqi Zhang and Yu Zhao   11:00-12:30 (Pier 7&8)

Machine translation technology has made great progress in recent years, but it cannot guarantee error-free results. Human translators perform post-editing on machine translations to correct errors in the scene of computer aided translation. In favor of expediting the post-editing process, many works have investigated machine translation in interactive modes, in which machines can automatically refine the rest of translations constrained by human's edits. Translation Suggestion (TS), as an interactive mode to assist human translators, requires machines to generate alternatives for specific incorrect words or phrases selected by human translators. In this paper, we utilize the parameterized objective function of neural machine translation (NMT) and propose a novel constrained decoding algorithm, namely Prefix-Suffix Guided Decoding (PSGD), to deal with the TS problem without additional training. Compared to state-of-the-art lexical-constrained decoding method, PSGD improves translation quality by an average of 10.6 BLEU and reduces time overhead by an average of 63.4% on benchmark datasets. Furthermore, on both the WeTS and the WMT 2022 Translation Suggestion datasets, it is superior over other supervised learning systems trained with TS annotated data.

CMOT: Cross-modal Mixup via Optimal Transport for Speech Translation

Yan Zhou, Qingkai Fang and Yang Feng   11:00-12:30 (Pier 7&8)

End-to-end speech translation (ST) is the task of translating speech signals in the source language into text in the target language. As a cross-modal task, end-to-end ST is difficult to train with limited data. Existing methods often try to transfer knowledge from machine translation (MT), but their performances are restricted by the modality gap between speech and text. In this paper, we propose Cross-modal Mixup via Optimal Transport (CMOT) to overcome the modality gap. We find the alignment between speech and text sequences via optimal transport and then mix up the sequences from different modalities at a token level using the alignment. Experiments on the MuST-C ST benchmark demonstrate that CMOT achieves an average BLEU of 30.0 in 8 translation directions, outperforming previous methods. Further analysis shows CMOT can adaptively find the alignment between modalities, which helps alleviate the modality gap between speech and text.

Extract and Attend: Improving Entity Translation in Neural Machine Translation

Zixin Zeng, Rui Wang, Yichong Leng, Junliang Guo, Shufang Xie, Xu Tan, Tao Qin and Tie-Yan Liu   11:00-12:30 (Pier 7&8)

While Neural Machine Translation (NMT) has achieved great progress in recent years, it still suffers from inaccurate translation of entities (e.g., person/organization name, location), due to the lack of entity training instances. When we humans encounter an unknown entity during translation, we usually first look up in a dictionary and then organize the entity translation together with the translations of other parts to form a smooth target sentence. Inspired by this translation process, we propose an Extract-and-Attend approach to enhance entity translation in NMT, where the translation candidates of source entities are first extracted from a dictionary and then attended to by the NMT model to generate the target sentence. Specifically, the translation candidates are extracted by first detecting the entities in a source sentence and then translating the entities through looking up in a dictionary. Then, the extracted candidates are added as a prefix of the decoder input to be attended to by the decoder when generating the target sentence through self-attention. Experiments conducted on En-Zh and En-Ru demonstrate that the proposed method is effective on improving both the translation accuracy of entities and the overall translation quality, with up to 35% reduction on entity error rate and 0.85 gain on BLEU and 13.8 gain on COMET.

Lego-MT: Learning Detachable Models for Massively Multilingual Machine Translation

Fei Yuan, Yinquan Lu, Wenhao Zhu, Lingpeng Kong, Lei Li, Yu Qiao and Jingjing Xu   11:00-12:30 (Pier 7&8)

Multilingual neural machine translation (MNMT) aims to build a unified model for many language directions. Existing monolithic models for

MNMT encounter two challenges: parameter interference among languages and inefficient inference for large models. In this paper, we revisit the classic multi-way structures and develop a detachable model by assigning each language (or group of languages) to an individual branch that supports plug-and-play training and inference. To address the needs of learning representations for all languages in a unified space, we propose a novel efficient training recipe, upon which we build an effective detachable model, Lego-MT. For a fair comparison, we collect data from OPUS and build a translation benchmark covering 433 languages and 1.3B parallel data. Experiments show that Lego-MT with 1.2B parameters brings an average gain of 3.2 spBLEU. It even outperforms M2M-100 with 12B parameters. The proposed training recipe brings a 28.2× speedup over the conventional multi-way training method.code and data repo: https://github.com/CONE-MT/Lego-MT.git.

Duplex Diffusion Models Improve Speech-to-Speech Translation
Xianchao Wu   11:00-12:30 (Pier 7&8)
Speech-to-speech translation is a typical sequence-to-sequence learning task that naturally has two directions. How to effectively leverage bidirectional supervision signals to produce high-fidelity audio for both directions? Existing approaches either train two separate models or a multitask-learned model with low efficiency and inferior performance. In this paper, we propose a duplex diffusion model that applies diffusion probabilistic models to both sides of a reversible duplex Conformer, so that either end can simultaneously input and output a distinct language's speech. Our model enables reversible speech translation by simply flipping the input and output ends. Experiments show that our model achieves the first success of reversible speech translation with significant improvements of ASR-BLEU scores compared with a list of state-of-the-art baselines.

CKDST: Comprehensively and Effectively Distill Knowledge from Machine Translation to End-to-End Speech Translation
Yikun Lei, Zhengshan Xue, Xiaohu Zhao, Haoran Sun, Shaolin Zhu, xiaodong lin and Deyi Xiong   11:00-12:30 (Pier 7&8)
Distilling knowledge from a high-resource task, e.g., machine translation, is an effective way to alleviate the data scarcity problem of end-to-end speech translation. However, previous works simply use the classical knowledge distillation that does not allow for adequate transfer of knowledge from machine translation. In this paper, we propose a comprehensive knowledge distillation framework for speech translation, CKDST, which is capable of comprehensively and effectively distilling knowledge from machine translation to speech translation from two perspectives: cross-modal contrastive representation distillation and simultaneous decoupled knowledge distillation. In the former, we leverage a contrastive learning objective to optimize the mutual information between speech and text representations for representation distillation in the encoder. In the later, we decouple the non-target class knowledge from target class knowledge for logits distillation in the decoder. Experiments on the MuST-C benchmark dataset demonstrate that our CKDST substantially improves the baseline by 1.2 BLEU on average in all translation directions, and outperforms previous state-of-the-art end-to-end and cascaded speech translation models.

Robustness of Multi-Source MT to Transcription Errors
Dominik Macháček, Peter Polak, Ondřej Bojar and Raj Dabre   11:00-12:30 (Pier 7&8)
Automatic speech translation is sensitive to speech recognition errors, but in a multilingual scenario, the same content may be available in various languages via simultaneous interpreting, dubbing or subtitling. In this paper, we hypothesize that leveraging multiple sources will improve translation quality if the sources complement one another in terms of correct information they contain. To this end, we first show that on a 10-hour ESIC corpus, the ASR errors in the original English speech and its simultaneous interpreting into German and Czech are mutually independent. We then use two sources, English and German, in a multi-source setting for translation into Czech to establish its robustness to ASR errors. Furthermore, we observe this robustness when translating both noisy sources together in a simultaneous translation setting. Our results show that multi-source neural machine translation has the potential to be useful in a real-time simultaneous translation setting, thereby motivating further investigation in this area.

CTC-based Non-autoregressive Speech Translation
Chen Xu, Xiaoqian Liu, Xiaowen Liu, Qingxuan Sun, Yuhao Zhang, Murun Yang, Qianqian Dong, Tom Ko, Mingxuan Wang, Tong Xiao, Anxiang Ma and Jingbo Zhu   11:00-12:30 (Pier 7&8)
Combining end-to-end speech translation (ST) and non-autoregressive (NAR) generation is promising in language and speech processing for their advantages of less error propagation and low latency. In this paper, we investigate the potential of connectionist temporal classification (CTC) for non-autoregressive speech translation (NAST). In particular, we develop a model consisting of two encoders that are guided by CTC to predict the source and target texts, respectively. Introducing CTC into NAST on both language sides has obvious challenges: 1) the conditional independent generation somewhat breaks the interdependency among tokens, and 2) the monotonic alignment assumption in standard CTC does not hold in translation tasks. In response, we develop a prediction-aware encoding approach and a cross-layer attention approach to address these issues. We also use curriculum learning to improve convergence of training. Experiments on the MuST-C ST benchmarks show that our NAST model achieves an average BLEU score of 29.5 with a speed-up of 5.67×, which is comparable to the autoregressive counterpart and even outperforms the previous best result of 0.9 BLEU points.

Implicit Memory Transformer for Computationally Efficient Simultaneous Speech Translation
Matthew Raffel and Lizhong Chen   11:00-12:30 (Pier 7&8)
Simultaneous speech translation is an essential communication task difficult for humans whereby a translation is generated concurrently with oncoming speech inputs. For such a streaming task, transformers using block processing to break an input sequence into segments have achieved state-of-the-art performance at a reduced cost. Current methods to allow information to propagate across segments, including left context and memory banks, have faltered as they are both insufficient representations and unnecessarily expensive to compute. In this paper, we propose an Implicit Memory Transformer that implicitly retains memory through a new left context method, removing the need to explicitly represent memory with memory banks. We generate the left context from the attention output of the previous segment and include it in the keys and values of the current segment's attention calculation. Experiments on the MuST-C dataset show that the Implicit Memory Transformer provides a substantial speedup on the encoder forward pass with nearly identical translation quality when compared with the state-of-the-art approach that employs both left context and memory banks.

Enhancing Event Causality Identification with Counterfactual Reasoning
Feiteng Mu and Wenjie Li   11:00-12:30 (Pier 7&8)
Existing methods for event causality identification (ECI) focus on mining potential causal signals, i.e., causal context keywords and event pairs. However, causal signals are ambiguous, which may lead to the context-keywords bias and the event-pairs bias. To solve this issue, we propose the *counterfactual reasoning* that explicitly estimates the influence of context keywords and event pairs in training, so that we are able to eliminate the biases in inference. Experiments are conducted on two datasets, the result demonstrates the effectiveness of our method.

Uncertainty-Aware Bootstrap Learning for Joint Extraction on Distantly-Supervised Data
Yufei Li, Xiao Yu, Yanchi Liu, Haifeng Chen and Cong Liu   11:00-12:30 (Pier 7&8)
Jointly extracting entity pairs and their relations is challenging when working on distantly-supervised data with ambiguous or noisy labels. To mitigate such impact, we propose uncertainty-aware bootstrap learning, which is motivated by the intuition that the higher uncertainty of an instance, the more likely the model confidence is inconsistent with the ground truths. Specifically, we first explore instance-level data uncertainty to create an initial high-confident examples. Such subset serves as filtering noisy instances and facilitating the model to converge

fast at the early stage. During bootstrap learning, we propose self-ensembling as a regularizer to alleviate inter-model uncertainty produced by noisy labels. We further define probability variance of joint tagging probabilities to estimate inner-model parametric uncertainty, which is used to select and build up new reliable training instances for the next iteration. Experimental results on two large datasets reveal that our approach outperforms existing strong baselines and related methods.

Bootstrapping Neural Relation and Explanation Classifiers

Zheng Tang and Mihai Surdeanu 11:00-12:30 (Pier 7&8)
We introduce a method that self trains (or bootstraps) neural relation and explanation classifiers. Our work expands the supervised approach of CITATION, which jointly trains a relation classifier with an explanation classifier that identifies context words important for the relation at hand, to semi-supervised scenarios. In particular, our approach iteratively converts the explainable models' outputs to rules and applies them to unlabeled text to produce new annotations. Our evaluation on the TACRED dataset shows that our method outperforms the rule-based model we started from by 15 F1 points, outperforms traditional self-training that relies just on the relation classifier by 5 F1 points, and performs comparatively with the prompt-based approach of CITATION (without requiring an additional natural language inference component).

UTC-IE: A Unified Token-pair Classification Architecture for Information Extraction

Hang Yan, Yu Sun, Xiaonan Li, Yunhua Zhou, Xuanjing Huang and Xipeng Qiu 11:00-12:30 (Pier 7&8)
Information Extraction (IE) spans several tasks with different output structures, such as named entity recognition, relation extraction and event extraction. Previously, those tasks were solved with different models because of diverse task output structures. Through re-examining IE tasks, we find that all of them can be interpreted as extracting spans and span relations. They can further be decomposed into token-pair classification tasks by using the start and end token of a span to pinpoint the span, and using the start-to-start and end-to-end token pairs of two spans to determine the relation. Based on the reformulation, we propose a Unified Token-pair Classification architecture for Information Extraction (UTC-IE), where we introduce Plusformer on top of the token-pair feature matrix. Specifically, it models axis-aware interaction with plus-shaped self-attention and local interaction with Convolutional Neural Network over token pairs. Experiments show that our approach outperforms task-specific and unified models on all tasks in 10 datasets, and achieves better or comparable results on 2 joint IE datasets. Moreover, UTC-IE speeds up over state-of-the-art models on IE tasks significantly in most datasets, which verifies the effectiveness of our architecture.

Guide the Many-to-One Assignment: Open Information Extraction via IoU-aware Optimal Transport

Kaiwen Wei, Yiran Yang, li jin, Xian Sun, Zequn Zhang, Jingyuan Zhang, Xiao yu Li, Linhao Zhang, Jintao Liu and Guo Zhi 11:00-12:30 (Pier 7&8)
Open Information Extraction (OIE) seeks to extract structured information from raw text without the limitations of close ontology. Recently, the detection-based OIE methods have received great attention from the community due to their parallelism. However, as the essential step of those models, how to assign ground truth labels to the parallelly generated tuple proposals remains under-exploited. The commonly utilized Hungarian algorithm for this procedure is restricted to handling one-to-one assignment among the desired tuples and tuple proposals, which ignores the correlation between proposals and affects the recall of the models. To solve this problem, we propose a dynamic many-to-one label assignment strategy named IOT. Concretely, the label assignment process in OIE is formulated as an Optimal Transport (OT) problem. We leverage the intersection-over-union (IoU) as the assignment quality measurement, and convert the problem of finding the best assignment solution to the one solving the optimal transport plan by maximizing the IoU values. To further utilize the knowledge from the assignment, we design an Assignment-guided Multi-granularity loss (AM) by simultaneously considering word-level and tuple-level information. Experiment results show the proposed method outperforms the state-of-the-art models on three benchmarks.

From Ultra-Fine to Fine: Fine-tuning Ultra-Fine Entity Typing Models to Fine-grained

Hongliang Dai and Ziqian Zeng 11:00-12:30 (Pier 7&8)
For the task of fine-grained entity typing (FET), due to the use of a large number of entity types, it is usually considered too costly to manually annotating a training dataset that contains an ample number of examples for each type. A common way to address this problem is to use distantly annotated training data that contains incorrect labels. However, the performance of models trained solely with such data can be limited by the errors in the automatic annotation. Recently, there are a few approaches that no longer follow this conventional way. But without using sufficient direct entity typing supervision may also cause them to yield inferior performance. In this paper, we propose a new approach that can avoid the need of creating distantly labeled data whenever there is a new type schema. We first train an entity typing model that have an extremely board type coverage by using the ultra-fine entity typing data. Then, when there is a need to produce a model for a newly designed fine-grained entity type schema. We can simply fine-tune the previously trained model with a small number of examples annotated under this schema. Experimental results show that our approach achieves outstanding performance for FET under the few-shot setting. It can also outperform state-of-the-art weak supervision based methods after fine-tuning the model with only a small size manually annotated training set.

Data Augmentation for Low-Resource Keyphrase Generation

Krishna K Garg, Jishnu Ray Chowdhury and Cornelia Caragea 11:00-12:30 (Pier 7&8)
Keyphrase generation is the task of summarizing the contents of any given article into a few salient phrases (or keyphrases). Existing works for the task mostly rely on large-scale annotated datasets, which are not easy to acquire. Very few works address the problem of keyphrase generation in low-resource settings, but they still rely on a lot of additional unlabeled data for pretraining and on automatic methods for pseudo-annotations. In this paper, we present data augmentation strategies specifically to address keyphrase generation in purely resource-constrained domains. We design techniques that use the full text of the articles to improve both present and absent keyphrase generation. We test our approach comprehensively on three datasets and show that the data augmentation strategies consistently improve the state-of-the-art performance. We release our source code at https://github.com/kgarg8/kpgen-lowres-data-aug.

Zero- and Few-Shot Event Detection via Prompt-Based Meta Learning

Zhenrui Yue, Huimin Zeng, Mengfei Lan, Heng Ji and Dong Wang 11:00-12:30 (Pier 7&8)
With emerging online topics as a source for numerous new events, detecting unseen / rare event types presents an elusive challenge for existing event detection methods, where only limited data access is provided for training. To address the data scarcity problem in event detection, we propose MetaEvent, a meta learning-based framework for zero- and few-shot event detection. Specifically, we sample training tasks from existing event types and perform meta training to search for optimal parameters that quickly adapt to unseen tasks. In our framework, we propose to use the cloze-based prompt and a trigger-aware soft verbalizer to efficiently project output to unseen event types. Moreover, we design a contrastive meta objective based on maximum mean discrepancy (MMD) to learn class-separating features. As such, the proposed MetaEvent can perform zero-shot event detection by mapping features to event types without any prior knowledge. In our experiments, we demonstrate the effectiveness of MetaEvent in both zero-shot and few-shot scenarios, where the proposed method achieves state-of-the-art performance in extensive experiments on benchmark datasets FewEvent and MAVEN.

Early Discovery of Disappearing Entities in Microblogs

Satoshi Akasaki, Naoki Yoshinaga and Masashi Toyoda 11:00-12:30 (Pier 7&8)
We make decisions by reacting to changes in the real world, particularly the emergence and disappearance of impermanent entities such as

restaurants, services, and events. Because we want to avoid missing out on opportunities or making fruitless actions after those entities have disappeared, it is important to know when entities disappear as early as possible. We thus tackle the task of detecting disappearing entities from microblogs where various information is shared timely. The major challenge is detecting uncertain contexts of disappearing entities from noisy microblog posts. To collect such disappearing contexts, we design time-sensitive distant supervision, which utilizes entities from the knowledge base and time-series posts. Using this method, we actually build large-scale Twitter datasets of disappearing entities. To ensure robust detection in noisy environments, we refine pretrained word embeddings for the detection model on microblog streams in a timely manner. Experimental results on the Twitter datasets confirmed the effectiveness of the collected labeled data and refined word embeddings; the proposed method outperformed a baseline in terms of accuracy, and more than 70% of the detected disappearing entities in Wikipedia are discovered earlier than the update on Wikipedia, with the average lead-time is over one month.

An AMR-based Link Prediction Approach for Document-level Event Argument Extraction
Yuqing Yang, Qipeng Guo, Xiangkun Hu, Yue Zhang, Xipeng Qiu and Zheng Zhang 11:00-12:30 (Pier 7&8)
Recent works have introduced Abstract Meaning Representation (AMR) for Document-level Event Argument Extraction (Doc-level EAE), since AMR provides a useful interpretation of complex semantic structures and helps to capture long-distance dependency. However, in these works AMR is used only implicitly, for instance, as additional features or training signals. Motivated by the fact that all event structures can be inferred from AMR, this work reformulates EAE as a link prediction problem on AMR graphs. Since AMR is a generic structure and does not perfectly suit EAE, we propose a novel graph structure, Tailored AMR Graph (TAG), which compresses less informative subgraphs and edge types, integrates span information, and highlights surrounding events in the same document. With TAG, we further propose a novel method using graph neural networks as a link prediction model to find event arguments. Our extensive experiments on WikiEvents and RAMS show that this simpler approach outperforms the state-of-the-art models by 3.63pt and 2.33pt F1, respectively, and do so with reduced 56% inference time.

Document-Level Event Argument Extraction With a Chain Reasoning Paradigm
Jian Liu, Chen Liang, Jinan Xu, Haoyan Liu and Zhe Zhao 11:00-12:30 (Pier 7&8)
Document-level event argument extraction aims to identify event arguments beyond sentence level, where a significant challenge is to model long-range dependencies. Focusing on this challenge, we present a new chain reasoning paradigm for the task, which can generate decomposable first-order logic rules for reasoning. This paradigm naturally captures long-range interdependence due to the chains' compositional nature, which also improves interpretability by explicitly modeling the reasoning process. We introduce T-norm fuzzy logic for optimization, which permits end-to-end learning and shows promise for integrating the expressiveness of logical reasoning with the generalization of neural networks. In experiments, we show that our approach outperforms previous methods by a significant margin on two standard benchmarks (over 6 points in F1). Moreover, it is data-efficient in low-resource scenarios and robust enough to defend against adversarial attacks.

Joint Document-Level Event Extraction via Token-Token Bidirectional Event Completed Graph
Qizhi Wan, Changxuan Wan, Keli Xiao, Dexi Liu, Chenliang Li, Bolong Zheng, Xiping Liu and Rong Hu 11:00-12:30 (Pier 7&8)
We solve the challenging document-level event extraction problem by proposing a joint exaction methodology that can avoid inefficiency and error propagation issues in classic pipeline methods. Essentially, we address the three crucial limitations in existing studies. First, the autoregressive strategy of path expansion heavily relies on the orders of argument role. Second, the number of events in documents must be specified in advance. Last, unexpected errors usually exist when decoding events based on the entity-entity adjacency matrix. To address these issues, this paper designs a Token-Token Bidirectional Event Completed Graph (TT-BECG) in which the relation eType-Role1-Role2 serves as the edge type, precisely revealing which tokens play argument roles in an event of a specific event type. Exploiting the token-token adjacency matrix of the TT-BECG, we develop an edge-enhanced joint document-level event extraction model. Guided by the target token-token adjacency matrix, the predicted token-token adjacency matrix can be obtained during the model training. Then, extracted events and event records in a document are decoded based on the predicted matrix, including the graph structure and edge type decoding. Extensive experiments are conducted on two public datasets, and the results confirm the effectiveness of our method and its superiority over the state-of-the-art baselines.

CoAug: Combining Augmentation of Labels and Labelling Rules
Rakesh R. Menon, Bingqing Wang, Jun Araki, Zhengyu Zhou, Zhe Feng and Liu Ren 11:00-12:30 (Pier 7&8)
Collecting labeled data for Named Entity Recognition (NER) tasks is challenging due to the high cost of manual annotations. Instead, researchers have proposed few-shot self-training and rule-augmentation techniques to minimize the reliance on large datasets. However, inductive biases and restricted logical language lexicon, respectively, can limit the ability of these models to perform well. In this work, we propose CoAug, a co-augmentation framework that allows us to improve few-shot models and rule-augmentation models by bootstrapping predictions from each model. By leveraging rules and neural model predictions to train our models, we complement the benefits of each and achieve the best of both worlds. In our experiments, we show that our best CoAug model can outperform strong weak-supervision-based NER models at least by 6.5 F1 points.

QueryForm: A Simple Zero-shot Form Entity Query Framework
Zifeng Wang, Zizhao Zhang, Jacob Devlin, Chen-Yu Lee, Guolong Su, Hao Zhang, Jennifer Dy, Vincent Perot and Tomas Pfister 11:00-12:30 (Pier 7&8)
Zero-shot transfer learning for document understanding is a crucial yet under-investigated scenario to help reduce the high cost involved in annotating document entities. We present a novel query-based framework, QueryForm, that extracts entity values from form-like documents in a zero-shot fashion. QueryForm contains a dual prompting mechanism that composes both the document schema and a specific entity type into a query, which is used to prompt a Transformer model to perform a single entity extraction task. Furthermore, we propose to leverage large-scale query-entity pairs generated from form-like webpages with weak HTML annotations to pre-train QueryForm. By unifying pre-training and fine-tuning into the same query-based framework, QueryForm enables models to learn from structured documents containing various entities and layouts, leading to better generalization to target document types without the need for target-specific training data. Query-Form sets new state-of-the-art average F1 score on both the XFUND (+4.6% 10.1%) and the Payment (+3.2% 9.5%) zero-shot benchmark, with a smaller model size and no additional image input.

ECG-QALM: Entity-Controlled Synthetic Text Generation using Contextual Q&A for NER
Karan Aggarwal, Henry Jin and Aitzaz Ahmad 11:00-12:30 (Pier 7&8)
Named Entity Recognition (NER) state-of-the-art methods requires high-quality labeled datasets. Issues such as scarcity of labeled data, under-representation of entities, and privacy concerns with using sensitive data for training, can be significant barriers. Generating synthetic data to train models is a promising solution to mitigate these problems. We propose ECG-QALM, a contextual question and answering approach using pre-trained language models to synthetically generate entity-controlled text. Generated text is then used to augment small labeled datasets for downstream NER tasks. We evaluate our method on two publicly available datasets. We find ECG-QALM is capable of producing full text samples with desired entities appearing in a controllable way, while retaining sentence coherence closest to the real world data. Evaluations on NER tasks show significant improvements (75% - 140%) in low-labeled data regimes.

PuMer: Pruning and Merging Tokens for Efficient Vision Language Models

Qingqing Cao, Bhargavi Paranjape and Hannaneh Hajishirzi 11:00-12:30 (Pier 7&8)
Large-scale vision language (VL) models use Transformers to perform cross-modal interactions between the input text and image. These cross-modal interactions are computationally expensive and memory-intensive due to the quadratic complexity of processing the input image and text. We present PuMer: a token reduction framework that uses text-informed Pruning and modality-aware Merging strategies to progressively reduce the tokens of input image and text, improving model inference speed and reducing memory footprint. PuMer learns to keep salient image tokens related to the input text and merges similar textual and visual tokens by adding lightweight token reducer modules at several cross-modal layers in the VL model. Training PuMer is mostly the same as finetuning the original VL model but faster. Our evaluation for two vision language models on four downstream VL tasks shows PuMer increases inference throughput by up to 2x and reduces memory footprint by over 50% while incurring less than a 1% accuracy drop.

Generating Hashtags for Short-form Videos with Guided Signals
*Tiezheng Yu, Hanchao Yu, Davis Liang, Yuning Mao, Shaoliang Nie, Po-Yao Huang, Madian Khabsa, Pascale Fung and Yi-Chia Wang*11:00-12:30 (Pier 7&8)
Short-form video hashtag recommendation (SVHR) aims to recommend hashtags to content creators from videos and corresponding descriptions. Most prior studies regard SVHR as a classification or ranking problem and select hashtags from a set of limited candidates. However, in reality, users can create new hashtags, and trending hashtags change rapidly over time on social media. Both of these properties cannot be easily modeled with classification approaches. To bridge this gap, we formulate SVHR as a generation task that better represents how hashtags are created naturally. Additionally, we propose the Guided Generative Model (GGM) where we augment the input features by retrieving relevant hashtags from a large-scale hashtag pool as extra guidance signals. Experimental results on two short-form video datasets show that our generative models outperform strong classification baselines, and the guidance signals further boost the performance by 8.11 and 2.17 absolute ROUGE-1 scores on average, respectively. We also perform extensive analyses including human evaluation, demonstrating that our generative model can create meaningful and relevant novel hashtags while achieving state-of-the-art performance on known hashtags

Generating Structured Pseudo Labels for Noise-resistant Zero-shot Video Sentence Localization
Minghang Zheng, Shaogang Gong, Hailin Jin, Yuxin Peng and Yang Liu 11:00-12:30 (Pier 7&8)
Video sentence localization aims to locate moments in an unstructured video according to a given natural language query. A main challenge is the expensive annotation costs and the annotation bias. In this work, we study video sentence localization in a zero-shot setting, which learns with only video data without any annotation. Existing zero-shot pipelines usually generate event proposals and then generate a pseudo query for each event proposal. However, their event proposals are obtained via visual feature clustering, which is query-independent and inaccurate; and the pseudo-queries are short or less interpretable. Moreover, existing approaches ignores the risk of pseudo-label noise when leveraging them in training. To address the above problems, we propose a Structure-based Pseudo Label generation (SPL), which first generate free-form interpretable pseudo queries before constructing query-dependent event proposals by modeling the event temporal structure. To mitigate the effect of pseudo-label noise, we propose a noise-resistant iterative method that repeatedly re-weight the training sample based on noise estimation to train a grounding model and correct pseudo labels. Experiments on the ActivityNet Captions and Charades-STA datasets demonstrate the advantages of our approach. Code can be found at https://github.com/minghangz/SPL.

Modularized Zero-shot VQA with Pre-trained Models
RUI CAO and Jing Jiang 11:00-12:30 (Pier 7&8)
Large-scale pre-trained models (PTMs) show great zero-shot capabilities. In this paper, we study how to leverage them for zero-shot visual question answering (VQA). Our approach is motivated by a few observations. First, VQA questions often require multiple steps of reasoning, which is still a capability that most PTMs lack. Second, different steps in VQA reasoning chains require different skills such as object detection and relational reasoning, but a single PTM may not possess all these skills. Third, recent work on zero-shot VQA does not explicitly consider multi-step reasoning chains, which makes them less interpretable compared with a decomposition-based approach. We propose a modularized zero-shot network that explicitly decomposes questions into sub reasoning steps and is highly interpretable. We convert sub reasoning tasks to acceptable objectives of PTMs and assign tasks to proper PTMs without any adaptation. Our experiments on two VQA benchmarks under the zero-shot setting demonstrate the effectiveness of our method and better interpretability compared with several baselines.

A Multi-Modal Context Reasoning Approach for Conditional Inference on Joint Textual and Visual Clues
Yunxin Li, Baotian Hu, Chen Xinyu, Yuxin Ding, Lin Ma and Min Zhang 11:00-12:30 (Pier 7&8)
Conditional inference on joint textual and visual clues is a multi-modal reasoning task that textual clues provide prior permutation or external knowledge, which are complementary with visual content and pivotal to deducing the correct option. Previous methods utilizing pretrained vision-language models (VLMs) have achieved impressive performances, yet they show a lack of multimodal context reasoning capability, especially for text-modal information. To address this issue, we propose a Multi-modal Context Reasoning approach, named ModCR. Compared to VLMs performing reasoning via cross modal semantic alignment, it regards the given textual abstract semantic and objective image information as the pre-context information and embeds them into the language model to perform context reasoning. Different from recent vision-aided language models used in natural language processing, ModCR incorporates the multi-view semantic alignment information between language and vision by introducing the learnable alignment prefix between image and text in the pretrained language model. This makes the language model well-suitable for such multi-modal reasoning scenario on joint textual and visual clues. We conduct extensive experiments on two corresponding data sets and experimental results show significantly improved performance (exact gain by 4.8% on PMR test set) compared to previous strong baselines.

UniFine: A Unified and Fine-grained Approach for Zero-shot Vision-Language Understanding
Rui Sun, Zhecan Wang, Haoxuan You, Noel Codella, Kai-Wei Chang and Shih-Fu Chang 11:00-12:30 (Pier 7&8)
Vision-language tasks, such as VQA, SNLI-VE, and VCR are challenging because they require the model's reasoning ability to understand the semantics of the visual world and natural language. Supervised methods working for vision-language tasks have been well-studied. However, solving these tasks in a zero-shot setting is less explored. Since Contrastive Language-Image Pre-training (CLIP) has shown remarkable zero-shot performance on image-text matching, previous works utilized its strong zero-shot ability by converting vision-language tasks into an image-text matching problem, and they mainly consider global-level matching (e.g., the whole image or sentence). However, we find visual and textual fine-grained information, e.g., keywords in the sentence and objects in the image, can be fairly informative for semantics understanding. Inspired by this, we propose a unified framework to take advantage of the fine-grained information for zero-shot vision-language learning, covering multiple tasks such as VQA, SNLI-VE, and VCR. Our experiments show that our framework outperforms former zero-shot methods on VQA and achieves substantial improvement on SNLI-VE and VCR. Furthermore, our ablation studies confirm the effectiveness and generalizability of our proposed model.

Prompt Tuning for Unified Multimodal Pretrained Models
Hao Yang, Junyang Lin, An Yang, Peng Wang and Chang Zhou 11:00-12:30 (Pier 7&8)
Prompt tuning has become a new paradigm for model tuning and it has demonstrated success in natural language pretraining and even vision pretraining. The parameter-efficient prompt tuning methods that optimize soft embeddings while keeping the pretrained model frozen

demonstrate advantages in low computation costs and almost lossless performance. In this work, we explore the transfer of prompt tuning to multimodal pretrained models. Specifically, we implement prompt tuning to a unified sequence-to-sequence pretrained model by adding a sequence of learnable embeddings to each layer and finetuning the pretrained model on downstream task with only the learnable embeddings being optimized. Experimental results on a series of multimodal understanding and generation tasks demonstrate that our method OFA-PT can achieve comparable performance with finetuning across a series of multimodal generation and understanding tasks. Additionally, it significantly outperforms the unified multimodal pretrained model with other parameter-efficient tuning methods, e.g., Adapter, BitFit. etc. Besides, in comparison with finetuned models, the prompt-tuned models demonstrate improved robustness against adversarial attacks. We further figure out that experimental factors, including prompt length, prompt depth, and reparameterization, have great impacts on the model performance, and thus we empirically provide a recommendation for the setups of prompt tuning.

AV-TranSpeech: Audio-Visual Robust Speech-to-Speech Translation
Rongjie Huang, Huadai Liu, Xize Cheng, Yi Ren, Linjun Li, Zhenhui Ye, Jinzheng He, Lichao Zhang, Jinglin Liu, Xiang Yin and Zhou Zhao
11:00-12:30 (Pier 7&8)
Direct speech-to-speech translation (S2ST) aims to convert speech from one language into another, and has demonstrated significant progress to date. Despite the recent success, current S2ST models still suffer from distinct degradation in noisy environments and fail to translate visual speech (i.e., the movement of lips and teeth). In this work, we present AV-TranSpeech, the first audio-visual speech-to-speech (AV-S2ST) translation model without relying on intermediate text. AV-TranSpeech complements the audio stream with visual information to promote system robustness and opens up a host of practical applications: dictation or dubbing archival films. To mitigate the data scarcity with limited parallel AV-S2ST data, we 1) explore self-supervised pre-training with unlabeled audio-visual data to learn contextual representation, and 2) introduce cross-modal distillation with S2ST models trained on the audio-only corpus to further reduce the requirements of visual data. Experimental results on two language pairs demonstrate that AV-TranSpeech outperforms audio-only models under all settings regardless of the type of noise. With low-resource audio-visual data (10h, 30h), cross-modal distillation yields an improvement of 7.6 BLEU on average compared with baselines.Audio samples are available at https://AV-TranSpeech.github.io/.

Listen, Decipher and Sign: Toward Unsupervised Speech-to-Sign Language Recognition
Liming Wang, Junrui Ni, Heting Gao, Jialu Li, Kai Chieh Chang, Xulin Fan, Junkai Wu, Mark Hasegawa-Johnson and Chang D. Yoo 11:00-12:30 (Pier 7&8)
Existing supervised sign language recognition systems rely on an abundance of well-annotated data. Instead, an unsupervised speech-to-sign language recognition (SSR-U) system learns to translate between spoken and sign languages by observing only non-parallel speech and sign-language corpora. We propose Speech2sign-U, a neural network-based approach capable of both character-level and word-level SSR-U. Our approach significantly outperforms baselines directly adapted from unsupervised speech recognition (ASR-U) models by as much as 50% recall@10 on several challenging American sign language corpora with various levels of sample sizes, vocabulary sizes, and audio and visual variability. The code is available at https://github.com/cactuswiththoughts/UnsupSpeech2Sign.gitcactuswiththoughts/UnsupSpeech2Sign.git.

Joint Speech Transcription and Translation: Pseudo-Labeling with Out-of-Distribution Data
Mozhdeh Gheini, Tatiana Likhomanenko, Matthias Sperber and Hendra Setiawan 11:00-12:30 (Pier 7&8)
Self-training has been shown to be helpful in addressing data scarcity for many domains, including vision, speech, and language. Specifically, self-training, or pseudo-labeling, labels unsupervised data and adds that to the training pool. In this work, we investigate and use pseudo-labeling for a recently proposed novel setup: joint transcription and translation of speech, which suffers from an absence of sufficient parallel data resources. We show that under such data-deficient circumstances, the unlabeled data can significantly vary in domain from the supervised data, which results in pseudo-label quality degradation. We investigate two categories of remedies that require no additional supervision and target the domain mismatch: pseudo-label filtering and data augmentation. We show that pseudo-label analysis and processing in this way results in additional gains on top of the vanilla pseudo-labeling setup providing a total improvement of up to 0.4% absolute WER and 2.1 BLEU points for En–De and 0.6% absolute WER and 2.2 BLEU points for En–Zh.

Zero-shot Visual Question Answering with Language Model Feedback
Yifan Du, Junyi Li, Tianyi Tang, Wayne Xin Zhao and Ji-Rong Wen 11:00-12:30 (Pier 7&8)
In this paper, we propose a novel language model guided captioning approach, LAMOC, for knowledge-based visual question answering (VQA). Our approach employs the generated captions by a captioning model as the context of an answer prediction model, which is a Pre-Trained Language model (PLM). As the major contribution, we leverage the guidance and feedback of the prediction model to improve the capability of the captioning model. In this way, the captioning model can become aware of the task goal and information need from the PLM. To develop our approach, we design two specific training stages, where the first stage adapts the captioning model to the prediction model (selecting more suitable caption propositions for training) and the second stage tunes the captioning model according to the task goal (learning from feedback of the PLM). Extensive experiments demonstrate the effectiveness of the proposed approach on the knowledge-based VQA task. Specifically, on the challenging A-OKVQA dataset, LAMOC outperforms several competitive zero-shot methods and even achieves comparable results to a fine-tuned VLP model. Our code is publicly available at https://github.com/RUCAIBox/LAMOC.

Masked Audio Text Encoders are Effective Multi-Modal Rescorers
Jinglun Cai, Monica Sunkara, Xilai Li, Anshu Bhatia, Xiao Pan and Sravan Bodapati 11:00-12:30 (Pier 7&8)
Masked Language Models (MLMs) have proven to be effective for second-pass rescoring in Automatic Speech Recognition (ASR) systems. In this work, we propose Masked Audio Text Encoder (MATE), a multi-modal masked language model rescorer which incorporates acoustic representations into the input space of MLM. We adopt contrastive learning for effectively aligning the modalities by learning shared representations. We show that using a multi-modal rescorer is beneficial for domain generalization of the ASR system when target domain data is unavailable. MATE reduces word error rate (WER) by 4%-16% on in-domain, and 3%-7% on out-of-domain datasets, over the text-only baseline. Additionally, with very limited amount of training data (0.8 hours) MATE achieves a WER reduction of 8%-23% over the first-pass baseline.

Modality Adaption or Regularization? A Case Study on End-to-End Speech Translation
Yuchen Han, Chen Xu, Tong Xiao and Jingbo Zhu 11:00-12:30 (Pier 7&8)
Pre-training and fine-tuning is a paradigm for alleviating the data scarcity problem in end-to-end speech translation (E2E ST). The commonplace "modality gap" between speech and text data often leads to inconsistent inputs between pre-training and fine-tuning. However, we observe that this gap occurs in the early stages of fine-tuning, but does not have a major impact on the final performance. On the other hand, we find that there has another gap, which we call the "capacity gap": high resource tasks (such as ASR and MT) always require a large model to fit, when the model is reused for a low resource task (E2E ST), it will get a sub-optimal performance due to the over-fitting. In a case study, we find that the regularization plays a more important role than the well-designed modality adaption method, which achieves 29.0 for en-de and 40.3 for en-fr on the MuST-C dataset.

Subword Segmental Machine Translation: Unifying Segmentation and Target Sentence Generation
Francois Meyer and Jan Buys 11:00-12:30 (Pier 7&8)

Subword segmenters like BPE operate as a preprocessing step in neural machine translation and other (conditional) language models. They are applied to datasets before training, so translation or text generation quality relies on the quality of segmentations. We propose a departure from this paradigm, called subword segmental machine translation (SSMT). SSMT unifies subword segmentation and MT in a single trainable model. It learns to segment target sentence words while jointly learning to generate target sentences. To use SSMT during inference we propose dynamic decoding, a text generation algorithm that adapts segmentations as it generates translations. Experiments across 6 translation directions show that SSMT improves chrF scores for morphologically rich agglutinative languages. Gains are strongest in the very low-resource scenario. SSMT also learns subwords that are closer to morphemes compared to baselines and proves more robust on a test set constructed for evaluating morphological compositional generalisation.

An Investigation of Noise in Morphological Inflection
Adam Wiemerslage, Changbing Yang, Garrett Nicolai, Miikka Silfverberg and Katharina Kann   11:00-12:30 (Pier 7&8)
With a growing focus on morphological inflection systems for languages where high-quality data is scarce, training data noise is a serious but so far largely ignored concern. We aim at closing this gap by investigating the types of noise encountered within a pipeline for truly unsupervised morphological paradigm completion and its impact on morphological inflection systems: First, we propose an error taxonomy and annotation pipeline for inflection training data. Then, we compare the effect of different types of noise on multiple state-of-the- art inflection models. Finally, we propose a novel character-level masked language modeling (CMLM) pretraining objective and explore its impact on the models' resistance to noise. Our experiments show that various architectures are impacted differently by separate types of noise, but encoder-decoders tend to be more robust to noise than models trained with a copy bias. CMLM pretraining helps transformers, but has lower impact on LSTMs.

XSemPLR: Cross-Lingual Semantic Parsing in Multiple Natural Languages and Meaning Representations
Yusen Zhang, Jun Wang, Zhiguo Wang and Rui Zhang   11:00-12:30 (Pier 7&8)
Cross-Lingual Semantic Parsing (CLSP) aims to translate queries in multiple natural languages (NLs) into meaning representations (MRs) such as SQL, lambda calculus, and logic forms. However, existing CLSP models are separately proposed and evaluated on datasets of limited tasks and applications, impeding a comprehensive and unified evaluation of CLSP on a diverse range of NLs and MRs. To this end, we present XSemPLR, a unified benchmark for cross-lingual semantic parsing featured with 22 natural languages and 8 meaning representations by examining and selecting 9 existing datasets to cover 5 tasks and 164 domains. We use XSemPLR to conduct a comprehensive benchmark study on a wide range of multilingual language models including encoder-based models (mBERT, XLM-R), encoder-decoder models (mBART, mT5), and decoder-based models (Codex, BLOOM). We design 6 experiment settings covering various lingual combinations (monolingual, multilingual, cross-lingual) and numbers of learning samples (full dataset, few-shot, and zero-shot). Our experiments show that encoder-decoder models (mT5) achieve the highest performance compared with other popular models, and multilingual training can further improve the average performance. Notably, multilingual large language models (e.g., BLOOM) are still inadequate to perform CLSP tasks. We also find that the performance gap between monolingual training and cross-lingual transfer learning is still significant for multilingual models, though it can be mitigated by cross-lingual few-shot training. Our dataset and code are available at https://github.com/psunlpgroup/XSemPLR.

Pre-Trained Language-Meaning Models for Multilingual Parsing and Generation
Chunliu Wang, Huiyuan Lai, Malvina Nissim and Johan Bos   11:00-12:30 (Pier 7&8)
Pre-trained language models (PLMs) have achieved great success in NLP and have recently been used for tasks in computational semantics. However, these tasks do not fully benefit from PLMs since meaning representations are not explicitly included. We introduce multilingual pre-trained language-meaning models based on Discourse Representation Structures (DRSs), including meaning representations besides natural language texts in the same model, and design a new strategy to reduce the gap between the pre-training and fine-tuning objectives. Since DRSs are language neutral, cross-lingual transfer learning is adopted to further improve the performance of non-English tasks. Automatic evaluation results show that our approach achieves the best performance on both the multilingual DRS parsing and DRS-to-text generation tasks. Correlation analysis between automatic metrics and human judgements on the generation task further validates the effectiveness of our model. Human inspection reveals that out-of-vocabulary tokens are the main cause of erroneous results.

Enhancing Unsupervised Semantic Parsing with Distributed Contextual Representations
Zixuan Ling, Xiaoqing Zheng, Jianhan Xu, Jinshu Lin, Kai-Wei Chang, Cho-Jui Hsieh and Xuanjing Huang   11:00-12:30 (Pier 7&8)
We extend a non-parametric Bayesian model of (Titov and Klementiev, 2011) to deal with homonymy and polysemy by leveraging distributed contextual word and phrase representations pre-trained on a large collection of unlabelled texts. Then, unsupervised semantic parsing is performed by decomposing sentences into fragments, clustering the fragments to abstract away syntactic variations of the same meaning, and predicting predicate-argument relations between the fragments. To better model the statistical dependencies between predicates and their arguments, we further conduct a hierarchical Pitman-Yor process. An improved Metropolis-Hastings merge-split sampler is proposed to speed up the mixing and convergence of Markov chains by leveraging pre-trained distributed representations. The experimental results show that the models achieve better accuracy on both question-answering and relation extraction tasks.

A Pilot Study on Dialogue-Level Dependency Parsing for Chinese
Gongyao Jiang, Shuang Liu, Meishan Zhang and Min Zhang   11:00-12:30 (Pier 7&8)
Dialogue-level dependency parsing has received insufficient attention, especially for Chinese. To this end, we draw on ideas from syntactic dependency and rhetorical structure theory (RST), developing a high-quality human-annotated corpus, which contains 850 dialogues and 199,803 dependencies. Considering that such tasks suffer from high annotation costs, we investigate zero-shot and few-shot scenarios. Based on an existing syntactic treebank, we adopt a signal-based method to transform seen syntactic dependencies into unseen ones between elementary discourse units (EDUs), where the signals are detected by masked language modeling. Besides, we apply single-view and multi-view data selection to access reliable pseudo-labeled instances. Experimental results show the effectiveness of these baselines. Moreover, we discuss several crucial points about our dataset and approach.

Importance of Synthesizing High-quality Data for Text-to-SQL Parsing
Yiqun Hu, Yiyun Zhao, Jiarong Jiang, Wuwei Lan, Henghui Zhu, Anuj Chauhan, Alexander Hanbo Li, Lin Pan, Jun Wang, Chung-Wei Hang, Sheng Zhang, Jiang Guo, Mingwen Dong, Joseph Lilien, Patrick Ng, Zhiguo Wang, Vittorio Castelli and Bing Xiang   11:00-12:30 (Pier 7&8)
There has been increasing interest in synthesizing data to improve downstream text-to-SQL tasks. In this paper, we examined the existing synthesized datasets and discovered that state-of-the-art text-to-SQL algorithms did not further improve on popular benchmarks when trained with augmented synthetic data. We observed three shortcomings: illogical synthetic SQL queries from independent column sampling, arbitrary table joins, and language gaps between the synthesized SQL and natural language question (NLQ) pair. To address these issues, we propose a novel synthesis framework that imposes strong typing constraints, incorporates key relationships from schema, and conducts schema-distance-weighted column sampling. We also adopt an intermediate representation (IR) for the SQL-to-text task to further improve the quality of the generated NLQ. When existing powerful text-to-SQL parsers are pretrained on our high-quality synthesized data, these models have significant accuracy boosts and achieve new state-of-the-art performance on Spider. We also demonstrate the effectiveness of our techniques with ablation studies

A Self-Supervised Integration Method of Pretrained Language Models and Word Definitions
Hwiyeol Jo   11:00-12:30 (Pier 7&8)
We investigate the representation of pretrained language models and humans, using the idea of word definition modeling–how well a word is represented by its definition, and vice versa. Our analysis shows that a word representation in pretrained language models does not successfully map its human-written definition and its usage in example sentences. We then present a simple method DefBERT that integrates pretrained models with word semantics in dictionaries. We show its benefits on newly-proposed tasks of definition ranking and definition sense disambiguation. Furthermore, we present the results on standard word similarity tasks and short text classification tasks where models are required to encode semantics with only a few words. The results demonstrate the effectiveness of integrating word definitions and pretrained language models.

Together We Make Sense–Learning Meta-Sense Embeddings
Haochen Luo, Yi Zhou and Danushka Bollegala   11:00-12:30 (Pier 7&8)
Sense embedding learning methods learn multiple vectors for a given ambiguous word, corresponding to its different word senses. For this purpose, different methods have been proposed in prior work on sense embedding learning that use different sense inventories, sense-tagged corpora and learning methods. However, not all existing sense embeddings cover all senses of ambiguous words equally well due to the discrepancies in their training resources. To address this problem, we propose the first-ever meta-sense embedding method – Neighbour Preserving Meta-Sense Embeddings, which learns meta-sense embeddings by combining multiple independently trained source sense embeddings such that the sense neighbourhoods computed from the source embeddings are preserved in the meta-embedding space. Our proposed method can combine source sense embeddings that cover different sets of word senses. Experimental results on Word Sense Disambiguation (WSD) and Word-in-Context (WiC) tasks show that the proposed meta-sense embedding method consistently outperforms several competitive baselines.An anonymised version of the source code implementation for our proposed method is submitted to reviewing system. Both source code and the learnt meta-sense embeddings will be publicly released upon paper acceptance.

Taxonomy of Problems in Lexical Semantics
Bradley M Hauer and Grzegorz Kondrak   11:00-12:30 (Pier 7&8)
Semantic tasks are rarely formally defined, and the exact relationship between them is an open question. We introduce a taxonomy that elucidates the connection between several problems in lexical semantics, including monolingual and cross-lingual variants. Our theoretical framework is based on the hypothesis of the equivalence of concept and meaning distinctions. Using algorithmic problem reductions, we demonstrate that all problems in the taxonomy can be reduced to word sense disambiguation (WSD), and that WSD itself can be reduced to some problems, making them theoretically equivalent. In addition, we carry out experiments that strongly support the soundness of the concept-meaning hypothesis, and the correctness of our reductions.

Improving Diachronic Word Sense Induction with a Nonparametric Bayesian method
Ashjan Alsulaimani and Erwan Moreau   11:00-12:30 (Pier 7&8)
Diachronic Word Sense Induction (DWSI) is the task of inducing the temporal representations of a word meaning from the context, as a set of senses and their prevalence over time. We introduce two new models for DWSI, based on topic modelling techniques: one is based on Hierarchical Dirichlet Processes (HDP), a nonparametric model; the other is based on the Dynamic Embedded Topic Model (DETM), a recent dynamic neural model. We evaluate these models against two state of the art DWSI models, using a time-stamped labelled dataset from the biomedical domain. We demonstrate that the two proposed models perform better than the state of the art. In particular, the HDP-based model drastically outperforms all the other models, including the dynamic neural model.

DMLM: Descriptive Masked Language Modeling
Edoardo Barba, Niccolò Campolungo and Roberto Navigli   11:00-12:30 (Pier 7&8)
Over the last few years, Masked Language Modeling (MLM) pre-training has resulted in remarkable advancements in many Natural Language Understanding (NLU) tasks, which sparked an interest in researching alternatives and extensions to the MLM objective. In this paper, we tackle the absence of explicit semantic grounding in MLM and propose Descriptive Masked Language Modeling (DMLM), a knowledge-enhanced reading comprehension objective, where the model is required to predict the most likely word in a context, being provided with the word's definition. For instance, given the sentence "I was going to the _", if we provided as definition "financial institution", the model would have to predict the word "bank"; if, instead, we provided "sandy seashore", the model should predict "beach". Our evaluation highlights the effectiveness of DMLM in comparison with standard MLM, showing improvements on a number of well-established NLU benchmarks, as well as other semantics-focused tasks, e.g., Semantic Role Labeling. Furthermore, we demonstrate how it is possible to take full advantage of DMLM to embed explicit semantics in downstream tasks, explore several properties of DMLM-based contextual representations and suggest a number of future directions to investigate.

Ranking-Enhanced Unsupervised Sentence Representation Learning
Yeon Seonwoo, Guoyin Wang, Changmin Seo, Sajal Choudhary, Jiwei Li, Xiang Li, Puyang Xu, Sunghyun Park and Alice Oh   11:00-12:30 (Pier 7&8)
Unsupervised sentence representation learning has progressed through contrastive learning and data augmentation methods such as dropout masking. Despite this progress, sentence encoders are still limited to using only an input sentence when predicting its semantic vector. In this work, we show that the semantic meaning of a sentence is also determined by nearest-neighbor sentences that are similar to the input sentence. Based on this finding, we propose a novel unsupervised sentence encoder, RankEncoder. RankEncoder predicts the semantic vector of an input sentence by leveraging its relationship with other sentences in an external corpus, as well as the input sentence itself. We evaluate RankEncoder on semantic textual benchmark datasets. From the experimental results, we verify that 1) RankEncoder achieves 80.07% Spearman's correlation, a 1.1% absolute improvement compared to the previous state-of-the-art performance, 2) RankEncoder is universally applicable to existing unsupervised sentence embedding methods, and 3) RankEncoder is specifically effective for predicting the similarity scores of similar sentence pairs.

Align-then-Enhance: Multilingual Entailment Graph Enhancement with Soft Predicate Alignment
Yuting Wu, Yutong Hu, Yansong Feng, Tianyi Li, Mark Steedman and Dongyan Zhao   11:00-12:30 (Pier 7&8)
Entailment graphs (EGs) with predicates as nodes and entailment relations as edges are typically incomplete, while EGs in different languages are often complementary to each other. In this paper, we propose a new task, multilingual entailment graph enhancement, which aims to utilize the entailment information from one EG to enhance another EG in a different language. The ultimate goal is to obtain an enhanced EG containing richer and more accurate entailment information. We present an align-then-enhance framework (ATE) to achieve accurate multilingual entailment graph enhancement, which first exploits a cross-graph guided interaction mechanism to automatically discover potential equivalent predicates between different EGs and then constructs more accurate enhanced entailment graphs based on soft predicate alignments. Extensive experiments show that ATE achieves better and more robust predicate alignment results between different EGs, and the enhanced entailment graphs generated by ATE outperform the original graphs for entailment detection.

PIP: Parse-Instructed Prefix for Syntactically Controlled Paraphrase Generation

Yixin Wan, Kuan-Hao Huang and Kai-Wei Chang   11:00-12:30 (Pier 7&8)

Syntactically controlled paraphrase generation requires language models to generate paraphrases for sentences according to specific syntactic structures. Existing fine-tuning methods on this task is costly, as all parameters of the model need to be updated during the training process. Inspired by recent studies on parameter-efficient learning, we propose Parse-Instructed Prefix (PIP), a novel adaptation of prefix-tuning to tune large pre-trained language models on syntactically controlled paraphrase generation task in a low-data setting with significantly less training cost. We introduce two methods to instruct a model's encoder prefix to capture syntax-related knowledge: direct initiation (PIP-Direct) and indirect optimization (PIP-Indirect). Comparing to traditional fine-tuning methods for this task, PIP is a compute-efficient alternative with 10 times less learnable parameters. Comparing to existing prefix-tuning methods, PIP excels at capturing syntax control information, achieving significantly higher performance at the same level of learnable parameter count.

Conjunct Lengths in English, Dependency Length Minimization, and Dependency Structure of Coordination
Adam Przepiórkowski and Michał Woźniak   11:00-12:30 (Pier 7&8)

This paper confirms that, in English binary coordinations, left conjuncts tend to be shorter than right conjuncts, regardless of the position of the governor of the coordination. We demonstrate that this tendency becomes stronger when length differences are greater, but only when the governor is on the left or absent, not when it is on the right. We explain this effect via Dependency Length Minimization and we show that this explanation provides support for symmetrical dependency structures of coordination (where coordination is multi-headed by all conjuncts, as in Word Grammar or in enhanced Universal Dependencies, or where it single-headed by the conjunction, as in the Prague Dependency Treebank), as opposed to asymmetrical structures (where coordination is headed by the first conjunct, as in the Meaning–Text Theory or in basic Universal Dependencies).

Automatic Readability Assessment for Closely Related Languages
Joseph Marvin Imperial and Ekaterina Kochmar   11:00-12:30 (Pier 7&8)

In recent years, the main focus of research on automatic readability assessment (ARA) has shifted towards using expensive deep learning-based methods with the primary goal of increasing models' accuracy. This, however, is rarely applicable for low-resource languages where traditional handcrafted features are still widely used due to the lack of existing NLP tools to extract deeper linguistic representations. In this work, we take a step back from the technical component and focus on how linguistic aspects such as mutual intelligibility or degree of language relatedness can improve ARA in a low-resource setting. We collect short stories written in three languages in the Philippines—Tagalog, Bikol, and Cebuano—to train readability assessment models and explore the interaction of data and features in various cross-lingual setups. Our results show that the inclusion of CrossNGO, a novel specialized feature exploiting n-gram overlap applied to languages with high mutual intelligibility, significantly improves the performance of ARA models compared to the use of off-the-shelf large multilingual language models alone. Consequently, when both linguistic representations are combined, we achieve state-of-the-art results for Tagalog and Cebuano, and baseline scores for ARA in Bikol.

Unsupervised Semantic Variation Prediction using the Distribution of Sibling Embeddings
Taichi Aida and Danushka Bollegala   11:00-12:30 (Pier 7&8)

Languages are dynamic entities, where the meanings associated with words constantly change with time. Detecting the semantic variation of words is an important task for various NLP applications that must make time-sensitive predictions. Existing work on semantic variation prediction have predominantly focused on comparing some form of an averaged contextualised representation of a target word computed from a given corpus. However, some of the previously associated meanings of a target word can become obsolete over time (e.g. meaning of gay as happy), while novel usages of existing words are observed (e.g. meaning of cell as a mobile phone). We argue that mean representations alone cannot accurately capture such semantic variations and propose a method that uses the entire cohort of the contextualised embeddings of the target word, which we refer to as the sibling distribution. Experimental results on SemEval-2020 Task 1 benchmark dataset for semantic variation prediction show that our method outperforms prior work that consider only the mean embeddings, and is comparable to the current state-of-the-art. Moreover, a qualitative analysis shows that our method detects important semantic changes in words that are not captured by the existing methods.

LMs stand their Ground: Investigating the Effect of Embodiment in Figurative Language Interpretation by Language Models
Philipp Wicke   11:00-12:30 (Pier 7&8)

Figurative language is a challenge for language models since its interpretation is based on the use of words in a way that deviates from their conventional order and meaning. Yet, humans can easily understand and interpret metaphors, similes or idioms as they can be derived from embodied metaphors. Language is a proxy for embodiment and if a metaphor is conventional and lexicalised, it becomes easier for a system without a body to make sense of embodied concepts. Yet, the intricate relation between embodiment and features such as concreteness or age of acquisition has not been studied in the context of figurative language interpretation concerning language models. Hence, the presented study shows how larger language models perform better at interpreting metaphoric sentences when the action of the metaphorical sentence is more embodied. The analysis rules out multicollinearity with other features (e.g. word length or concreteness) and provides initial evidence that larger language models conceptualise embodied concepts to a degree that facilitates figurative language understanding.

Language acquisition: do children and language models follow similar learning stages?
Linnea Evanson, Yair Lakretz and Jean Rémi king   11:00-12:30 (Pier 7&8)

During language acquisition, children follow a typical sequence of learning stages, whereby they first learn to categorize phonemes before they develop their lexicon and eventually master increasingly complex syntactic structures. However, the computational principles that lead to this learning trajectory remain largely unknown. To investigate this, we here compare the learning trajectories of deep language models to those of human children. Specifically, we test whether, during its training, GPT-2 exhibits stages of language acquisition comparable to those observed in children aged between 18 months and 6 years. For this, we train 48 GPT-2 models from scratch and evaluate their syntactic and semantic abilities at each training step, using 96 probes curated from the BLiMP, Zorro and BIG-Bench benchmarks. We then compare these evaluations with the behavior of 54 children during language production. Our analyses reveal three main findings. First, similarly to children, the language models tend to learn linguistic skills in a systematic order. Second, this learning scheme is parallel: the language tasks that are learned last improve from the very first training steps. Third, some – but not all – learning stages are shared between children and these language models. Overall, these results shed new light on the principles of language acquisition, and highlight important divergences in how humans and modern algorithms learn to process natural language.

UniCoRN: Unified Cognitive Signal ReconstructioN bridging cognitive signals and human language
Nuwa Xi, Sendong Zhao, Haochun Wang, Chi Liu, Bing Qin and Ting Liu   11:00-12:30 (Pier 7&8)

Decoding text stimuli from cognitive signals (e.g. fMRI) enhances our understanding of the human language system, paving the way for building versatile Brain-Computer Interface. However, existing studies largely focus on decoding individual word-level fMRI volumes from a restricted vocabulary, which is far too idealized for real-world application. In this paper, we propose fMRI2text, the first open-vocabulary task aiming to bridge fMRI time series and human language. Furthermore, to explore the potential of this new task, we present a baseline solution, UniCoRN: the Unified Cognitive Signal ReconstructioN for Brain Decoding. By reconstructing both individual time points and time series, UniCoRN establishes a robust encoder for cognitive signals (fMRI & EEG). Leveraging a pre-trained language model as decoder, UniCoRN

proves its efficacy in decoding coherent text from fMRI series across various split settings. Our model achieves a 34.77% BLEU score on fMRI2text, and a 37.04% BLEU when generalized to EEG-to-text decoding, thereby surpassing the former baseline. Experimental results indicate the feasibility of decoding consecutive fMRI volumes, and the effectiveness of decoding different cognitive signals using a unified structure.

Distributed Marker Representation for Ambiguous Discourse Markers and Entangled Relations
Dongyu Ru, Lin Qiu, Xipeng Qiu, Yue Zhang and Zheng Zhang　　　　　　　　　　　　　　　　11:00-12:30 (Pier 7&8)
Discourse analysis is an important task because it models intrinsic semantic structures between sentences in a document. Discourse markers are natural representations of discourse in our daily language. One challenge is that the markers as well as pre-defined and human-labeled discourse relations can be ambiguous when describing the semantics between sentences. We believe that a better approach is to use a contextual-dependent distribution over the markers to express discourse information. In this work, we propose to learn a Distributed Marker Representation (DMR) by utilizing the (potentially) unlimited discourse marker data with a latent discourse sense, thereby bridging markers with sentence pairs. Such representations can be learned automatically from data without supervision, and in turn provide insights into the data itself. Experiments show the SOTA performance of our DMR on the implicit discourse relation recognition task and strong interpretability. Our method also offers a valuable tool to understand complex ambiguity and entanglement among discourse markers and manually defined discourse relations.

SERENGETI: Massively Multilingual Language Models for Africa
Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed and Alcides Alcoba Inciarte　　　11:00-12:30 (Pier 7&8)
Multilingual pretrained language models (mPLMs) acquire valuable, generalizable linguistic information during pretraining and have advanced the state of the art on task-specific finetuning. To date, only 31 out of 2,000 African languages are covered in existing language models. We ameliorate this limitation by developing SERENGETI, a set of massively multilingual language model that covers 517 African languages and language varieties. We evaluate our novel models on eight natural language understanding tasks across 20 datasets, comparing to 4 mPLMs that cover 4-23 African languages. SERENGETI outperforms other models on 11 datasets across the eights tasks, achieving 82.27 average F_1. We also perform analyses of errors from our models, which allows us to investigate the influence of language genealogy and linguistic similarity when the models are applied under zero-shot settings. We will publicly release our models for research.Anonymous link

Language Agnostic Multilingual Information Retrieval with Contrastive Learning
Xiyang Hu, Xinchi Chen, Peng Qi, Deguang Kong, Kunlun Liu, William Yang Wang and zhiheng huang　　11:00-12:30 (Pier 7&8)
Multilingual information retrieval (IR) is challenging since annotated training data is costly to obtain in many languages. We present an effective method to train multilingual IR systems when only English IR training data and some parallel corpora between English and other languages are available. We leverage parallel and non-parallel corpora to improve the pretrained multilingual language models' cross-lingual transfer ability. We design a semantic contrastive loss to align representations of parallel sentences that share the same semantics in different languages, and a new language contrastive loss to leverage parallel sentence pairs to remove language-specific information in sentence representations from non-parallel corpora. When trained on English IR data with these losses and evaluated zero-shot on non-English data, our model demonstrates significant improvement to prior work on retrieval performance, while it requires much less computational effort. We also demonstrate the value of our model for a practical setting when a parallel corpus is only available for a few languages, but a lack of parallel corpora resources persists for many other low-resource languages. Our model can work well even with a small number of parallel sentences, and be used as an add-on module to any backbones and other tasks.

Can Cross-Lingual Transferability of Multilingual Transformers Be Activated Without End-Task Data?
Zewen Chi, Heyan Huang and Xian-Ling Mao　　　　　　　　　　　　　　　　　　　　11:00-12:30 (Pier 7&8)
Pretrained multilingual Transformers have achieved great success in cross-lingual transfer learning. Current methods typically activate the cross-lingual transferability of multilingual Transformers by fine-tuning them on end-task data. However, the methods cannot perform cross-lingual transfer when end-task data are unavailable. In this work, we explore whether the cross-lingual transferability can be activated without end-task data. We propose a cross-lingual transfer method, named PlugIn-X. PlugIn-X disassembles monolingual and multilingual Transformers into sub-modules, and reassembles them back to be the multilingual end-task model. After representation adaptation, PlugIn-X finally performs cross-lingual transfer in a plug-and-play style. Experimental results show that PlugIn-X successfully activates the cross-lingual transferability of multilingual Transformers without accessing end-task data. Moreover, we analyze how the cross-model representation alignment affects the cross-lingual transferability.

Language Anisotropic Cross-Lingual Model Editing
Yang Xu, Yutai Hou, Wanxiang Che and Min Zhang　　　　　　　　　　　　　　　　　11:00-12:30 (Pier 7&8)
Multilingual pre-trained language models can learn task-specific abilities or memorize facts across multiple languages but inevitably make undesired predictions with specific inputs. Under similar observation, model editing aims to post-hoc calibrate a model targeted to specific inputs with keeping the model's raw behavior. However, existing work only studies the monolingual scenario, which lacks the cross-lingual transferability to perform editing simultaneously across languages. In this work, we focus on cross-lingual model editing. Firstly, we define the cross-lingual model editing task and corresponding metrics, where an edit in one language propagates to the others. Next, we propose a framework to naturally adapt monolingual model editing approaches to the cross-lingual scenario using parallel corpus. Further, we propose language anisotropic editing to improve cross-lingual editing by amplifying different subsets of parameters for each language. On the newly defined cross-lingual model editing task, we empirically demonstrate the failure of monolingual baselines in propagating the edit to multiple languages and the effectiveness of the proposed language anisotropic model editing. Our code is publicly available at https://github.com/franklear/LiME.

Why Does Zero-Shot Cross-Lingual Generation Fail? An Explanation and a Solution
Tianjian Li and Kenton Murray　　　　　　　　　　　　　　　　　　　　　　　　11:00-12:30 (Pier 7&8)
Zero-shot cross-lingual transfer is when a multilingual model is trained to perform a task in one language and then is applied to another language. Although the zero-shot cross-lingual transfer approach has achieved success in various classification tasks, its performance on natural language generation tasks falls short in quality and sometimes outputs an incorrect language. In our study, we show that the fine-tuning process learns language invariant representations, which is beneficial for classification tasks but harmful for generation tasks. Motivated by this, we propose a simple method to regularize the model from learning language invariant representations and a method to select model checkpoints without a development set in the target language, both resulting in better generation quality. Experiments on three semantically diverse generation tasks show that our method reduces the accidental translation problem by 68% and improves the ROUGE-L score by 1.5 on average.

Multi-VALUE: A Framework for Cross-Dialectal English NLP
Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta and Diyi Yang　　　11:00-12:30 (Pier 7&8)
Dialect differences caused by regional, social, and economic factors cause performance discrepancies for many groups of language technology users. Inclusive and equitable language technology must critically be dialect invariant, meaning that performance remains constant over

dialectal shifts. Current systems often fall short of this ideal since they are designed and tested on a single dialect: Standard American English (SAE). We introduce a suite of resources for evaluating and achieving English dialect invariance. The resource is called Multi-VALUE, a controllable rule-based translation system spanning 50 English dialects and 189 unique linguistic features. Multi-VALUE maps SAE to synthetic forms of each dialect. First, we use this system to stress tests question answering, machine translation, and semantic parsing. Stress tests reveal significant performance disparities for leading models on non-standard dialects. Second, we use this system as a data augmentation technique to improve the dialect robustness of existing systems. Finally, we partner with native speakers of Chicano and Indian English to release new gold-standard variants of the popular CoQA task. To execute the transformation code, run model checkpoints, and download both synthetic and gold-standard dialectal benchmark datasets, see http://value-nlp.org.

Cross-Lingual Retrieval Augmented Prompt for Low-Resource Languages
Ercong Nie, Sheng Liang, Helmut Schmid and Hinrich Schütze   11:00-12:30 (Pier 7&8)
Multilingual Pretrained Language Models (MPLMs) perform strongly in cross-lingual transfer. We propose Prompts Augmented by Retrieval Crosslingually (PARC) to improve zero-shot performance on low-resource languages (LRLs) by augmenting the context with prompts consisting of semantically similar sentences retrieved from a high-resource language (HRL). PARC improves zero-shot performance on three downstream tasks (sentiment classification, topic categorization, natural language inference) with multilingual parallel test sets across 10 LRLs covering 6 language families in unlabeled (+5.1%) and labeled settings (+16.3%). PARC also outperforms finetuning by 3.7%. We find a significant positive correlation between cross-lingual transfer performance on one side, and the similarity between high- and low-resource languages as well as the amount of low-resource pretraining data on the other side. A robustness analysis suggests that PARC has the potential to achieve even stronger performance with more powerful MPLMs.

Tokenization Impacts Multilingual Language Modeling: Assessing Vocabulary Allocation and Overlap Across Languages
Tomasz Limisiewicz, Jiří Balhar and David Mareček   11:00-12:30 (Pier 7&8)
Multilingual language models have recently gained attention as a promising solution for representing multiple languages in a single model. In this paper, we propose new criteria to evaluate the quality of lexical representation and vocabulary overlap observed in sub-word tokenizers. Our findings show that the overlap of vocabulary across languages can be actually detrimental to certain downstream tasks (POS, dependency tree labeling). In contrast, NER and sentence-level tasks (cross-lingual retrieval, NLI) benefit from sharing vocabulary. We also observe that the coverage of the language-specific tokens in the multilingual vocabulary significantly impacts the word-level tasks. Our study offers a deeper understanding of the role of tokenizers in multilingual language models and guidelines for future model developers to choose the most suitable tokenizer for their specific application before undertaking costly model pre-training.

Code-Switched Text Synthesis in Unseen Language Pairs
I-Hung Hsu, Avik Ray, Shubham Garg, Nanyun Peng and Jing Huang   11:00-12:30 (Pier 7&8)
Existing efforts on text synthesis for code-switching mostly require training on code-switched texts in the target language pairs, limiting the deployment of the models to cases lacking code-switched data. In this work, we study the problem of synthesizing code-switched texts for language pairs absent from the training data. We introduce GLOSS, a model built on top of a pre-trained multilingual machine translation model (PMMTM) with an additional code-switching module. This module, either an adapter or extra prefixes, learns code-switching patterns from code-switched data during training, while the primary component of GLOSS, i.e., the PMMTM, is frozen. The design of only adjusting the code-switching module prevents our model from overfitting to the constrained training data for code-switching. Hence, GLOSS exhibits the ability to generalize and synthesize code-switched texts across a broader spectrum of language pairs. Additionally, we develop a self-training algorithm on target language pairs further to enhance the reliability of GLOSS. Automatic evaluations on four language pairs show that GLOSS achieves at least 55% relative BLEU and METEOR scores improvements compared to strong baselines. Human evaluations on two language pairs further validate the success of GLOSS.

Exploring Anisotropy and Outliers in Multilingual Language Models for Cross-Lingual Semantic Sentence Similarity
Katharina Haemmerl, Alina Fastowski, Jindřich Libovický and Alexander Fraser   11:00-12:30 (Pier 7&8)
Previous work has shown that the representations output by contextual language models are more anisotropic than static type embeddings, and typically display outlier dimensions. This seems to be true for both monolingual and multilingual models, although much less work has been done on the multilingual context. Why these outliers occur and how they affect the representations is still an active area of research. We investigate outlier dimensions and their relationship to anisotropy in multiple pre-trained multilingual language models. We focus on cross-lingual semantic similarity tasks, as these are natural tasks for evaluating multilingual representations. Specifically, we examine sentence representations. Sentence transformers which are fine-tuned on parallel resources (that are not always available) perform better on this task, and we show that their representations are more isotropic. However, we aim to improve multilingual representations in general. We investigate how much of the performance difference can be made up by only transforming the embedding space without fine-tuning, and visualise the resulting spaces. We test different operations: Removing individual outlier dimensions, cluster-based isotropy enhancement, and ZCA whitening. We publish our code for reproducibility.

Frustratingly Easy Label Projection for Cross-lingual Transfer
Yang Chen, Chao Jiang, Alan Ritter and Wei Xu   11:00-12:30 (Pier 7&8)
Translating training data into many languages has emerged as a practical solution for improving cross-lingual transfer. For tasks that involve span-level annotations, such as information extraction or question answering, an additional label projection step is required to map annotated spans onto the translated texts. Recently, a few efforts have utilized a simple mark-then-translate method to jointly perform translation and projection by inserting special markers around the labeled spans in the original sentence. However, as far as we are aware, no empirical analysis has been conducted on how this approach compares to traditional annotation projection based on word alignment. In this paper, we present an extensive empirical study across 57 languages and three tasks (QA, NER, and Event Extraction) to evaluate the effectiveness and limitations of both methods, filling an important gap in the literature. Experimental results show that our optimized version of mark-then-translate, which we call EasyProject, is easily applied to many languages and works surprisingly well, outperforming the more complex word alignment-based methods. We analyze several key factors that affect the end-task performance, and show EasyProject works well because it can accurately preserve label span boundaries after translation. We will publicly release all our code and data.

Adversarial Training for Low-Resource Disfluency Correction
Vineet Bhat, Preethi Jyothi and Pushpak Bhattacharyya   11:00-12:30 (Pier 7&8)
Disfluencies commonly occur in conversational speech. Speech with disfluencies can result in noisy Automatic Speech Recognition (ASR) transcripts, which affects downstream tasks like machine translation. In this paper, we propose an adversarially-trained sequence-tagging model for Disfluency Correction (DC) that utilizes a small amount of labeled real disfluent data in conjunction with a large amount of unlabeled data. We show the benefit of our proposed technique, which crucially depends on synthetically generated disfluent data, by evaluating it for DC in three Indian languages- Bengali, Hindi, and Marathi (all from the Indo-Aryan family). Our technique also performs well in removing stuttering disfluencies in ASR transcripts introduced by speech impairments. We achieve an average 6.15 points improvement in F1-score over competitive baselines across all three languages mentioned. To the best of our knowledge, we are the first to utilize adversarial training for DC and use it to correct stuttering disfluencies in English, establishing a new benchmark for this task.

CoLaDa: A Collaborative Label Denoising Framework for Cross-lingual Named Entity Recognition
Tingting Ma, Qianhui Wu, Huiqiang Jiang, Börje F. Karlsson, Tiejun Zhao and Chin-Yew Lin   11:00-12:30 (Pier 7&8)
Cross-lingual named entity recognition (NER) aims to train an NER system that generalizes well to a target language by leveraging labeled data in a given source language. Previous work alleviates the data scarcity problem by translating source-language labeled data or performing knowledge distillation on target-language unlabeled data. However, these methods may suffer from label noise due to the automatic labeling process. In this paper, we propose CoLaDa, a Collaborative Label Denoising Framework, to address this problem. Specifically, we first explore a model-collaboration-based denoising scheme that enables models trained on different data sources to collaboratively denoise pseudo labels used by each other. We then present an instance-collaboration-based strategy that considers the label consistency of each token's neighborhood in the representation space for denoising. Experiments on different benchmark datasets show that the proposed CoLaDa achieves superior results compared to previous methods, especially when generalizing to distant languages.

Enhancing Few-shot Cross-lingual Transfer with Target Language Peculiar Examples
Hwichan Kim and Mamoru Komachi   11:00-12:30 (Pier 7&8)
Few-shot cross-lingual transfer, fine-tuning Multilingual Masked Language Model (MMLM) with source language labeled data and a small amount of target language labeled data, provides excellent performance in the target language. However, if no labeled data in the target language are available, they need to be created through human annotations. In this study, we devise a metric to select annotation candidates from an unlabeled data pool that efficiently enhance accuracy for few-shot cross-lingual transfer. It is known that training a model with hard examples is important to improve the model's performance. Therefore, we first identify examples that MMLM cannot solve in a zero-shot cross-lingual transfer setting and demonstrate that it is hard to predict peculiar examples in the target language, i.e., the examples distant from the source language examples in cross-lingual semantic space of the MMLM. We then choose high peculiarity examples as annotation candidates and perform few-shot cross-lingual transfer. In comprehensive experiments with 20 languages and 6 tasks, we demonstrate that the high peculiarity examples improve the target language accuracy compared to other candidate selection methods proposed in previous studies.

Improving Pretraining Techniques for Code-Switched NLP
Richeek Das, Sahasra Ranjan, Shreya Pathak and Preethi Jyothi   11:00-12:30 (Pier 7&8)
Pretrained models are a mainstay in modern NLP applications. Pretraining requires access to large volumes of unlabeled text. While monolingual text is readily available for many of the world's languages, access to large quantities of code-switched text (i.e., text with tokens of multiple languages interspersed within a sentence) is much more scarce. Given this resource constraint, the question of how pretraining using limited amounts of code-switched text could be altered to improve performance for code-switched NLP becomes important to tackle. In this paper, we explore different masked language modeling (MLM) pretraining techniques for code-switched text that are cognizant of language boundaries prior to masking. The language identity of the tokens can either come from human annotators, trained language classifiers, or simple relative frequency-based estimates. We also present an MLM variant by introducing a residual connection from an earlier layer in the pretrained model that uniformly boosts performance on downstream tasks. Experiments on two downstream tasks, Question Answering (QA) and Sentiment Analysis (SA), involving four code-switched language pairs (Hindi-English, Spanish-English, Tamil-English, Malayalam-English) yield relative improvements of up to 5.8 and 2.7 F1 scores on QA (Hindi-English) and SA (Tamil-English), respectively, compared to standard pretraining techniques. To understand our task improvements better, we use a series of probes to study what additional information is encoded by our pretraining techniques and also introduce an auxiliary loss function that explicitly models language identification to further aid the residual MLM variants.

A Hierarchical Explanation Generation Method Based on Feature Interaction Detection
Yiming Ju, Yuanzhe Zhang, Kang Liu and Jun Zhao   11:00-12:30 (Pier 7&8)
The opaqueness of deep NLP models has motivated efforts to explain how deep models predict. Recently, work has introduced hierarchical attribution explanations, which calculate attribution scores for compositional text hierarchically to capture compositional semantics. Existing work on hierarchical attributions tends to limit the text groups to a continuous text span, which we call the connecting rule. While easy for humans to read, limiting the attribution unit to a continuous span might lose important long-distance feature interactions for reflecting model predictions. In this work, we introduce a novel strategy for capturing feature interactions and employ it to build hierarchical explanations without the connecting rule. The proposed method can convert ubiquitous non-hierarchical explanations (e.g., LIME) into their corresponding hierarchical versions. Experimental results show the effectiveness of our approach in building high-quality hierarchical explanations.

Emergent Modularity in Pre-trained Transformers
Zhengyan Zhang, Zhiyuan Zeng, Yankai Lin, Chaojun Xiao, Xiaozhi Wang, Xu Han, Zhiyuan Liu, Ruobing Xie, Maosong Sun and Jie Zhou
11:00-12:30 (Pier 7&8)
This work examines the presence of modularity in pre-trained Transformers, a feature commonly found in human brains and thought to be vital for general intelligence. In analogy to human brains, we consider two main characteristics of modularity: (1) functional specialization of neurons: we evaluate whether each neuron is mainly specialized in a certain function, and find that the answer is yes. (2) function-based neuron grouping: we explore to find a structure that groups neurons into modules by function, and each module works for its corresponding function. Given the enormous amount of possible structures, we focus on Mixture-of-Experts as a promising candidate, which partitions neurons into experts and usually activates different experts for different inputs. Experimental results show that there are functional experts, where clustered are the neurons specialized in a certain function. Moreover, perturbing the activations of functional experts significantly affects the corresponding function. Finally, we study how modularity emerges during pre-training, and find that the modular structure is stabilized at the early stage, which is faster than neuron stabilization. It suggests that Transformer first constructs the modular structure and then learns fine-grained neuron functions. Our code and data are available at https://github.com/THUNLP/modularity-analysis.

Counterfactuals of Counterfactuals: a back-translation-inspired approach to analyse counterfactual editors
George Filandrianos, Edmund G Dervakos, Orfeas Menis Mastromichalakis, Chrysoula Zerva and Giorgos Stamou   11:00-12:30 (Pier 7&8)
In the wake of responsible AI, interpretability methods, which attempt to provide an explanation for the predictions of neural models have seen rapid progress. In this work, we are concerned with explanations that are applicable to natural language processing (NLP) models and tasks, and we focus specifically on the analysis of counterfactual, contrastive explanations. We note that while there have been several explainers proposed to produce counterfactual explanations, their behaviour can vary significantly and the lack of a universal ground truth for the counterfactual edits imposes an insuperable barrier on their evaluation. We propose a new back-translation-inspired evaluation methodology that utilises earlier outputs of the explainer as ground truth proxies to investigate the consistency of explainers. We show that by iteratively feeding the counterfactual to the explainer we can obtain valuable insights into the behaviour of both the predictor and the explainer models, and infer patterns that would be otherwise obscured. Using this methodology, we conduct a thorough analysis and propose a novel metric to evaluate the consistency of counterfactual generation approaches with different characteristics across available performance indicators.

Fighting Bias With Bias: Promoting Model Robustness by Amplifying Dataset Biases
Yuval Reif and Roy Schwartz   11:00-12:30 (Pier 7&8)
NLP models often rely on superficial cues known as dataset biases to achieve impressive performance, and can fail on examples where these

biases do not hold. Recent work sought to develop robust, unbiased models by filtering biased examples from training sets. In this work, we argue that such filtering can obscure the true capabilities of models to overcome biases, which might never be removed in full from the dataset. We suggest that in order to drive the development of models robust to subtle biases, dataset biases should be amplified in the training set. We introduce an evaluation framework defined by a bias-amplified training set and an anti-biased test set, both automatically extracted from existing datasets. Experiments across three notions of bias, four datasets and two models show that our framework is substantially more challenging for models than the original data splits, and even more challenging than hand-crafted challenge sets. Our evaluation framework can use any existing dataset, even those considered obsolete, to test model robustness. We hope our work will guide the development of robust models that do not rely on superficial biases and correlations. To this end, we publicly release our code and data.

A Close Look into the Calibration of Pre-trained Language Models
Yangyi Chen, Lifan Yuan, Ganqu CUI, Zhiyuan Liu and Heng Ji   11:00-12:30 (Pier 7&8)
Pre-trained language models (PLMs) may fail in giving reliable estimates of their predictive uncertainty. We take a close look into this problem, aiming to answer two questions: (1) Do PLMs learn to become calibrated in the training process? (2) How effective are existing calibration methods? For the first question, we conduct fine-grained control experiments to study the dynamic change in PLMs' calibration performance in training. We consider six factors as control variables, including dataset difficulty, available training samples, training steps, the number of tunable parameters, model scale, and pretraining. We observe a consistent change in calibration performance across six factors. We find that PLMs don't learn to become calibrated in training, evidenced by the continual increase in confidence, no matter whether the predictions are correct or not. We highlight that our finding somewhat contradicts two established conclusions: (a) Larger PLMs are more calibrated; (b) Pretraining improves model calibration. Next, we study the effectiveness of existing calibration methods in mitigating the overconfidence issue. Besides unlearnable calibration methods (e.g., label smoothing), we adapt and extend two recently proposed learnable methods that directly collect data to train models to have reasonable confidence estimations. Experimental results show that learnable methods significantly reduce PLMs' confidence in wrong predictions.

Robustness of Learning from Task Instructions
Jiasheng Gu, Hongyu Zhao, Hanzi Xu, Liangyu Nie, Hongyuan Mei and Wenpeng Yin   11:00-12:30 (Pier 7&8)
Traditional supervised learning mostly works on individual tasks and requires training on a large set of task-specific examples. This paradigm seriously hinders the development of task generalization since preparing a task-specific example set is costly. To build a system that can quickly and easily generalize to new tasks, task instructions have been adopted as an emerging trend of supervision recently. These instructions give the model the definition of the task and allow the model to output the appropriate answer based on the instructions and inputs. However, task instructions are often expressed in different forms, which can be interpreted from two threads: first, some instructions are short sentences and are pretrained language model (PLM) oriented, such as prompts, while other instructions are paragraphs and are human-oriented, such as those in Amazon MTurk; second, different end-users very likely explain the same task with instructions of different textual expressions. A robust system for task generalization should be able to handle any new tasks regardless of the variability of instructions.

However, the system robustness in dealing with instruction-driven task generalization is still unexplored. This work investigates the system robustness when the instructions of new tasks are (i) manipulated, (ii) paraphrased, or (iii) from different levels of conciseness. To our knowledge, this is the first work that systematically studies how robust a PLM is when it is supervised by instructions with different factors of variability.

SenteCon: Leveraging Lexicons to Learn Human-Interpretable Language Representations
Victoria Lin and Louis-Philippe Morency   11:00-12:30 (Pier 7&8)
Although deep language representations have become the dominant form of language featurization in recent years, in many settings it is important to understand a model's decision-making process. This necessitates not only an interpretable model but also interpretable features. In particular, language must be featurized in a way that is interpretable while still characterizing the original text well. We present SenteCon, a method for introducing human interpretability in deep language representations. Given a passage of text, SenteCon encodes the text as a layer of interpretable categories in which each dimension corresponds to the relevance of a specific category. Our empirical evaluations indicate that encoding language with SenteCon provides high-level interpretability at little to no cost to predictive performance on downstream tasks. Moreover, we find that SenteCon outperforms existing interpretable language representations with respect to both its downstream performance and its agreement with human characterizations of the text.

Adversarial Robustness of Prompt-based Few-Shot Learning for Natural Language Understanding
Venkata Prabhakara Sarath Nookala, Gaurav Verma, Subhabrata Mukherjee and Srijan Kumar   11:00-12:30 (Pier 7&8)
State-of-the-art few-shot learning (FSL) methods leverage prompt-based fine-tuning to obtain remarkable results for natural language understanding (NLU) tasks. While much of the prior FSL methods focus on improving downstream task performance, there is a limited understanding of the adversarial robustness of such methods. In this work, we conduct an extensive study of several state-of-the-art FSL methods to assess their robustness to adversarial perturbations. To better understand the impact of various factors towards robustness (or the lack of it), we evaluate prompt-based FSL methods against fully fine-tuned models for aspects such as the use of unlabeled data, multiple prompts, number of few-shot examples, model size and type. Our results on six GLUE tasks indicate that compared to fully fine-tuned models, vanilla FSL methods lead to a notable relative drop in task performance (i.e., are less robust) in the face of adversarial perturbations. However, using (i) unlabeled data for prompt-based FSL and (ii) multiple prompts flip the trend – the few-shot learning approaches demonstrate a lesser drop in task performance than fully fine-tuned models. We further demonstrate that increasing the number of few-shot examples and model size lead to increased adversarial robustness of vanilla FSL methods. Broadly, our work sheds light on the adversarial robustness evaluation of prompt-based FSL methods for NLU tasks.

NOTABLE: Transferable Backdoor Attacks Against Prompt-based NLP Models
Kai Mei, Zheng Li, Zhenting Wang, Yang Zhang and Shiqing Ma   11:00-12:30 (Pier 7&8)
Prompt-based learning is vulnerable to backdoor attacks. Existing backdoor attacks against prompt-based models consider injecting backdoors into the entire embedding layers or word embedding vectors. Such attacks can be easily affected by retraining on downstream tasks and with different prompting strategies, limiting the transferability of backdoor attacks. In this work, we propose transferable backdoor attacks against prompt-based models, called NOTABLE, which is independent of downstream tasks and prompting strategies. Specifically, NOTABLE injects backdoors into the encoders of PLMs by utilizing an adaptive verbalizer to bind triggers to specific words (i.e., anchors). It activates the backdoor by pasting input with triggers to reach adversary-desired anchors, achieving independence from downstream tasks and prompting strategies. We conduct experiments on six NLP tasks, three popular models, and three prompting strategies. Empirical results show that NOTABLE achieves superior attack performance (i.e., attack success rate over 90% on all the datasets), and outperforms two state-of-the-art baselines. Evaluations on three defenses show the robustness of NOTABLE. Our code can be found at https://github.com/RU-System-Software-and-Security/Notable.

Don't Retrain, Just Rewrite: Countering Adversarial Perturbations by Rewriting Text
Ashim Gupta, Carter Wood Blum, Temma Choji, Yingjie Fei, Shalin Shah, Alakananda Vempala and Vivek Srikumar   11:00-12:30 (Pier 7&8)
Can language models transform inputs to protect text classifiers against adversarial attacks? In this work, we present ATINTER, a model

that intercepts and learns to rewrite adversarial inputs to make them non-adversarial for a downstream text classifier. Our experiments on four datasets and five attack mechanisms reveal that ATINTER is effective at providing better adversarial robustness than existing defense approaches, without compromising task accuracy. For example, on sentiment classification using the SST-2 dataset, our method improves the adversarial accuracy over the best existing defense approach by more than 4% with a smaller decrease in task accuracy (0.5 % vs 2.5%). Moreover, we show that ATINTER generalizes across multiple downstream tasks and classifiers without having to explicitly retrain it for those settings. For example, we find that when ATINTER is trained to remove adversarial perturbations for the sentiment classification task on the SST-2 dataset, it even transfers to a semantically different task of news classification (on AGNews) and improves the adversarial robustness by more than 10%.

Contrastive Learning with Adversarial Examples for Alleviating Pathology of Language Model
Pengwei Zhan, Jing Yang, Xiao Huang, Chunlei Jing, Jingying Li and Liming Wang   11:00-12:30 (Pier 7&8)
Neural language models have achieved superior performance. However, these models also suffer from the pathology of overconfidence in the out-of-distribution examples, potentially making the model difficult to interpret and the interpretation methods fail to provide faithful attributions. In this paper, we explain the model pathology from the view of sentence representation and argue that the counter-intuitive bias degree and direction of the out-of-distribution examples' representation cause the pathology. We propose a Contrastive learning regularization method using Adversarial examples for Alleviating the Pathology (ConAAP), which calibrates the sentence representation of out-of-distribution examples. ConAAP generates positive and negative examples following the attribution results and utilizes adversarial examples to introduce direction information in regularization. Experiments show that ConAAP effectively alleviates the model pathology while slightly impacting the generalization ability on in-distribution examples and thus helps interpretation methods obtain more faithful results.

Claim-Dissector: An Interpretable Fact-Checking System with Joint Re-ranking and Veracity Prediction
Martin Fajcik, Petr Motlicek and Pavel Smrz   11:00-12:30 (Pier 7&8)
We present Claim-Dissector: a novel latent variable model for fact-checking and analysis, which given a claim and a set of retrieved evidence jointly learns to identify: (i) the relevant evidences to the given claim (ii) the veracity of the claim. We propose to disentangle the per-evidence relevance probability and its contribution to the final veracity probability in an interpretable way — the final veracity probability is proportional to a linear ensemble of per-evidence relevance probabilities. In this way, the individual contributions of evidences towards the final predicted probability can be identified. In per-evidence relevance probability, our model can further distinguish whether each relevant evidence is supporting (S) or refuting (R) the claim. This allows to quantify how much the S/R probability contributes to final verdict or to detect disagreeing evidence. Despite its interpretable nature, our system achieves results competetive with state-of-the-art on the FEVER dataset, as compared to typical two-stage system pipelines, while using significantly fewer parameters. Furthermore, our analysis shows that our model can learn fine-grained relevance cues while using coarse-grained supervision and we demonstrate it in 2 ways. (i) We show that our model can achieve competitive sentence recall while using only paragraph-level relevance supervision. (ii) Traversing towards the finest granularity of relevance, we show that our model is capable of identifying relevance at the token level. To do this, we present a new benchmark TLR-FEVER focusing on token-level interpretability — humans annotate tokens in relevant evidences they considered essential when making their judgment. Then we measure how similar are these annotations to the tokens our model is focusing on.

Instruction Induction: From Few Examples to Natural Language Task Descriptions
Or Honovich, Uri Shaham, Samuel R. Bowman and Omer Levy   11:00-12:30 (Pier 7&8)
Large language models are able to perform a task by conditioning on a few input-output demonstrations - a paradigm known as in-context learning. We show that language models can explicitly infer an underlying task from a few demonstrations by prompting them to generate a natural language instruction that fits the examples. To explore this ability, we introduce the instruction induction challenge, compile a dataset consisting of 24 tasks, and define a novel evaluation metric based on executing the generated instruction. We discover that, to a large extent, the ability to generate instructions does indeed emerge when using a model that is both large enough and aligned to follow instructions; InstructGPT achieves 65.7% of human performance in our execution-based metric, while the original GPT-3 model reaches only 9.8% of human performance. This surprising result suggests that instruction induction might be a viable learning paradigm in and of itself, where instead of fitting a set of latent continuous parameters to the data, one searches for the best description in the natural language hypothesis space.

Fine-tuning Happens in Tiny Subspaces: Exploring Intrinsic Task-specific Subspaces of Pre-trained Language Models
Zhong Zhang, Bang Liu and Junming Shao   11:00-12:30 (Pier 7&8)
Pre-trained language models (PLMs) are known to be overly parameterized and have significant redundancy, indicating a small degree of freedom of the PLMs. Motivated by the observation, in this paper, we study the problem of re-parameterizing and fine-tuning PLMs from a new perspective: Discovery of intrinsic task-specific subspace. Specifically, by exploiting the dynamics of the fine-tuning process for a given task, the parameter optimization trajectory is learned to uncover its intrinsic task-specific subspace. A key finding is that PLMs can be effectively fine-tuned in the subspace with a small number of free parameters. Beyond, we observe some outlier dimensions emerging during fine-tuning in the subspace. Disabling these dimensions degrades the model performance significantly. This suggests that these dimensions are crucial to induce task-specific knowledge to downstream tasks.

PromptAttack: Probing Dialogue State Trackers with Adversarial Prompts
Xiangjue Dong, Yun He, Ziwei Zhu and James Caverlee   11:00-12:30 (Pier 7&8)
A key component of modern conversational systems is the Dialogue State Tracker (or DST), which models a user's goals and needs. Toward building more robust and reliable DSTs, we introduce a prompt-based learning approach to automatically generate effective adversarial examples to probe DST models. Two key characteristics of this approach are: (i) it only needs the output of the DST with no need for model parameters, and (ii) it can learn to generate natural language utterances that can target any DST. Through experiments over state-of-the-art DSTs, the proposed framework leads to the greatest reduction in accuracy and the best attack success rate while maintaining good fluency and a low perturbation ratio. We also show how much the generated adversarial examples can bolster a DST through adversarial training. These results indicate the strength of prompt-based attacks on DSTs and leave open avenues for continued refinement.

Reinforcement Learning for Topic Models
Jeremy J Costello and Marek Z Reformat   11:00-12:30 (Pier 7&8)
We apply reinforcement learning techniques to topic modeling by replacing the variational autoencoder in ProdLDA with a continuous action space reinforcement learning policy. We train the system with a policy gradient algorithm REINFORCE. Additionally, we introduced several modifications: modernize the neural network architecture, weight the ELBO loss, use contextual embeddings, and monitor the learning process via computing topic diversity and coherence for each training step. Experiments are performed on 11 data sets. Our unsupervised model outperforms all other unsupervised models and performs on par with or better than most models using supervised labeling. Our model is outperformed on certain data sets by a model using supervised labeling and contrastive learning. We have also conducted an ablation study to provide empirical evidence of performance improvements from changes we made to ProdLDA and found that the reinforcement learning formulation boosts performance. We open-source our code implementation.

Gender-tuning: Empowering Fine-tuning for Debiasing Pre-trained Language Models

Somayeh Ghanbarzadeh, Yan Huang, Hamid Palangi, Radames Saul Cruz Moreno and Hamed Khanpour   11:00-12:30 (Pier 7&8)
Recent studies have revealed that the widely-used Pre-trained Language Models (PLMs) propagate societal biases from the large unmoderated pre-training corpora. Existing solutions require debiasing training processes and datasets for debiasing, which are resource-intensive and costly. Furthermore, these methods hurt the PLMs' performance on downstream tasks. In this study, we propose Gender-tuning, which debiases the PLMs through fine-tuning on downstream tasks' datasets. For this aim, Gender-tuning integrates Masked Language Modeling (MLM) training objectives into fine-tuning's training process. Comprehensive experiments show that Gender-tuning outperforms the state-of-the-art baselines in terms of average gender bias scores in PLMs while improving PLMs' performance on downstream tasks solely using the downstream tasks' dataset. Also, Gender-tuning is a deployable debiasing tool for any PLM that works with original fine-tuning.

FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models
Shramay Palta and Rachel Rudinger   11:00-12:30 (Pier 7&8)
It is common sense that one should prefer to eat a salad with a fork rather than with a chainsaw. However, for eating a bowl of rice, the choice between a fork and a pair of chopsticks is culturally relative. We introduce FORK, a small, manually-curated set of CommonsenseQA-style questions for probing cultural biases and assumptions present in commonsense reasoning systems, with a specific focus on food-related customs. We test several CommonsenseQA systems on FORK, and while we see high performance on questions about the US culture, the poor performance of these systems on questions about non-US cultures highlights systematic cultural assumptions aligned with US over non-US cultures.

Foveate, Attribute, and Rationalize: Towards Physically Safe and Trustworthy AI
Alex Mei, Sharon Levy and William Yang Wang   11:00-12:30 (Pier 7&8)
Users' physical safety is an increasing concern as the market for intelligent systems continues to grow, where unconstrained systems may recommend users dangerous actions that can lead to serious injury. Covertly unsafe text is an area of particular interest, as such text may arise from everyday scenarios and are challenging to detect as harmful. We propose FARM, a novel framework leveraging external knowledge for trustworthy rationale generation in the context of safety. In particular, FARM foveates on missing knowledge to qualify the information required to reason in specific scenarios and retrieves this information with attribution to trustworthy sources. This knowledge is used to both classify the safety of the original text and generate human-interpretable rationales, shedding light on the risk of systems to specific user groups and helping both stakeholders manage the risks of their systems and policymakers to provide concrete safeguards for consumer safety. Our experiments show that FARM obtains state-of-the-art results on the SafeText dataset, showing absolute improvement in safety classification accuracy by 5.9%.

Social-Group-Agnostic Bias Mitigation via the Stereotype Content Model
Ali Omrani, Alireza Salkhordeh Ziabari, Charles Yu, Preni Golazizian, Brendan Kennedy, Mohammad Atari, Heng Ji and Morteza Dehghani
11:00-12:30 (Pier 7&8)
Existing bias mitigation methods require social-group-specific word pairs (e.g., "man" – "woman") for each social attribute (e.g., gender), restricting the bias mitigation to only one specified social attribute. Further, this constraint renders such methods impractical and costly for mitigating bias in understudied and/or unmarked social groups. We propose that the Stereotype Content Model (SCM) — a theoretical framework developed in social psychology for understanding the content of stereotyping — can help debiasing efforts to become social-group-agnostic by capturing the underlying connection between bias and stereotypes. SCM proposes that the content of stereotypes map to two psychological dimensions of warmth and competence. Using only pairs of terms for these two dimensions (e.g., warmth: "genuine" – "fake"; competence: "smart" – "stupid"), we perform debiasing with established methods on both pre-trained word embeddings and large language models. We demonstrate that our social-group-agnostic, SCM-based debiasing technique performs comparably to group-specific debiasing on multiple bias benchmarks, but has theoretical and practical advantages over existing approaches.

Shielded Representations: Protecting Sensitive Attributes Through Iterative Gradient-Based Projection
Shadi Iskander, Kira Radinsky and Yonatan Belinkov   11:00-12:30 (Pier 7&8)
Natural language processing models tend to learn and encode social biases present in the data. One popular approach for addressing such biases is to eliminate encoded information from the model's representations. However, current methods are restricted to removing only linearly encoded information. In this work, we propose Iterative Gradient-Based Projection (IGBP), a novel method for removing non-linear encoded concepts from neural representations. Our method consists of iteratively training neural classifiers to predict a particular attribute we seek to eliminate, followed by a projection of the representation on a hypersurface, such that the classifiers become oblivious to the target attribute. We evaluate the effectiveness of our method on the task of removing gender and race information as sensitive attributes. Our results demonstrate that IGBP is effective in mitigating bias through intrinsic and extrinsic evaluations, with minimal impact on downstream task accuracy.

A Comparative Study on the Impact of Model Compression Techniques on Fairness in Language Models
Krithika Ramesh, Arnav Chavan, Shrey Pandit and Sunayana Sitaram   11:00-12:30 (Pier 7&8)
Compression techniques for deep learning have become increasingly popular, particularly in settings where latency and memory constraints are imposed. Several methods, such as pruning, distillation, and quantization, have been adopted for compressing models, each providing distinct advantages. However, existing literature demonstrates that compressing deep learning models could affect their fairness. Our analysis involves a comprehensive evaluation of pruned, distilled, and quantized language models, which we benchmark across a range of intrinsic and extrinsic metrics for measuring bias in text classification. We also investigate the impact of using multilingual models and evaluation measures. Our findings highlight the significance of considering both the pre-trained model and the chosen compression strategy in developing equitable language technologies. The results also indicate that compression strategies can have an adverse effect on fairness measures.

Causal Intervention for Mitigating Name Bias in Machine Reading Comprehension
Jiazheng Zhu, shaojuan wu, Xiaowang Zhang, Yuexian Hou and Zhiyong Feng   11:00-12:30 (Pier 7&8)
Machine Reading Comprehension (MRC) is to answer questions based on a given passage, which has made great achievements using pre-trained Language Models (LMs). We study the robustness of MRC models to names which is flexible and repeatability. MRC models based on LMs may overuse the name information to make predictions, which causes the representation of names to be non-interchangeable, called name bias. In this paper, we propose a novel Causal Interventional paradigm for MRC (CI4MRC) to mitigate name bias. Specifically, we uncover that the pre-trained knowledge concerning names is indeed a confounder by analyzing the causalities among the pre-trained knowledge, context representation and answers based on a Structural Causal Model (SCM). We develop effective CI4MRC algorithmic implementations to constrain the confounder based on the neuron-wise and token-wise adjustments. Experiments demonstrate that our proposed CI4MRC effectively mitigates the name bias and achieves competitive performance on the original SQuAD. Moreover, our method is general to various pre-trained LMs and performs robustly on the adversarial datasets.

T2IAT: Measuring Valence and Stereotypical Biases in Text-to-Image Generation
Jialu Wang, Xinyue Gabby Liu, Zonglin Di, Yang Liu and Xin Eric Wang   11:00-12:30 (Pier 7&8)
Warning: This paper contains several contents that may be toxic, harmful, or offensive.

In the last few years, text-to-image generative models have gained remarkable success in generating images with unprecedented quality accompanied by a breakthrough of inference speed. Despite their rapid progress, human biases that manifest in the training examples, particularly with regard to common stereotypical biases, like gender and skin tone, still have been found in these generative models. In this work, we seek to measure more complex human biases exist in the task of text-to-image generations. Inspired by the well-known Implicit Association Test (IAT) from social psychology, we propose a novel Text-to-Image Association Test (T2IAT) framework that quantifies the implicit stereotypes between concepts and valence, and those in the images. We replicate the previously documented bias tests on generative models, including morally neutral tests on flowers and insects as well as demographic stereotypical tests on diverse social attributes. The results of these experiments demonstrate the presence of complex stereotypical behaviors in image generations.

DP-BART for Privatized Text Rewriting under Local Differential Privacy
Timour Igamberdiev and Ivan Habernal   11:00-12:30 (Pier 7&8)
Privatized text rewriting with local differential privacy (LDP) is a recent approach that enables sharing of sensitive textual documents while formally guaranteeing privacy protection to individuals. However, existing systems face several issues, such as formal mathematical flaws, unrealistic privacy guarantees, privatization of only individual words, as well as a lack of transparency and reproducibility. In this paper, we propose a new system 'DP-BART' that largely outperforms existing LDP systems. Our approach uses a novel clipping method, iterative pruning, and further training of internal representations which drastically reduces the amount of noise required for DP guarantees. We run experiments on five textual datasets of varying sizes, rewriting them at different privacy guarantees and evaluating the rewritten texts on downstream text classification tasks. Finally, we thoroughly discuss the privatized text rewriting approach and its limitations, including the problem of the strict text adjacency constraint in the LDP paradigm that leads to the high noise requirement.

Modeling the Q-Diversity in a Min-max Play Game for Robust Optimization
Ting Wu, Rui Zheng, Tao Gui, Qi Zhang and Xuanjing Huang   11:00-12:30 (Pier 7&8)
Models trained with empirical risk minimization (ERM) are revealed to easily rely on spurious correlations, resulting in poor generalization. Group distributionally robust optimization (group DRO) can alleviate this problem by minimizing the worst-case loss over pre-defined groups. While promising, in practice factors like expensive annotations and privacy preclude the availability of group labels. More crucially, when taking a closer look at the failure modes of out-of-distribution generalization, the typical procedure of reweighting in group DRO loses efficiency. Hinged on the limitations, in this work, we reformulate the group DRO framework by proposing Q-Diversity. Characterized by an interactive training mode, Q-Diversity relaxes the group identification from annotation into direct parameterization. Furthermore, a novel mixing strategy across groups is presented to diversify the under-represented groups. In a series of experiments on both synthetic and real-world text classification tasks, results demonstrate that Q-Diversity can consistently improve worst-case accuracy under different distributional shifts, outperforming state-of-the-art alternatives.

Modular and On-demand Bias Mitigation with Attribute-Removal Subnetworks
Lukas Hauzenberger, Shahed Masoudian, Deepak Kumar, Markus Schedl and Navid Rekabsaz   11:00-12:30 (Pier 7&8)
Societal biases are reflected in large pre-trained language models and their fine-tuned versions on downstream tasks. Common in-processing bias mitigation approaches, such as adversarial training and mutual information removal, introduce additional optimization criteria, and update the model to reach a new debiased state. However, in practice, end-users and practitioners might prefer to switch back to the original model, or apply debiasing only on a specific subset of protected attributes. To enable this, we propose a novel modular bias mitigation approach, consisting of stand-alone highly sparse debiasing subnetworks, where each debiasing module can be integrated into the core model on-demand at inference time. Our approach draws from the concept of diff pruning, and proposes a novel training regime adaptable to various representation disentanglement optimizations. We conduct experiments on three classification tasks with gender, race, and age as protected attributes. The results show that our modular approach, while maintaining task performance, improves (or at least remains on-par with) the effectiveness of bias mitigation in comparison with baseline finetuning. Particularly on a two-attribute dataset, our approach with separately learned debiasing subnetworks shows effective utilization of either or both the subnetworks for selective bias mitigation.

Debiasing should be Good and Bad: Measuring the Consistency of Debiasing Techniques in Language Models
Robert A. Morabito, Jad Kabbara and Ali Emami   11:00-12:30 (Pier 7&8)
Debiasing methods that seek to mitigate the tendency of Language Models (LMs) to occasionally output toxic or inappropriate text have recently gained traction. In this paper, we propose a standardized protocol which distinguishes methods that yield not only desirable results, but are also consistent with their mechanisms and specifications. For example, we ask, given a debiasing method that is developed to reduce toxicity in LMs, if the definition of toxicity used by the debiasing method is reversed, would the debiasing results also be reversed? We used such considerations to devise three criteria for our new protocol: Specification Polarity, Specification Importance, and Domain Transferability. As a case study, we apply our protocol to a popular debiasing method, Self-Debiasing, and compare it to one we propose, called Instructive Debiasing, and demonstrate that consistency is as important an aspect to debiasing viability as is simply a desirable result. We show that our protocol provides essential insights into the generalizability and interpretability of debiasing methods that may otherwise go overlooked.

Stereotypes and Smut: The (Mis)representation of Non-cisgender Identities by Text-to-Image Models
Eddie L. Ungless, Bjorn Ross and Anne Lauscher   11:00-12:30 (Pier 7&8)
Cutting-edge image generation has been praised for producing high-quality images, suggesting a ubiquitous future in a variety of applications. However, initial studies have pointed to the potential for harm due to predictive bias, reflecting and potentially reinforcing cultural stereotypes. In this work, we are the first to investigate how multimodal models handle diverse gender identities. Concretely, we conduct a thorough analysis in which we compare the output of three image generation models for prompts containing cisgender vs. non-cisgender identity terms. Our findings demonstrate that certain non-cisgender identities are consistently (mis)represented as less human, more stereotyped and more sexualised. We complement our experimental analysis with (a) a survey among non-cisgender individuals and (b) a series of interviews, to establish which harms affected individuals anticipate, and how they would like to be represented. We find respondents are particularly concerned about misrepresentation, and the potential to drive harmful behaviours and beliefs. Simple heuristics to limit offensive content are widely rejected, and instead respondents call for community involvement, curated training data and the ability to customise. These improvements could pave the way for a future where change is led by the affected community, and technology is used to positively "[portray] queerness in ways that we haven't even thought of'" rather than reproducing stale, offensive stereotypes.

The Devil is in the Details: On the Pitfalls of Event Extraction Evaluation
Hao Peng, Xiaozhi Wang, Feng Yao, Kaisheng Zeng, Lei Hou, Juanzi Li, Zhiyuan Liu and Weixing Shen   11:00-12:30 (Pier 7&8)
Event extraction (EE) is a crucial task aiming at extracting events from texts, which includes two subtasks: event detection (ED) and event argument extraction (EAE). In this paper, we check the reliability of EE evaluations and identify three major pitfalls: (1) The data preprocessing discrepancy makes the evaluation results on the same dataset not directly comparable, but the data preprocessing details are not widely noted and specified in papers. (2) The output discrepancy of different model paradigms makes different-paradigm EE models lack grounds for comparison and also leads to unclear mapping issues between predictions and annotations. (3) The absence of pipeline evaluation of many EAE-only works makes them hard to be directly compared with EE works and may not well reflect the model performance in real-world pipeline scenarios. We demonstrate the significant influence of these pitfalls through comprehensive meta-analyses of recent papers and em-

pirical experiments. To avoid these pitfalls, we suggest a series of remedies, including specifying data preprocessing, standardizing outputs, and providing pipeline evaluation results. To help implement these remedies, we develop a consistent evaluation framework OmniEvent, which can be obtained from https://github.com/THU-KEG/OmniEvent.

Is Anisotropy Truly Harmful? A Case Study on Text Clustering
Mira Ait-Saada and Mohamed Nadif   11:00-12:30 (Pier 7&8)
In the last few years, several studies have been devoted to dissecting dense text representations in order to understand their effectiveness and further improve their quality. Particularly, the anisotropy of such representations has been observed, which means that the directions of the word vectors are not evenly distributed across the space but rather concentrated in a narrow cone. This has led to several attempts to counteract this phenomenon both on static and contextualized text representations. However, despite this effort, there is no established relationship between anisotropy and performance. In this paper, we aim to bridge this gap by investigating the impact of different transformations on both the isotropy and the performance in order to assess the true impact of anisotropy. To this end, we rely on the clustering task as a means of evaluating the ability of text representations to produce meaningful groups. Thereby, we empirically show a limited impact of anisotropy on the expressiveness of sentence representations both in terms of directions and L2 closeness.

A Call for Standardization and Validation of Text Style Transfer Evaluation
Phil Sidney Ostheimer, Mayank Kumar Nagda, Marius Kloft and Sophie Fellenz   11:00-12:30 (Pier 7&8)
Text Style Transfer (TST) evaluation is, in practice, inconsistent. Therefore, we conduct a meta-analysis on human and automated TST evaluation and experimentation that thoroughly examines existing literature in the field. The meta-analysis reveals a substantial standardization gap in human and automated evaluation. In addition, we also find a validation gap: only few automated metrics have been validated using human experiments. To this end, we thoroughly scrutinize both the standardization and validation gap and reveal the resulting pitfalls. This work also paves the way to close the standardization and validation gap in TST evaluation by calling out requirements to be met by future research.

This prompt is measuring <mask>: evaluating bias evaluation in language models
Seraphina Goldfarb-Tarrant, Eddie L. Ungless, Esma Balkir and Su Lin Blodgett   11:00-12:30 (Pier 7&8)
Bias research in NLP seeks to analyse models for social biases, thus helping NLP practitioners uncover, measure, and mitigate social harms. We analyse the body of work that uses prompts and templates to assess bias in language models. We draw on a measurement modelling framework to create a taxonomy of attributes that capture what a bias test aims to measure and how that measurement is carried out. By applying this taxonomy to 90 bias tests, we illustrate qualitatively and quantitatively that core aspects of bias test conceptualisations and operationalisations are frequently unstated or ambiguous, carry implicit assumptions, or be mismatched. Our analysis illuminates the scope of possible bias types the field is able to measure, and reveals types that are as yet under-researched. We offer guidance to enable the community to explore a wider section of the possible bias space, and to better close the gap between desired outcomes and experimental design, both for bias and for evaluating language models more broadly.

Numeric Magnitude Comparison Effects in Large Language Models
Raj Sanjay Shah, Vijay Marupudi, Reba Koenen, Khushi Bhardwaj and Sashank Varma   11:00-12:30 (Pier 7&8)
Large Language Models (LLMs) do not differentially represent numbers, which are pervasive in text. In contrast, neuroscience research has identified distinct neural representations for numbers and words. In this work, we investigate how well popular LLMs capture the magnitudes of numbers (e.g., that 4<5) from a behavioral lens. Prior research on the representational capabilities of LLMs evaluates whether they show human-level performance, for instance, high overall accuracy on standard benchmarks. Here, we ask a different question, one inspired by cognitive science: How closely do the number representations of LLMscorrespond to those of human language users, who typically demonstrate the distance, size, and ratio effects? We depend on a linking hypothesis to map the similarities among the model embeddings of number words and digits to human response times. The results reveal surprisingly human-like representations across language models of different architectures, despite the absence of the neural circuitry that directly supports these representations in the human brain. This research shows the utility of understanding LLMs using behavioral benchmarks and points the way to future work on the number of representations of LLMs and their cognitive plausibility.

Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation
Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow and Yanai Elazar   11:00-12:30 (Pier 7&8)
Few-shot fine-tuning and in-context learning are two alternative strategies for task adaptation of pre-trained language models. Recently, in-context learning has gained popularity over fine-tuning due to its simplicity and improved out-of-domain generalization, and because extensive evidence shows that fine-tuned models pick up on spurious correlations. Unfortunately, previous comparisons of the two approaches were done using models of different sizes. This raises the question of whether the observed weaker out-of-domain generalization of fine-tuned models is an inherent property of fine-tuning or a limitation of the experimental setup. In this paper, we compare the generalization of few-shot fine-tuning and in-context learning to challenge datasets, while controlling for the models used, the number of examples, and the number of parameters, ranging from 125M to 30B. Our results show that fine-tuned language models can in fact generalize well out-of-domain. We find that both approaches generalize similarly; they exhibit large variation and depend on properties such as model size and the number of examples, highlighting that robust task adaptation remains a challenge.

GUMSum: Multi-Genre Data and Evaluation for English Abstractive Summarization
Yang Janet Liu and Amir Zeldes   11:00-12:30 (Pier 7&8)
Automatic summarization with pre-trained language models has led to impressively fluent results, but is prone to 'hallucinations', low performance on non-news genres, and outputs which are not exactly summaries. Targeting ACL 2023's 'Reality Check' theme, we present GUMSum, a small but carefully crafted dataset of English summaries in 12 written and spoken genres for evaluation of abstractive summarization. Summaries are highly constrained, focusing on substitutive potential, factuality, and faithfulness. We present guidelines and evaluate human agreement as well as subjective judgments on recent system outputs, comparing general-domain untuned approaches, a fine-tuned one, and a prompt-based approach, to human performance. Results show that while GPT3 achieves impressive scores, it still underperforms humans, with varying quality across genres. Human judgments reveal different types of errors in supervised, prompted, and human-generated summaries, shedding light on the challenges of producing a good summary.

Towards Benchmarking and Improving the Temporal Reasoning Capability of Large Language Models
Qingyu Tan, Hwee Tou Ng and Lidong Bing   11:00-12:30 (Pier 7&8)
Reasoning about time is of fundamental importance. Many facts are time-dependent. For example, athletes change teams from time to time, and different government officials are elected periodically. Previous time-dependent question answering (QA) datasets tend to be biased in either their coverage of time spans or question types. In this paper, we introduce a comprehensive probing dataset TempReason to evaluate the temporal reasoning capability of large language models. Our dataset includes questions of three temporal reasoning levels. In addition, we also propose a novel learning framework to improve the temporal reasoning capability of large language models, based on temporal span extraction and time-sensitive reinforcement learning. We conducted experiments in closed book QA, open book QA, and reasoning QA set-

tings and demonstrated the effectiveness of our approach.

Controlling Learned Effects to Reduce Spurious Correlations in Text Classifiers
Parikshit Bansal and Amit Sharma   11:00-12:30 (Pier 7&8)
To address the problem of NLP classifiers learning spurious correlations between training features and target labels, a common approach is to make the model's predictions invariant to these features. However, this can be counter-productive when the features have a non-zero causal effect on the target label and thus are important for prediction. Therefore, using methods from the causal inference literature, we propose an algorithm to regularize the learnt effect of the features on the model's prediction to the estimated effect of feature on label. This results in an automated augmentation method that leverages the estimated effect of a feature to appropriately change the labels for new augmented inputs. On toxicity and IMDB review datasets, the proposed algorithm minimises spurious correlations and improves the minority group (i.e., samples breaking spurious correlations) accuracy, while also improving the total accuracy compared to standard training.

How do humans perceive adversarial text? A reality check on the validity and naturalness of word-based adversarial attacks
Salijona Dyrmishi, Salah GHAMIZI and Maxime Cordy   11:00-12:30 (Pier 7&8)
Natural Language Processing (NLP) models based on Machine Learning (ML) are susceptible to adversarial attacks – malicious algorithms that imperceptibly modify input text to force models into making incorrect predictions. However, evaluations of these attacks ignore the property of imperceptibility or study it under limited settings. This entails that adversarial perturbations would not pass any human quality gate and do not represent real threats to human-checked NLP systems. To bypass this limitation and enable proper assessment (and later, improvement) of NLP model robustness, we have surveyed 378 human participants about the perceptibility of text adversarial examples produced by state-of-the-art methods. Our results underline that existing text attacks are impractical in real-world scenarios where humans are involved. This contrasts with previous smaller-scale human studies, which reported overly optimistic conclusions regarding attack success. Through our work, we hope to position human perceptibility as a first-class success criterion for text attacks, and provide guidance for research to build effective attack algorithms and, in turn, design appropriate defence mechanisms.

Moving Beyond Downstream Task Accuracy for Information Retrieval Benchmarking
Keshav Santhanam, Jon Saad-Falcon, Martin Franz, Omar Khattab, Avi Sil, Radu Florian, Md Arafat Sultan, Salim Roukos, Matei Zaharia and Christopher Potts   11:00-12:30 (Pier 7&8)
Neural information retrieval (IR) systems have progressed rapidly in recent years, in large part due to the release of publicly available benchmarking tasks. Unfortunately, some dimensions of this progress are illusory: the majority of the popular IR benchmarks today focus exclusively on downstream task accuracy and thus conceal the costs incurred by systems that trade away efficiency for quality. Latency, hardware cost, and other efficiency considerations are paramount to the deployment of IR systems in user-facing settings. We propose that IR benchmarks structure their evaluation methodology to include not only metrics of accuracy, but also efficiency considerations such as a query latency and the corresponding cost budget for a reproducible hardware setting. For the popular IR benchmarks MS MARCO and XOR-TyDi, we show how the best choice of IR system varies according to how these efficiency considerations are chosen and weighed. We hope that future benchmarks will adopt these guidelines toward more holistic IR evaluation.

9

Workshops

Overview

During the days of the workshops, **Registration** will be held from 08:00.

Thursday, July 13, 2023

Friday, July 14, 2023

W1 - The 17th International Workshop on Semantic Evaluation (SemEval)

Organizers:
Ritesh Kumar, Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish
Tayyar Madabushi

https://semeval.github.io/SemEval2023/
Venue: Queens Quay
Thursday, July 13, 2023

The 17th edition of SemEval features 12 TASKS on a range of topics, including tasks on idiomaticy detection and embedding, sarcasm detection, multilingual news similarity, and linking mathematical symbols to their descriptions. Several tasks are multilingual, and others ask for multimodal approaches.

W2 - The 12th Joint Conference on Lexical and Computational Semantics (*SEM)

Organizers:

Mohammad Taher Pilehvar, Jose Camacho-Collados, Alexis Palmer, Malihe Alikhani, Mert Inan

`https://sites.google.com/view/starsem2023`
Venue: Pier 5
Thursday, July 13, 2023

The 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023) is organized and sponsored by SIGLEX, the Special Interest Group of the ACL. *SEM brings together researchers interested in the semantics of natural languages and its computational modeling. The conference embraces data-driven, neural, and probabilistic approaches, as well as symbolic approaches and everything in between; practical applications as well as theoretical contributions are welcome. The long-term goal of *SEM is to provide a stable forum for the growing number of NLP researchers working on all aspects of semantics of (many and diverse!) natural languages.

W3 - The 4th Workshop on Computational Approaches to Discourse (CODI)

Organizers:

Chloé Braud, Christian Hardmeier, Junyi Jessy Li, Sharid Loáiciga, Michael Strube, Amir Zeldes

The last ten years have seen a dramatic improvement in the ability of NLP systems to understand and produce words and sentences. This development has created a renewed interest in discourse phenomena as researchers move towards the processing of long-form text and conversations. There is a surge of activity in discourse parsing, coherence models, text summarization, corpora for discourse level reading comprehension, and discourse related/aided representation learning, to name a few, but the problems in computational approaches to discourse are still substantial. At this juncture, we have organized three Workshops on Computational Approaches to Discourse (CODI) at EMNLP 2020, EMNLP 2021 and COLING 2022 to bring together discourse experts and upcoming researchers. These workshops have catalyzed work to improve the speed and knowledge needed to solve such problems and have served as a forum for the discussion of suitable datasets and reliable evaluation methods.

W4 - The 20th International Conference on Spoken Language Translation (IWSLT)

Organizers:

Marine Carpuat, Marcello Federico, Alex Waibel, Jan Niehues, Sebastian Stüker,
Elizabeth Salesky, Atul Kr. Ojha

https://iwslt.org/2023/
Venue: Dockside 1
Thursday, July 13, 2023

The International Conference on Spoken Language Translation (IWSLT) is an annual scientific conference, associated with an open evaluation campaign on spoken language translation, where both scientific papers and system descriptions are presented.

W5 - The 8th Workshop on Representation Learning for NLP (RepL4NLP)

Organizers:

Burcu Can, Maximilian Mozes, Samuel Cahyawijaya, Naomi Saphra, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Chen Zhao

https://sites.google.com/view/repl4nlp2023
Venue: Harbour B
Thursday, July 13, 2023

The 8th Workshop on Representation Learning for NLP aims to continue the success of the Repl4NLP workshop series, with the 1st Workshop on Representation Learning for NLP having received about 50 submissions and over 250 attendees - the second most attended collocated event at ACL'16 after WMT. The workshop was introduced as a synthesis of several years of independent *CL workshops focusing on vector space models of meaning, compositionality, and the application of deep neural networks and spectral methods to NLP. It provides a forum for discussing recent advances on these topics, as well as future research directions in linguistically motivated vector-based models in NLP. The workshop will take place in a hybrid setting, and, as in previous years, feature interdisciplinary keynotes, paper presentations, posters, as well as a panel discussion.

W6 - The 4th Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)

Organizers:

Nafise Sadat Moosavi, Iryna Gurevych, Yufang Hou, Gyuwan Kim, Young Jin Kim, Tal Schuster, Ameeta Agrawal

The Natural Language Processing (NLP) community has, in recent years, demonstrated a notable focus on improving higher scores on standard benchmarks and taking the lead on community-wide leaderboards (e.g., GLUE, SentEval). While this aspiration has led to improvements in benchmark performance of (predominantly neural) models, it has also came at a cost, i.e., increased model complexity and the ever-growing amount of computational resources required for training and using the current state-of-the-art models. Moreover, the recent research efforts have, for the most part, failed to identify sources of empirical gains in models, often failing to empirically justify the model complexity beyond benchmark performance.

Because of these easily observable trends, we have proposed the SustaiNLP workshop with the goal of promoting more sustainable NLP research and practices, with two main objectives: (1) encouraging development of more efficient NLP models; and (2) providing simpler architectures and empirical justification of model complexity. For both aspects, we will encourage submissions from all topical areas of NLP.

W7 - The 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)

Organizers:
Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, Torsten Zesch

The BEA Workshop is a leading venue for NLP innovation in the context of educational applications. It is one of the largest one-day workshops in the ACL community with over 100 registered attendees in the past several years. The growing interest in educational applications and a diverse community of researchers involved resulted in the creation of the Special Interest Group in Educational Applications (SIGEDU) in 2017, which currently has over 300 members.

W8 - The 1st Workshop on Natural Language Reasoning and Structured Explanations

Organizers:

Peter Clark, Ellie Pavlick, Denny Zhou, Noah Goodman, Sarah Wiegreffe, Felix Hill

With recent scaling of large pre-trained Transformer language models (LLMs), the scope of feasible NLP tasks has broadened. Significant recent work has focused on tasks that require some kind of natural language reasoning. A trajectory in question answering has led us from extraction-oriented datasets like SQuAD to "multi-hop" reasoning datasets like HotpotQA and StrategyQA. Although LLMs have shown remarkable performance on most NLP tasks, it is often unclear why their answers follow from what they know. To address this gap, a new class of explanation techniques has emerged which play an integral part in structuring the reasoning necessary to solve these datasets. For example, the chain-of-thought paradigm leverages explanations as vehicles for LLMs to mimic human reasoning processes. Entailment trees offer a way to ground multi-step reasoning in a collection of verifiable steps. Frameworks like SayCan bridge high-level planning in language and with low-level action trajectories. As a result, we see a confluence of methods blending explainable machine learning/NLP, classical AI (especially theorem proving), and cognitive science (how do humans structure explanations?). This workshop aims to bring together a diverse set of perspectives from these different traditions and attempt to establish common ground for how these various kinds of explanation structures can tackle a broad class of reasoning problems in natural language and beyond.

W9 - The 7th Workshop on Online Abuse and Harms (WOAH)

Organizers:
Yi-Ling Chung, Aida Mostafazadeh Davani, Debora Nozza, Paul Röttger, Zeerak
Talat

The goal of The Workshop on Online Abuse and Harms (WOAH) is to advance research that develops, interrogates and applies computational methods for detecting, classifying and modelling online abuse.

W10 - The 3rd Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc)

Organizers:

Roee Aharoni, Nouha Dziri, Song Feng, Yongbin Li, Yu Li, Hui Wan

https://doc2dial.github.io/workshop2023/
Venue: Dockside 2
Thursday, July 13, 2023

The DialDoc workshop focuses on Document-Grounded Dialogue and Conversational Question Answering. Given the vast amount of content created every day in various mediums, it is a meaningful yet challenging task not only to make such content accessible to end users via various conversational interfaces, but also to make sure the responses provided by the models are grounded and faithful with respect to the knowledge sources.

W11 - The 1st Workshop on Matching From Unstructured and Structured Data (MATCHING)

Organizers:
Dunia Mladenić, Estevam Hruschka, Marko Grobelnik, Sajjadur Rahman, Tom Mitchell

Matching Entities from structured and unstructured sources is an important task in many domains and applications such as HR and E-commerce. For example, in HR platforms/services, it is important to match resumes to job descriptions and job seekers to companies. Similarly in web platforms/services, it is important to match customers to businesses such as hotels and restaurant, among others. In such domains, it is also relevant to match "textual customer reviews" to customers queries, and sentences (or phrases) as answers to customer questions. Recent advances in Natural Language Processing, Natural Language Understanding, Conversational AI, Language Generation, Machine Learning, Deep Learning, Data Management, Information Extraction, Knowledge Bases/Graphs, (MultiSingle Hop/Commonsense) Inference/Reasoning, Recommendation Systems, and others, have demonstrated promising results in different Matching tasks related (but not limited) to the previously mentioned domains. We believe that there is tremendous opportunity to further exploit and explore the use of advanced NLP (and language related) techniques applied to Matching tasks. Therefore, the goal of this workshop is to bring together the research communities (from academia and industry) of these related areas, that are interested in the development and the application of novel natural-language-based approaches/models/systems to address challenges around different Matching tasks.

W12 - The 17th Workshop on Linguistic Annotation (LAW)

Organizers:
Annemarie Friedrich, Jakob Prange, Amir Zeldes, Ines Rehbein

Linguistic annotation of natural language corpora is the backbone of supervised methods of statistical natural language processing. The Linguistic Annotation Workshop (LAW) is the annual workshop of the ACL Special Interest Group on Annotation (SIGANN), and it provides a forum for the presentation and discussion of innovative research on all aspects of linguistic annotation, including the creation and evaluation of annotation schemes, methods for automatic and manual annotation, use and evaluation of annotation software and frameworks, representation of linguistic data and annotations, semi-supervised "human in the loop" methods of annotation, crowd-sourcing approaches, and more. As in the past, the LAW will provide a forum for annotation researchers to work towards standardization, best practices, and interoperability of annotation information and software.

W13 - The 22nd Workshop on Biomedical Natural Language Processing and Shared Tasks (BioNLP-ST)

Organizers:

Kevin Bretonnel Cohen, Dina Demner-Fushman, Sophia Ananiadou, Jun-ichi Tsujii

The BioNLP workshop associated with the ACL SIGBIOMED special interest group has established itself as the primary venue for presenting foundational research in language processing for the biological and medical domains. The workshop is running every year since 2002 and continues getting stronger. BioNLP welcomes and encourages work on languages other than English, and inclusion and diversity. BioNLP truly encompasses the breadth of the domain and brings together researchers in bio- and clinical NLP from all over the world. The workshop will continue presenting work on a broad and interesting range of topics in NLP. The interest to biomedical language has broadened significantly due to the COVID-19 pandemic and continues to grow: as access to information becomes easier and more people generate and access health-related text, it becomes clearer that only language technologies can enable and support adequate use of the biomedical text.

W14 - The 5th Workshop on NLP for Conversational AI

Organizers:
Abhinav Rastogi, Georgios Spithourakis, Yun-Nung (Vivian) Chen, Bing Liu, Yu Li,
Elnaz Nouri, Alon Albalak, Alexandros Papangelis

Over the past decades, mathematicians, linguists, and computer scientists have dedicated their efforts towards empowering human-machine communication in natural language. While in recent years the emergence of virtual personal assistants such as Siri, Alexa, Google Assistant, Cortana, and ChatGPT has pushed the field forward, they may still have numerous challenges.

Following the success of the 4th NLP for Conversational AI workshop at ACL, The 5th NLP4ConvAI will be a one-day workshop, co-located with ACL 2023 in Toronto, Canada. The goal of this workshop is to bring together researchers and practitioners to discuss impactful research problems in this area, share findings from real-world applications, and generate ideas for future research directions.

The workshop will include keynotes, posters, panel sessions, and a shared task. In keynote talks, senior technical leaders from industry and academia will share insights on the latest developments in the field. We would like to encourage researchers and students to share their prospects and latest discoveries. There will also be a panel discussion with noted conversational AI leaders focused on the state of the field, future directions, and open problems across academia and industry.

W15 - The 3rd Workshop on Trustworthy NLP (TrustNLP)

Organizers:

Yada Pruksachatkun, Ninareh Mehrabi, Kai-Wei Chang, Aram Galystan, Jwala Dhamala, Anaelia Ovalle, Apurv Verma, Yang Trista Cao, Anoop Kumar, Rahul Gupta

Recent advances in Natural Language Processing, and the emergence of pretrained Large Language Models (LLM) specifically, have made NLP systems omnipresent in various aspects of our everyday life. In addition to traditional examples such as personal voice assistants, recommender systems, etc, more recent developments include content-generation models such as ChatGPT, text-to-image models (Dall-E), and so on. While these emergent technologies have an unquestionable potential to power various innovative NLP and AI applications, they also pose a number of challenges in terms of their safe and ethical use. To address such challenges, NLP researchers have formulated various objectives, e.g., intended to make models more fair, safe, and privacy-preserving. However, these objectives are often considered separately, which is a major limitation since it is often important to understand the interplay and/or tension between them. For instance, meeting a fairness objective might require access to users' demographic information, which creates tension with privacy objectives. The goal of this workshop is to move toward a more comprehensive notion of Trustworthy NLP, by bringing together researchers working on those distinct yet related topics, as well as their intersection.

W16 - The 13th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA)

Organizers:
Jeremy Barnes, Orphée De Clercq, Roman Klinger, Valentin Barriere, Salvatore Giorgi, Joaõ Sedoc, Shabnam Tafreshi, Iqra Ameer, Necva Bölücü, Hua Xu, Ali Al Bataineh

https://wassa-workshop.github.io/
Venue: Hourbour C
Friday, July 14, 2023

Subjectivity and Sentiment Analysis has become a highly developed research area, ranging from binary classification of reviews to the detection of complex emotion structures between entities found in text. This field has expanded both on a practical level, finding numerous successful applications in business, as well as on a theoretical level, allowing researchers to explore more complex research questions related to affective computing. Its continuing importance is also shown by the interest it generates in other disciplines such as Economics, Sociology, Psychology, Marketing, Crisis Management & Digital Humanities.
newline The aim of WASSA 2023 is to bring together researchers working on Subjectivity, Sentiment Analysis, Emotion Detection and Classification and their applications to other NLP or real-world tasks (e.g. public health messaging, fake news, media impact analysis, social media mining, computational literary studies) and researchers working on interdisciplinary aspects of affect computation from text.

W17 - The 5th Clinical Natural Language Processing Workshop (Clinical NLP)

Organizers:

Asma Ben Abacha, Steven Bethard, Tristan Naumann, Kirk Roberts, Anna Rumshisky

https://clinical-nlp.github.io/2023/
Venue: Pier 7 and 8
Friday, July 14, 2023

Clinical text is growing rapidly as electronic health records become pervasive. Much of the information recorded in a clinical encounter is located exclusively in provider narrative notes, which makes them indispensable for supplementing structured clinical data in order to better understand patient state and care provided. The methods and tools developed for the clinical domain have historically lagged behind the scientific advances in the general-domain NLP. Despite the substantial recent strides in clinical NLP, a substantial gap remains. The goal of this workshop is to address this gap by establishing a regular event in CL conferences that brings together researchers interested in developing state-of-the-art methods for the clinical domain. The focus is on improving NLP technology to enable clinical applications, and specifically, information extraction and modeling of narrative provider notes from electronic health records, patient encounter transcripts, and other clinical narratives.

W18 - The 1st Workshop on Social Influence in Conversations (SICon)

Organizers:
Kushal Chawla, Weiyan Shi, Maximillian Chen, Liang Qiu, Yu Li, James Hale, Alexandros Papangelis, Gale Lucas, Zhou Yu

https://sites.google.com/view/sicon-2023/home
Venue: Harbour A
Friday, July 14, 2023

Social influence is the change in an individual's thoughts, feelings, attitudes, or behaviors that results from interaction with another individual or a group. For example, a buyer uses social influence skills to engage in trade-offs and build rapport when bargaining with a seller. A therapist uses social influence skills like persuasion to motivate a patient towards physical exercise. Social influence is a core function of human communication, and such scenarios are ubiquitous in everyday life, from negotiations to argumentation to behavioral interventions. Consequently, realistic human-machine conversations must reflect these social influence dynamics, making it essential to systematically model and understand them in dialogue research. This requires perspectives not only from NLP and AI research but also from game theory, emotion, communication, and psychology.

newline We are excited to host the First Workshop on Social Influence in Conversations (SICon 2023). SICon 2023 will be a one-day hybrid event, co-located with ACL 2023. It would be the first venue that uniquely fosters a dedicated discussion on social influence within NLP while involving researchers from other disciplines such as affective computing and the social sciences. SICon 2023 features keynote talks, panel discussions, poster sessions, and lightning talks for accepted papers. We hope to bring together researchers and practitioners from a wide variety of disciplines to discuss important problems related to social influence, as well as share findings and recent advances. We encourage researchers of all stages and backgrounds to share their exciting work!

W19 - The 1st Workshop on Computation and Written Language (CAWL)

Organizers:
Kyle Gorman, Brian Roark, Richard Sproat

`https://cawl.wellformedness.com/`
Venue: Pier 2
Friday, July 14, 2023

Most work on NLP focuses on language in its canonical written form. This has often led researchers to ignore the differences between written and spoken language or, worse, to conflate the two. Instances of conflation are statements like "Chinese is a logographic language" or "Persian is a right-to-left language", variants of which can be found frequently in the ACL anthology. These statements confuse properties of the language with properties of its writing system. Ignoring differences between written and spoken language leads, among other things, to conflating different words that are spelled the same (e.g., English bass), or treating as different, words that have multiple spellings.

newline text enFurthermore, methods for dealing with written language issues (e.g., various kinds of normalization or conversion) or for recognizing text input (e.g. OCR & handwriting recognition or text entry methods) are often regarded as precursors to NLP rather than as fundamental parts of the enterprise, despite the fact that most NLP methods rely centrally on representations derived from text rather than (spoken) language. This general lack of consideration of writing has led to much of the research on such topics to largely appear outside of ACL venues, in conferences or journals of neighboring fields such as speech technology (e.g., text normalization) or human-computer interaction (e.g., text entry).

newline We will invite submissions on the relationship between written and spoken language, the properties of written language, the ways in which writing systems encode language, and applications specifically focused on characteristics of writing systems.

W20 - The 3rd Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)

Organizers:

Manuel Mager, Arturo Oncevay, Enora Rice, Abteen Ebrahimi, Shruti Rijhwani,
Alexis Palmer, Katharina Kann

AmericasNLP aims to (a) encourage research on NLP, computational linguistics, corpus linguistics, and speech around the globe to work on native American languages; (b))connect researchers and professionals from underrepresented communities and native speakers of endangered languages with the machine learning and natural language processing communities; and (c))promote research on both neural and non-neural machine learning approaches suitable for low-resource languages.

W21 - The 5th Workshop on Narrative Understanding (WNU)

Organizers:

Nader Akoury, Faeze Brahman, Khyathi Chandu, Snigdha Chaturvedi, Elizabeth Clark, Mohit Iyyer

`https://sites.google.com/umass.edu/wnu2023`
Venue: Dockside 2
Friday, July 14, 2023

This is the 5th iteration of the Narrative Understanding Workshop, which brings together an interdisciplinary group of researchers from AI, ML, NLP, Computer Vision and other related fields, as well as scholars from the humanities to discuss methods to improve automatic narrative understanding capabilities. The workshop will consist of talks from invited speakers, a panel of researchers and writers, and talks and posters from accepted papers.

W22 - The 20th Workshop on Computational Morphology and Phonology (SIGMORPHON)

Organizers:
Garrett Nicolai, Eleanor Chodroff, Çağrı Çöltekin, Fred Mailhot

`https://sigmorphon.github.io/workshops/2023/`
Venue: Dockside 3
Friday, July 14, 2023

SIGMORPHON aims to bring together researchers interested in applying computational techniques to problems in morphology, phonology, and phonetics. Work that addresses orthographic issues is also welcome. Papers will be on substantial, original, and unpublished research on these topics, potentially including strong work in progress.

10

Venue Information

The Westin Harbour Castle

Location: Located in Toronto (Downtown Toronto), The Westin Harbour Castle, Toronto is within a 15-minute walk of Scotiabank Arena and Harbourfront Centre. This 4-star hotel is 0.7 mi (1.2 km) from Ripley's Aquarium of Canada and 0.8 mi (1.2 km) from CN Tower.

Rooms: Make yourself at home in one of the 977 guestrooms featuring refrigerators and LCD televisions. Your room comes with a pillowtop bed. Wireless Internet access (surcharge) keeps you connected, and cable programming is available for your entertainment. Private bathrooms with shower/tub combinations feature complimentary toiletries and hair dryers. Canada power adapters operate on a 120V supply voltage and 60Hz.

Amenities: Enjoy a range of recreational amenities, including outdoor tennis courts, a health club, and an indoor pool. This hotel also features complimentary wireless Internet access, concierge services, and an arcade/game room.

Dining: Enjoy a meal at Mizzen or snacks in the coffee shop/cafe. The hotel also offers room service (during limited hours). Wrap up your day with a drink at the bar/lounge. Local cuisine breakfasts are available daily from 07:00 to 11:00 for a fee.
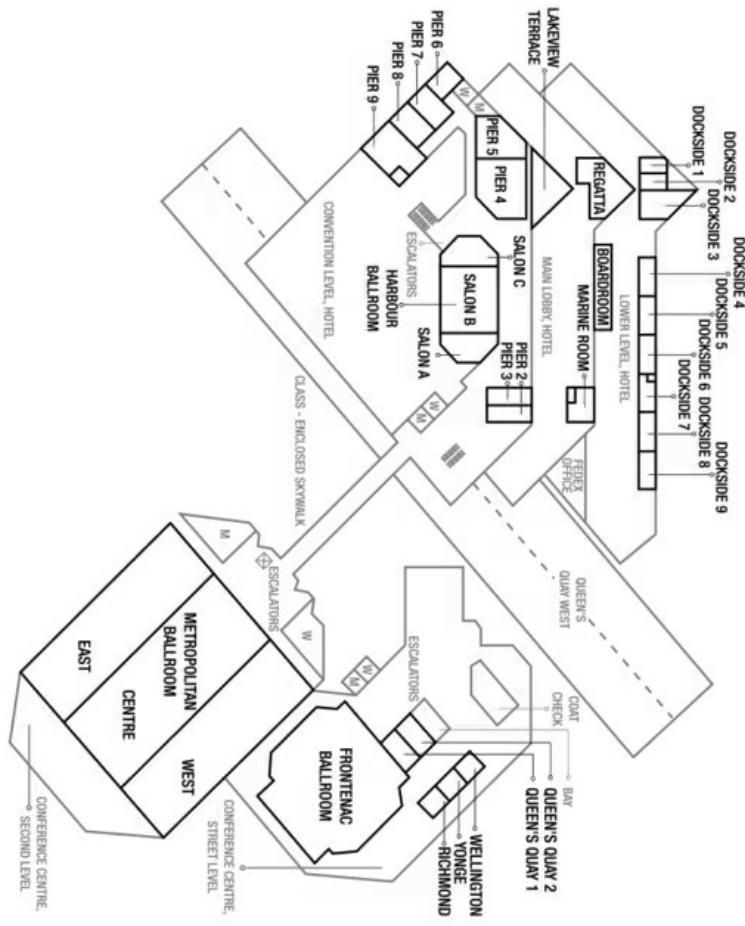
Business Amenities: Featured amenities include a 24-hour business center, express check-in, and express check-out. Planning an event in Toronto? This hotel has facilities measuring 70000 square feet (6503 square meters), including a conference center.

Renovations: The property will be renovating from October 7, 2022 to October 6, 2024 (completion date subject to change). The following areas are affected: Hallway Select guest rooms - During renovations, the hotel will make every effort to minimize noise and disturbance.
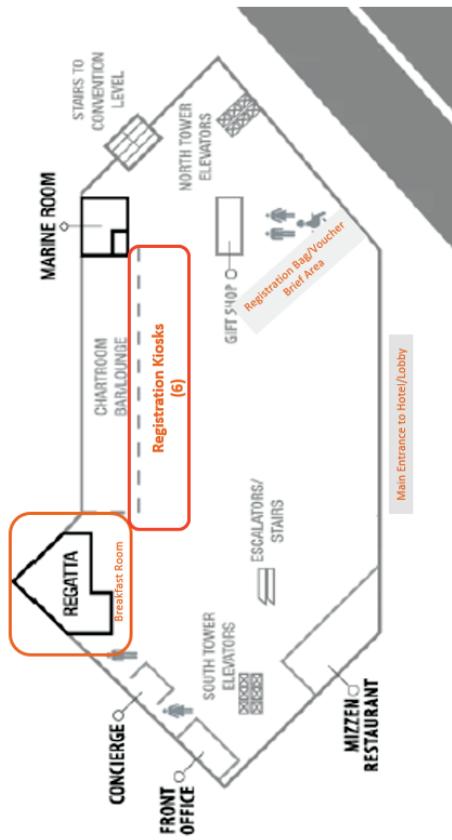
Spoken Languages: English, albeit with a Canadian accent.
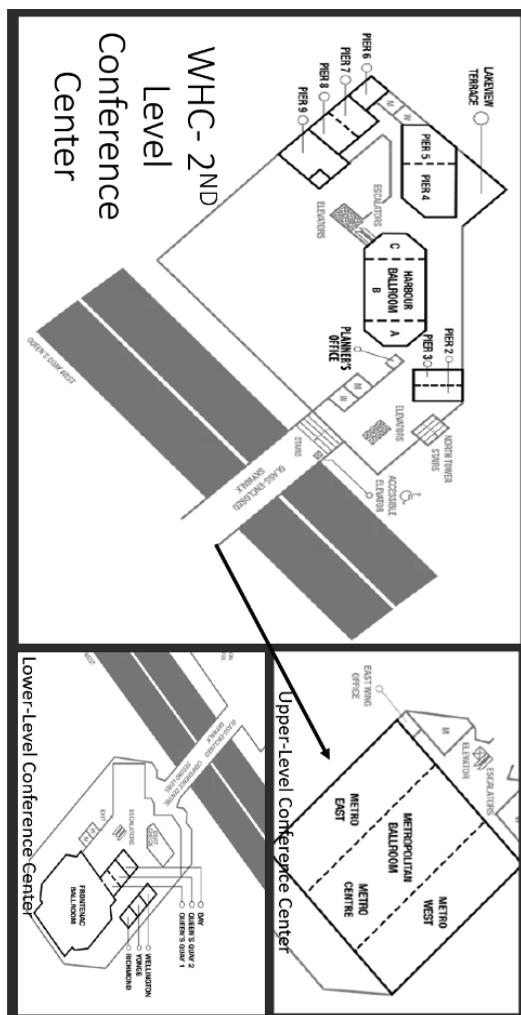
Onsite Payments: Visa, Diners Club, Debit cards not accepted, Discover, American Express, Cash, Mastercard, UnionPay.
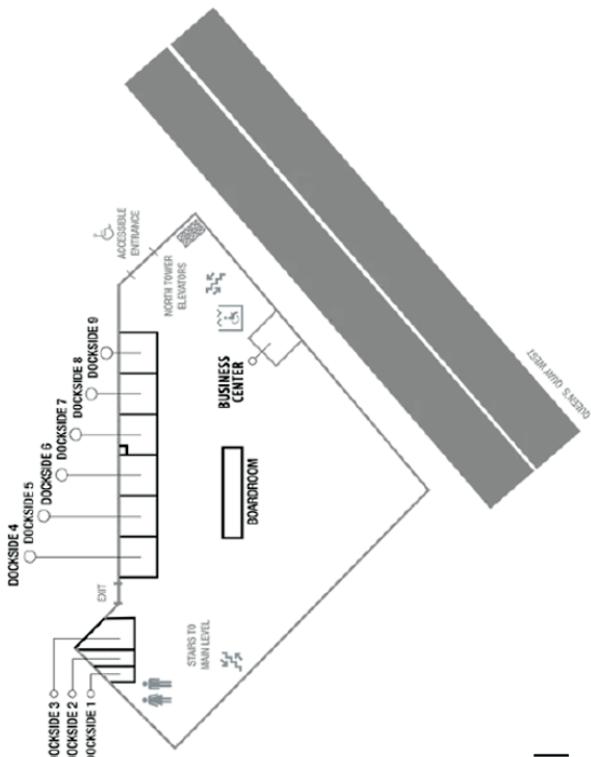
MEETING ROOMS

WHC – MAIN LOBBY (ML)

MARINE ROOM

STAIRS TO CONVENTION LEVEL

NORTH TOWER ELEVATORS

GIFT SHOP

Registration Bag/Voucher Brief Area

CHARTROOM BAR/LOUNGE

Registration Kiosks (6)

Main Entrance to Hotel/Lobby

REGATTA

Breakfast Room

ESCALATORS/ STAIRS

SOUTH TOWER ELEVATORS

CONCIERGE

FRONT OFFICE

MIZZEN RESTAURANT

WHC – 2ND Level Conference Center

Lower-Level Conference Center

Upper-Level Conference Center

WHC - 3rd
Lower Level

Index

Give technology language

Cohere builds high performance, secure language models for the enterprise.

Build incredible products with generative AI

Developers can access Cohere's large language models through its APIs:

Generative: Create product descriptions, blog posts, articles, and more based on adjustable parameters.

Embeddings: Capture the semantic meaning of text for semantic search engines, content moderation, and intent recognition tools.

Exploring the unknown together

Cohere For AI is a research lab that seeks to solve complex ML problems and diversify entry points into ML research through open collaboration.
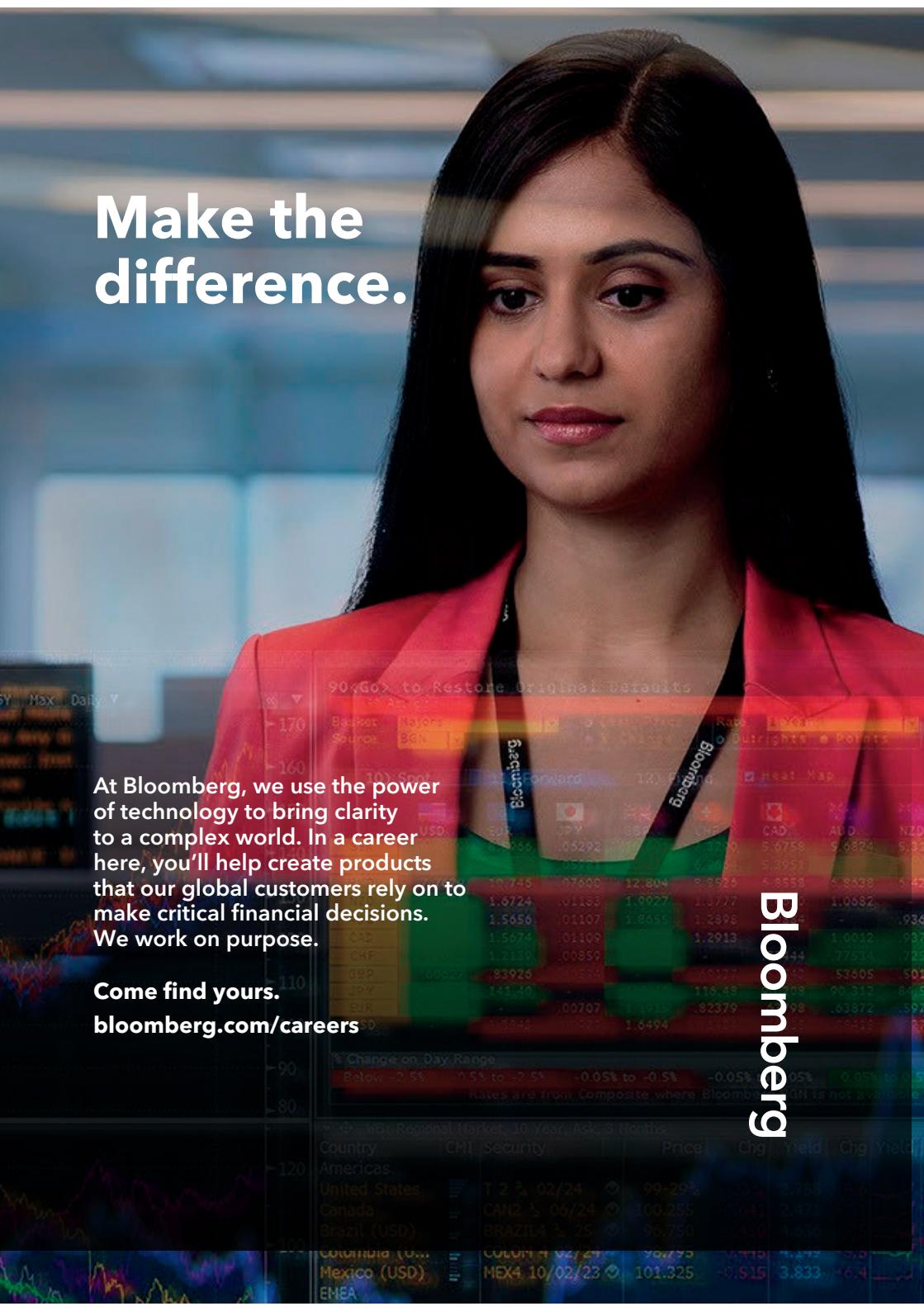
Join us: cohere.for.ai

Level up with LLM University

Learn how LLMs work, how they're useful, and how to deploy them in our free online curriculum.

docs.cohere.com/docs/llmu

cohere

Research

What's Next

IBM Research is home to 3,000 scientists and researchers who deeply believe in the power of the scientific method to invent what's next in computing for our company, for our clients and the world. We are pioneering technologies that will transform industries and society, including the future of AI, hybrid cloud, quantum computing and semiconductors. We've been here since the earliest days of computing, and we're leading the charge for the future. Since our first lab opened in 1945, we've authored more than 110,000 research publications. Our researchers have won six Nobel Prizes, six Turing Awards, and IBM has been granted more than 150,000 patents.

research.ibm.com

HUAWEI
NOAH'S ARK LAB

About us

The Noah's Ark Lab is the AI research center for Huawei Technologies Co. Ltd, located in Hong Kong, Shenzhen, Beijing, Shanghai, Xi'an, Nanjing, Hefei, London, Paris, Toronto, Montreal, Edmonton, etc.

The mission of the lab is to make significant contributions to both the company and society by innovating in artificial intelligence, data mining and related fields.

Research areas of the lab mainly include computer vision, speech and language processing, search and recommendation, decision and reasoning, AI theory and AI system engineering.
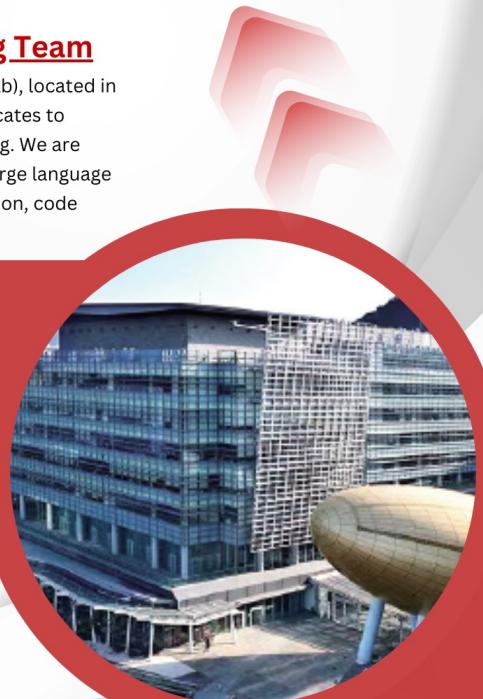
Founded in 2012, the lab has now grown to be a research organization with many significant achievements in both academia and industry. We welcome talented researchers and engineers to join us to realize their dreams.

The Speech and Language Processing Team

The Speech and Language Processing team (of Noah's Ark Lab), located in Hong Kong, Shenzhen, Montreal, London, and Toronto, dedicates to research and applications of speech and language processing. We are developing advanced technologies of the research topics: large language models, speech recognition and synthesis, machine translation, code generation, dialogue and question answering, etc.

APPLICATION:

- Interested candidates should send application materials including resume to haoran.shi@huawei.com
- More information is available at http://www.noahlab.com.hk

HUAWEI   NOAH'S ARK LAB

Supercharge enterprise growth and efficiency with generative AI-powered features

We've been pioneering digital conversational technology for over 27 years. Today, our award-winning Conversational Cloud™ platform empowers hundreds of the world's leading brands to deliver Curiously Human™ experiences that drive extraordinary results.

Drive scientific innovation with Curiously Human LLMs

Discover the transformative power of our Curiously Human approach in maximizing LLMs for data science advancements and effective Conversational AI.

DATA-DRIVEN EXCELLENCE
Enhance personalization and relevance.
Elevate your LLMs with the world's largest conversational dataset, sourced from over a billion monthly interactions.

This wealth of data empowers our AI to understand your customers and provide uniquely tailored experiences.

AI WITH A HUMAN TOUCH
Boost customer satisfaction and retention.
Maintain grounded, factual, and industry-specific conversations with the support of over 350,000 skilled humans in the loop, who continuously refine our models.

Our AI ensures your customer interactions are both accurate and engaging.

ACTIONABLE INSIGHTS
Optimize performance and drive results.
Harness the power of enterprise-level analytics and reporting that automatically delivers actionable insights.

LivePerson's approach to conversational intelligence helps you make data-driven decisions to optimize customer experiences and drive results.

RESPONSIBLE AI
Build trust and ensure compliance.
Minimize the risk of bias and ensure ethical AI implementation by partnering with LivePerson, the founders of Equal AI.
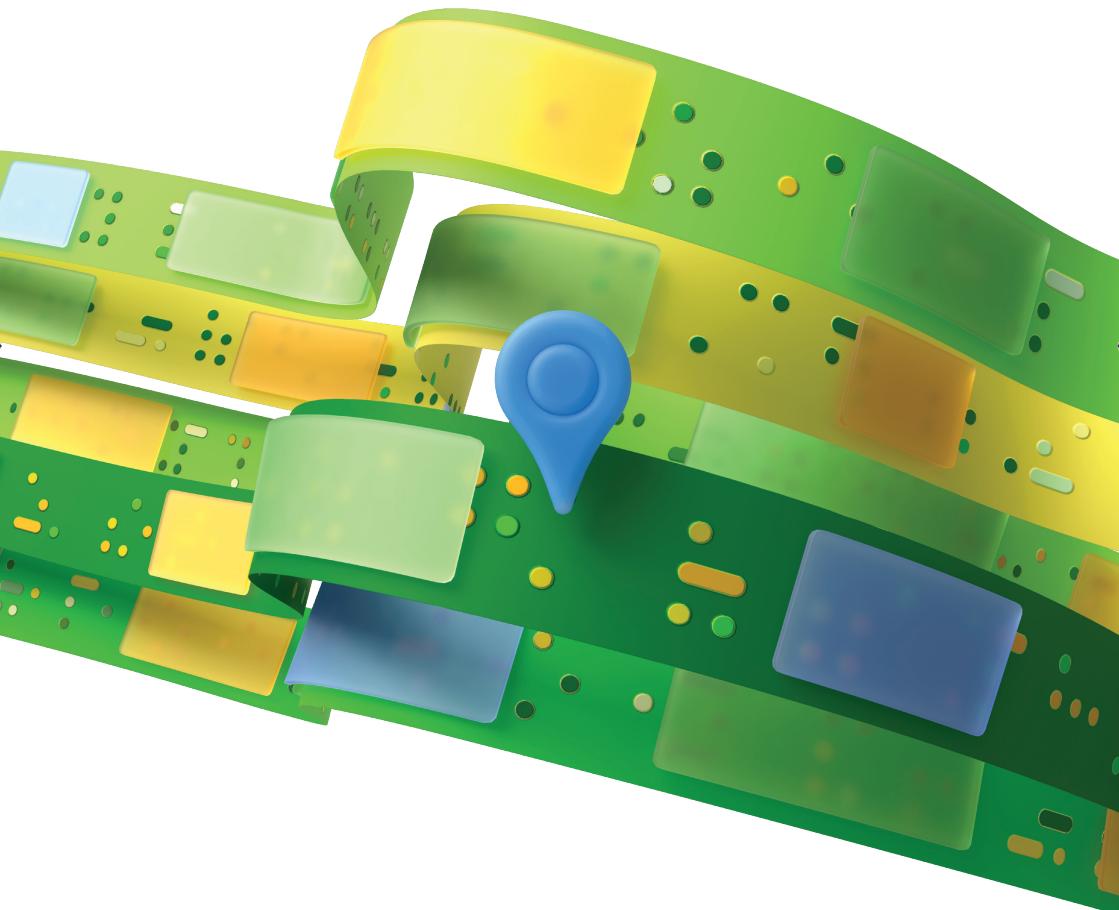
We've been spearheading standards and certification for responsible, safe, and secure AI since 2019.

Discover the LivePerson advantage
Visit our AI hub to learn more https://www.liveperson.com/ai/resources/

We're unlocking the possibilities of AI, together.

Meta AI

Microsoft Research is where leading scientists and engineers have the freedom and support to propel discovery and innovation. Here, they pursue and publish curiosity-driven research in a range of scientific and technical disciplines that can be translated into products. With access to vast computing power, global multi-disciplinary teams tackle complex problems that drive breakthrough technologies and improve lives.

Careers

Imagine having the freedom and resources to pursue and publish curiosity-driven research that tackles complex problems to improve lives. **aka.ms/msrcareers**

Events

Connect with our researchers at conferences and Microsoft Research events around the world. **aka.ms/msrevent**

Microsoft Research Blog

Read in-depth technical and notable articles from our researchers, scientists, and engineers. **aka.ms/msrblog**

Microsoft Research Podcast

Listen in on conversations that bring you closer to the cutting-edge of technology research and the scientists behind it. **aka.ms/msrpod**

Programs

Further your research with fellowships, grants, and opportunities. **aka.ms/msrprog**

Connect with us:

MicrosoftResearch

@MSFTResearch

microsoftresearch

Microsoft Research Group

@msft_research

#msftresearch

Microsoft

Come build the future with us

At Amazon, we believe scientific innovation is essential to building Earth's most customer-centric company. Our scientists are conducting cutting-edge research in areas ranging from conversational AI to machine learning, operations, quantum computing, robotics, and more.

Learn more about our research and papers published at ACL by scanning the QR code, visiting Amazon.Science, or sending us a note at acl-conference-2023@amazon.com
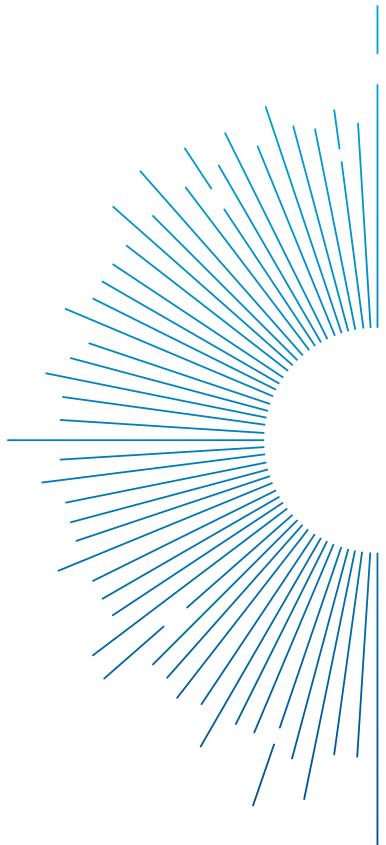
BAIDU NLP

BAIDU NATURAL LANGUAGE PROCESSING

On a mission to enable machines to understand language and acquire intelligence so as to make the world better, Baidu NLP is dedicated to core NLP technologies, leading technology platforms and innovative products that are set to serve users across the globe and make the complex world simpler.

grammarly

All innovation at Grammarly begins with our commitment to building technology to solve real user problems. We responsibly apply advances in machine learning, NLP, and generative AI to develop the world's most comprehensive digital communication assistance technology at scale.

Grammarly helps 30 million people and 50,000 teams write more clearly and effectively every day.

We are a values-driven team of more than 900 across North America and Europe, and we're growing. Join us!

grammarly.com/jobs

Megagon Labs

Knowledge Representation & Reasoning

AI for Data Management

Human Centered AI

Natural Language Processing

We empower people with better information to make their best decision.

www.megagon.ai

Ant Group

Ant Group traces its roots back to Alipay, which was established in 2004 to create trust between online sellers and buyers. Over the years, Ant Group has grown to become one of the world's leading open Internet platforms.

Through technological innovation, we support our partners in providing inclusive, convenient digital life and digital financial services to consumers and SMEs. In addition, we have been introducing new technologies and products to support the digital transformation of industries and facilitate collaboration. Working together with global partners, we enable merchants and consumers to make and receive payments and remit around the world.

Digital Payment   **Digital Connectivity**   **Digital Finance**

Digital Technologies   **Globalization**

🌐 https://www.antgroup.com/

✉ AntResearch@antgroup.com

COMCAST | xfinity

The Comcast Applied AI & Discovery Team invents the technological foundations for the Xfinity experiences of the future.

Voice Control

Content Discovery & Media Analytics

Smart Home & Connectivity

Conversational Agents

Scan to explore open roles and to learn more.

Unable to scan the code?
Visit Comcastcareers.com/ml-ai-team-page
on any mobile device or computer.

Welcome Sponsor

cohere

Diamond Sponsors

Bloomberg Engineering

Apple

Google Research

IBM

Meta AI

LIVEPERSON

Microsoft

Platinum Sponsors

amazon | science

Baidu 百度

ByteDance

HUAWEI

Flitto

Google DeepMind

grammarly

KAUST Artificial Intelligence Initiative

DATAOCEAN AI
YOUR GLOBAL DATA PARTNER

Megagon Labs

Gold Sponsors

ANT GROUP

COMCAST

JPMorgan Chase & Co.

NEC

Tencent

Silver Sponsors

Alibaba Group

aiXplain

BOSCH
Invented for life

duolingo

translated.

Bronze Sponsors

Adobe

servicenow

Babelscape

Diversity & Inclusion Champions

amazon | science

LIVEPERSON

Google DeepMind