# EACL 2023

## 2–6 May | Dubrovnik

# Contents

# 1

# Conference Information

## Message from the General Chair

Welcome to the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL). This is the flagship European conference dedicated to European and international researchers, covering a broad spectrum of research areas of Computational Linguistics and Natural Language Processing.

Organizing a scientific conference of the prestige and size of EACL is always a great honor associated with several challenges. Our team had to tackle unusual complexities: this conference was one of the first scheduled to be in person after the long period of online conferences forced by COVID pandemic. The bidding process for a location, which typically takes place several years before the actual start of the conference, is mainly driven by the aim of expanding and involving the science community of all European countries: EACL selected Kyiv, Ukraine, as the physical location. As you all know, in February 2022, an unpredictable and dramatic event happened, the war between Russian and Ukraine, which made the organization in Kyiv impossible.

Considering the importance of physical interaction among researchers, especially after the restrictions imposed by the COVID pandemic, we worked hard with the EACL and ACL boards to find an alternative location, able to delight our attendees. Our team achieved this seemingly impossible goal of organizing a conference in a new location a few months before its start: we selected Dubrovnik, Croatia, while preserving the original aim of strengthening the connection with the Ukrainian community. In this respect, the Ukraine local committee will feature a dedicated panel session, "Low-resource languages in NLP products", and a workshop to highlight work on Ukrainian language technologies. Following the latest conference, EACL 2023 will be "hybrid," serving both virtual and in-person participants. As our official local chairs are not from the physical location, we needed a local team from Croatia for helping with the logistics. As a result, the main unexpected novelty of EACL 2023 is to have two local organizing committees from two different European countries.

In the remainder of this preface, I would like to thank EACL contributors chronologically with respect to my work timeline for EACL: Roberto Basili and Shuly Wintner, the new and former Presidents of ACL, along with the EACL board – thanks for having trusted me to manage the organization of the conference in rather complicated times. I started to be confident that we would have done a good job after Isabelle Augenstein and Andreas Vlachos accepted the role of PC Chairs. They have performed amazing work, creating an outstanding program, and also helping me in recruiting our fantastic organization team. A special thank is due to Preslav Nakov (EACL officer) for his support: thanks to his action, the proactiveness of David Yarowsky, and the fairless effort of Jennifer Rachford (our new secretary of the ACL business office), we successfully implemented the apparently unrealistic idea of switching from the already planned online conference to a hybrid setting with a physical location in Dubrovnik. Regarding the online side of our hybrid conference, we partnered with Underline (Sol Rosenberg, Damira Mrsic and Luka Simic), who also gave us support for managing the entire conference. While finalizing the location, we started to activate the different sections of the conference, for which my acknowledgements are again in chronological order:

- Ukraine Local Committee, Viktoria Kolomiets, Mariana Romanyshyn, Oleksii Molchanovskyi, Oles Dobosevych, was instrumental in preserving our initial goal of connecting the Ukraine research community, organizing a panel and a workshop.
- The website chairs, Pepa Atanasova and Julius Cheng, started immediately to design our website, even when almost no information was available.
- The workshop chairs, Zeerak Talat and Antonio Toral, selected our conferences and led the selection of workshops for the joint ACL call.
- The tutorial chairs, Sameer Pradhan and Fabio Massimo Zanzotto, together with the ACL chairs, took care of the tutorial selection for the ACL related conferences.
- The demonstration chairs, Danilo Croce and Luca Soldaini, created a parallel conference program to select exciting demos.
- The Publicity Chairs, Laura Biester, Leshem Choshen and Joel Tetrault, have been our interface with the science community through social media platforms.
- The Publication Chairs, Carolina Scarton and Ryan Cotterell, produced high-quality proceedings, thanks to their competence and experience.
- The diversity and inclusion chairs, Sara Tonelli, Elena Cabrio, Verena Rieser, Spandana Gella, took care of DI and performed an amazing job, also working on hundreds applications.
- The Local Organising Committee of Croatia, Marko Tadić, Krešimir Šojat, and Daša Farkaš, gave essential help for the logistics, Visa, and student volunteers.
- Student Research Workshop Chairs, Matthias Lindemann, Alban Petit, and Elisa Bassignana, along with their faculty advisors Valerio Basile and Natalie Schluter, helped in setting the bases for forming great NLP researchers of the future.
- Our entire program committee, Senior Area Chairs, Area Chairs, reviewers, and best paper committee, was essential for obtaining our high-quality scientific program.
- The ACL's sponsorship director Chris Callison-Burch took care of our sponsorships.
- The student volunteers, as usual, are essential for a successful conference execution.
- Priscilla Rasmussen, our former ACL business office secretary, continued to provide us with useful advice.

Finally, I would like to thank our sponsors for helping us to fund scholarships and DI initiatives.

Alessandro Moschitti
Amazon Alexa AI, Los Angeles, USA
EACL 2023 General Chair

# ACL Statement on the Ukraine situation

March 11, 2022

The Association for Computational Linguistics (ACL) condemns in the strongest possible terms the actions of the Russian Federation government in invading the sovereign state of Ukraine and engaging in war against the Ukrainian people. We stand together with Ukrainian NLP colleagues, the Ukrainian people, Russian NLP colleagues and Russian people who condemn the actions of the Russian Federation government, and all those around the world who have been impacted by the invasion.

As a small token of our solidarity with the Ukrainian people, the ACL has decided to temporarily sever its ties with Russia-based organizations, while at the same time allowing Russian scientists to remain part of the ACL community. In practice, this means that the ACL will refrain from accepting any sponsorship or allowing any exhibits from Russian-headquartered entities at ACL-run events. Russian scholars are still welcome to participate in ACL events and publish at ACL venues.

The ACL is committed to peace and condemns any form of violence and harassment. We are also committed to peaceful co-operation, mutual understanding, and tolerance across borders. NLP scholars from both Ukraine and Russia are welcome to get in touch with the ACL with any concerns.

Tim Baldwin, on behalf of the ACL Executive

# Message from the Program Chairs

Welcome to the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL). After the last edition in 2021 having been held fully online due to the COVID pandemic, EACL 2023 is being held in "hybrid" mode this year, serving both virtual and in-person participants in Dubrovnik, Croatia. While the original plan was to hold the conference in Kyiv (which was the plan originally for EACL 2021), the ongoing war made the organisation in Ukraine impossible. In order to ensure that the original aim of strengthening the connections with the Ukrainian community is still served, our program features a dedicated session and a workshop to highlight work on Ukrainian language technologies.

# Submission and Acceptance

EACL 2023 accepted direct submissions, as well as submissions via ARR. For direct submissions, abstracts were needed to be registered one week prior to the submission date.

In total, EACL 2023 received 1550 submissions, the largest number to date, with the 2021 edition having received 1400 submissions. Out of those, 1045 were long and 505 were short paper submissions. 81 were ARR papers that were committed to EACL. 249 submissions were withdrawn throughout the reviewing process, including before the full paper submission deadline. 55 papers were desk rejected for various reasons (missing the limitations section, anonymity policy, multiple submission policy, plagiarism or formatting violations).

By the time we as the programme chairs made acceptance decisions, 1166 submissions were still active in the system. We kept the acceptance rate in line with previous *ACL conferences, resulting in 281 papers accepted to the main conference (24.1%), and 201 papers accepted to the Findings of EACL (17.2%), with the remaining 58.7% being rejected. One paper accepted to the main conference and four papers accepted to Findings were subsequently withdrawn. Out of the final set of accepted main conference papers, we invited 178 to be presented orally, and all 281 papers accepted to the main conference to be presented during in-person sessions, as well as a plenary virtual poster session. The EACL 2023 program also features six papers from the Transactions of the Association for Computational Linguistics (TACL) journal, and one from the Computational Linguistics (CL) journal.

# Limitations Section

Following EMNLP 2022, 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.

# Areas

To ensure a smooth process, the submissions to EACL 2023 were divided into 21 areas. The areas mostly followed these of previous EACL, and more broadly *ACL conferences, reflecting the typical divisions in the field. We also had a special area for papers for which both SACs had a conflict of interest. Those papers were reviewed by the reviewers and ACs in their original areas, but the paper recommendations were made by a dedicated SAC, who was a senior member of the NLP community. The most popular areas with over 100 submissions were "Generation and Summarization", "Language Resources and Evaluation", and "Machine Learning in NLP".

# Best Paper Awards

From the papers submitted to EACL 2023, we selected 25 papers accepted to the main conference as candidates for a Best Paper award, based on nominations by the reviewers. These papers were assessed by the Best Paper Award Committee, who also determined the types of paper awards, following the ACL Conference Awards Policy. The selected best papers and runner-ups will be announced in a dedicated plenary session for Best Paper Awards on 4 May 2023.

# Programme Committee Structure and Reviewing

Similar to prior NLP conferences, we adopted the hierarchical program committee structure, where for each area we invited 1-2 Senior Area Chairs (SACs), who worked with a team of Area Chairs (ACs), and a larger team of reviewers. We relied on statistics from prior years to estimate how many SACs, ACs and reviewers would be needed and ended up with 43 SACs, 118 ACs and 1634 reviewers. For identifying ACs and reviewers, we used the reviewer lists from prior *ACL conferences, and also encouraged all EACL 2023 authors to serve as reviewers, using a mandatory form requesting further information on their ability to serve as ACs, reviewers or emergency reviewers, which authors had to fill in on Softconf when registering their abstracts. We passed this information on to SACs, who were responsible for recruiting ACs and reviewers.

Rather than making assignments using a matching algorithm, we asked ACs and reviewers to bid on registered abstracts within their areas, to achieve a better fit. We went with this solution as the number of papers per area was relatively small, and we wanted to avoid poor reviewing assignments as much as possible. We then made an initial paper assignment, in which we ensured that each paper would be reviewed by at least one reviewer who bidded "yes" for the submission, and by no reviewers who bidded "no" for the submission.

Afterwards, we asked the SACs to fine-tune the allocations, and ensure each paper had one AC and three reviewers assigned to it.

To ensure the review quality, we provided detailed guidelines about what reviewers should and shouldn't do in a review, based on the EMNLP 2022 guidelines. We also asked reviewers to flag papers for potential ethical concerns.

For pre-reviewed ARR papers, we asked SACs to not rely mainly on the reviewer scores, but to make their recommendations based on the text of the reviews, meta-reviews and the papers themselves. For making acceptance decisions, we mostly followed SAC recommendations, though also taking into account the overall quality of papers submitted to the conference. Where recommendations seemed overly harsh or lenient given the reviewers' scores, reviews, author responses, or discussions amongst reviewers, we engaged in a dialogue with the respective SACs to make the final decision about the papers in question.

# Ethics Committee

We also formed an Ethics Committee (EC) dedicated to ethical issues. The ethics committee considered 21 papers that were flagged by the technical reviewing committee for ethical concerns. Out of these, 10 were conditionally accepted, meaning the ethics issues had to be addressed in the camera-ready version, to be verified by the EC prior to final acceptance, and the other 11 were accepted as is. The authors of all conditionally accepted papers submitted the camera-ready version and a short response that explained how they had made the changes requested by the EC. The EC double-checked these revised submissions and responses, and confirmed that the ethical concerns had been addressed. As a result, all conditionally accepted papers were accepted to the main conference or Findings.

# ACL Rolling Review

ACL Rolling Review (ARR) is an initiative of the Association for Computational Linguistics, where the reviewing and acceptance of papers to publication venues are done in a two-step process: (1) centralized rolling review and (2) the ability to commit the reviewed papers to be considered for publication by a publication venue. For EACL 2023, we decided to follow EMNLP 2022's example and run a process which is separate from ARR, but also allows for ARR submissions. Specifically, authors could either submit papers to EACL 2023 directly, or commit ARR reviewed papers by a certain date. We coordinated with the ARR team to extract the submission, review and meta-review from the OpenReview system, according to a submission link that the author provided when committing their ARR submission to EACL. The ARR commitment deadline was set one month after the direct submission deadline since the ARR submissions already have their reviews and meta-recommendation. These ARR papers were then ranked by the SACs together with the direct submissions in the track, and based on the reviews and meta-reviews from ARR. Overall, EACL had 81 papers committed from ARR, of these 24 were accepted to the main conference and 20 were accepted to Findings of EACL.

# Presentation Mode

We made the decision on which papers would be invited for oral poster presentations based on several factors: the relative rank of the paper according to SAC recommendation, whether the paper had been recommended for a best paper award by at least one reviewer, and for TACL and CL papers, the authors' preference of presentation mode.

# Keynotes and Panels

Another highlight of our program are the plenary sessions, for which we scheduled three talks, as well a panel:

- a keynote talk by Joyce Chai (University of Michigan) on “Language Use in Embodied AI"
- a keynote talk by Edward Greffenstette (Cohere AI and University College London) on “Going beyond the benefits of scale by reasoning about data”
- a keynote talk by Kevin Munger (Penn State University) on "Chatbots for Good and Evil"
- a panel on “low-resource languages in NLP products” led by Mariana Romanyshyn with Viktoria Kolomiets (Grammarly), Mariana Romanyshyn (Grammarly), Oleksii Molchanovskyi (Ukrainian Catholic University) and Oles Dobosevych (Ukrainian Catholic University)

# Thank Yous

EACL 2023 is the result of a collaborative effort and a supportive community, and we want to acknowledge the efforts of so many people with whom we worked directly and made significant efforts in putting together the programme for EACL 2023!

- Our General Chair, Alessandro Moschitti, who led the whole organising team, and helped with many of the decision processes;
- Our 43 Senior Area Chairs, who were instrumental in every aspect of the review process, from recruiting Area Chairs, correcting reviewer assignments, to making paper acceptances;
- Our 118 Area Chairs, who had the role of interacting with the reviewers, leading paper review discussions, and writing meta-reviews;
- The 1634 reviewers, who provided valuable feedback to the authors; The emergency reviewers, who provided their support at the last minute to ensure a timely reviewing process;

- Our Best Paper Selection Committee, who selected the best papers and the outstanding papers: Jonathan Kummerfeld (chair), Joakim Nivre, Bonnie Webber, Thamar Solorio and Hanna Hajishirzi;

- Our Ethics Committee, chaired by Zeerak Talat, for their hard work to ensure that all the accepted papers addressed the ethical issues appropriately, under a very tight schedule;

- Our amazing Publication Chairs, Carolina Scarton and Ryan Cotterell for compiling the proceedings in good time for the conference;

- Our Publicity Chairs, Laura Biester, Leshem Choshen and Joel Tetrault, for their work on managing the communications on social media platforms;

- Our website chairs, Pepa Atanasova and Julius Cheng for putting together the website for the conference and keeping it up to date;

- Damira Mrsic from Underline, for her support in developing the virtual conference platform;

- Jennifer Rachford, who has worked tirelessly online and on-site to ensure that EACL 2023 is a success.

We're looking forward to a great EACL 2023!

Isabelle Augenstein (University of Copenhagen, Denmark)
Andreas Vlachos (University of Cambridge, UK)
EACL 2023 Programme Committee Co-Chairs

# Organizing Committee

**General Chair**

Alessandro Moschitti, Amazon Alexa

**Program Chairs**

Isabelle Augenstein, University of Copenhagen
Andreas Vlachos, University of Cambridge

**Publications Chairs**

Ryan Cotterell, ETH Zürich
Carolina Scarton, University of Sheffield

**Workshop Chairs**

Zeerak Talat, Simon Fraser University
Antonio Toral, University of Groningen

**Tutorials Chairs**

Sameer Pradhan, University of Pennsylvania
Fabio Massimo Zanzotto, University of Rome, “Tor Vergata”

**Demonstrations Chairs**

Danilo Croce, University of Rome, “Tor Vergata”
Luca Soldaini, Allen Institute for AI

**Publicity Chairs**

Joel Tetreault, Dataminr
Leshem Choshen, IBM AI research; Hebrew University of Jerusalem
Laura Biester, University of Michigan

**Website Chairs**

Pepa Atanasova, University of Copenhagen
Julius Cheng, University of Cambridge

**Sponsorship Director**

Chris Callison-Burch, University of Pennsylvania

**Diversity and Inclusion Chairs**

Elena Cabrio, Université Côte d’Azur, Inria, CNRS, I3S
Sara Tonelli, Fondazione Bruno Kessler
Verena Rieser, Heriot-Watt University
Spandana Gella, Amazon Alexa

**Student Research Workshop Chairs**

Matthias Lindemann, University of Edinburgh
Alban Petit, Université Paris-Saclay
Elisa Bassignana, IT University of Copenhagen

**Student Research Workshop Faculty Advisors**

Valerio Basile, University of Turin
Natalie Schluter, IT University of Copenhagen; Apple

**Local Organising Committee**

Marko Tadić, University of Zagreb
Krešimir Šojat, University of Zagreb
Daša Farkaš, University of Zagreb

**Ukraine Local Committee**

Viktoria Kolomiets, Grammarly
Mariana Romanyshyn, Grammarly
Oleksii Molchanovskyi, Ukrainian Catholic University
Oles Dobosevych, Ukrainian Catholic University

# Program Committee

**Anaphora, Discourse and Pragmatics**

Bonnie Webber, University of Edinburgh
Michael Strube, Heidelberg Institute for Theoretical Studies

**Computational Social Science and Social Media**

Maria Liakata, Queen Mary University of London
Kalina Bontcheva, University of Sheffield

**Conflicts of Interests**

Joakim Nivre, Research Institutes of Sweden

**Dialogue and Interactive Systems**

Diarmuid Ó Séaghdha, Apple
Matthew Purver, Queen Mary University of London

**Document analysis, Text Categorization and Topic Models**

Nikolaos Aletras, University of Sheffield
Ekaterina Shutova, University of Amsterdam

**Ethical and Sustainable NLP**

Nafise Sadat Moosavi, Department of Computer Science, The University of Sheffield
Yonatan Belinkov, Technion

**Ethics Review**

Zeerak Talat, Simon Fraser University

**Generation and Summarization**

Ondrej Dusek, Charles University
Chenghua Lin, Department of Computer Science, University of Sheffield

**Information Extraction**

Roberto Navigli, Sapienza University of Rome
Mrinmaya Sachan, ETH Zurich

**Information Retrieval and Search**

Bruno Martins, IST and INESC-ID
Fabrizio Silvestri, Sapienza, University of Rome

**Interpretability and Model Analysis**

Dong Nguyen, Utrecht University
Roi Reichart, Technion - Israel Institute of Technology

**Language Grounding and Multi-Modality**

Grzegorz Chrupała, Tilburg University
Desmond Elliott, University of Copenhagen

**Language Resources and Evaluation**

Roman Klinger, University of Stuttgart
Omri Abend, The Hebrew University of Jerusalem

**Linguistic Theories, Cognitive Modeling and Psycholinguistics**

Barry Devereux, Queen's University, Belfast
Natalie Schluter, IT University of Copenhagen

**Machine Learning for NLP**

James Henderson, Idiap Research Institute
Vlad Niculae, University of Amsterdam

**Machine Translation**

Wilker Aziz, University of Amsterdam
Rico Sennrich, University of Zurich

**Multidisciplinary and other NLP Applications**

Annie Priyadarshini Louis, Google Research UK
Yulan He, King's College London

**Multilinguality**

Ivan Vulić, University of Cambridge
Alexander Fraser, Ludwig-Maximilians-Universität München

**Phonology, Morphology, and Word Segmentation**

Thierry Poibeau, LATTICE (CNRS and ENS/PSL)
François Yvon, ISIR CNRS and Sorbonne Université

**Question Answering**

Jonathan Berant, Tel Aviv University and AI2
Pontus Stenetorp, University College London

**Semantics: lexical**

Chris Biemann, Universität Hamburg
Mark Stevenson, University of Sheffield

**Semantics: sentence level and other areas**

Aliaksei Severyn, Google
Douwe Kiela, Hugging Face

**Sentiment Analysis and Argument Mining**

Veronique Hoste, LT3, Ghent University
Ivan Habernal, Technical University of Darmstadt

**Tagging, Chunking, Syntax and Parsing**

Marco Kuhlmann, Linköping University
Shay B. Cohen, University of Edinburgh

**Area Chairs**

Khalid Al Khatib, Malihe Alikhani, Mikel Artetxe, Akari Asai, Duygu Ataman, Niranjan Balasubramanian, Jeremy Barnes, Max Bartolo, Valerio Basile, Laurent Besacier, Iacer Calixto, Kris Cao, Tanmoy Chakraborty, Bharathi Raja Chakravarthi, Guanyi Chen, Wenhu Chen, Eleanor Chodroff, Caio Corro, Çağrı Çöltekin, Orphee De Clercq, Miryam De Lhoneux, Tejaswini Deoskar, Rotem Dror, Yanai Elazar, Arash Eshghi, Amir Feder, Yang Feng, Radu Florian, Anette Frank, Markus Freitag, André Freitas, Annemarie Friedrich, Jie Fu, Wei Gao, Mor Geva, Dan Goldwasser, Yulia Grishina, Lin Gui, Sara Hooker, Shujian Huang, Patrick Huber, Dieuwke Hupkes, Peter Jansen, Kristen Johnson, David Jurgens, Simon Keizer, Casey Kennington, Daniel Khashabi, Tushar Khot, Germán Kruszewski, Lun-wei Ku, Matthieu Labeau, Gerasimos Lampouras, Mirella Lapata, Gabriella Lapesa, Anne Lauscher, Piji Li, Bin Liang, Bang Liu, Craig Macdonald, Pranava Madhyastha, Andrea Madotto, Saad Mahamood, Jonathan Mallinson, Zita Marinho, Eugenio Martínez Cámara, Florian Metze, Sabrina Mielke, Pushkar Mishra, Saif Mohammad, Preslav Nakov, Shashi Narayan, Debora Nozza, Alexander Panchenko, Alexandros Papangelis, Nikolaos Pappas, Panupong Pasupat, Gabriele Pergola, Jonas Pfeiffer, Yuval Pinter, Edoardo Maria Ponti, Christopher Potts, Daniel Preotiuc-pietro, Emily Prud'hommeaux, Simon Razniewski, Siva Reddy, Sara Rosenthal, Alla Rozovskaya, Keisuke Sakaguchi, Tanja Samardzic, Minjoon Seo, Ehsan Shareghi, Ravi Shekhar, Carina Silberer, Miikka Silfverberg, Felix Stahlberg, Svetlana Stoyanchev, Saku Sugawara, Hiroya Takamura, Niket Tandon, Sara Tonelli, Nicola Tonellotto, Lonneke Van Der Plas, David Vandyke, David Vilares, Elena Voita, Svitlana Volkova, Shuai Wang, Derry Tanti Wijaya, Adina Williams, Wei Wu, Fei Xia, Deyi Xiong, Ikuya Yamada, Marcos Zampieri, Sina Zarrieß, Chrysoula Zerva, Arkaitz Zubiaga

**Reviewers**

Sadaf Abdul Rauf, Muhammad Abdul-mageed, Ibrahim Abu Farha, Lasha Abzianidze, Angus Addlesee, David Adelani, Stergos Afantenos, Severine Affeldt, Rodrigo Agerri, Piush Aggarwal, Željko Agić, Ameeta Agrawal, Roee Aharoni, Wasi Uddin Ahmad, Sina Ahmadi, Natalie Ahn, Xi Ai, Laura Aina, Akiko Aizawa, Md. Shad Akhtar, Hend Al-khalifa, Rami Al-rfou, Ahmed Alajrami, David Alfter, Bashar Alhafni, Hamed Alhoori, Hafsa Ali, Afra Alishahi, Miguel A. Alonso, Sultan Alrowili, Bharat Ram Ambati, Silvio Amir, Samuel Amouyal, Reinald Kim

Amplayo, Jisun An, Vishal Anand, Raviteja Anantha, Antonios Anastasopoulos, Tim Anderson, Melanie Andresen, Anelia Angelova, Alan Ansell, Francesco Antici, Diego Antognini, Maria Antoniak, Dimosthenis Antypas, Reut Apel, Emilia Apostolova, Jun Araki, Oscar Araque, Arturo Argueta, Akhil Arora, Ekaterina Artemova, Elliott Ash, Md.sadek Hossain Asif, Arian Askari, Zhenisbek Assylbekov, Aitziber Atutxa Salazar, Eleftherios Avramidis, Cem Rifki Aydin, Mahmoud Azab, Bharathi B, Jinheon Baek, Selene Baez Santamaria, Parsa Bagherzadeh, Vikas Bahirwani, Fan Bai, Jinyeong Bak, Timothy Baldwin, Miguel Ballesteros, Forrest Sheng Bao, Edoardo Barba, Francesco Barbieri, Ander Barrena, Pierpaolo Basile, Roberto Basili, Ali Basirat, Riza Batista-navarro, Timo Baumann, Rachel Bawden, Christos Baziotis, Ian Beaver, Nadia Bebeshina, Frederic Bechet, Tilman Beck, Beata Beigman Klebanov, Tadesse Destaw Belay, Meriem Beloucif, Farah Benamara, Luca Benedetto, Joshua Bensemann, Gábor Berend, Thales Bertaglia, Michele Bevilacqua, Rasika Bhalerao, Rohan Bhambhoria, Rishabh Bhardwaj, Sumit Bhatia, Arnab Bhattacharya, Rajarshi Bhowmik, Zhen Bi, Iman Munire Bilal, Alexandra Birch, Debmalya Biswas, Eduardo Blanco, Nate Blaylock, Su Lin Blodgett, Jelke Bloem, William Boag, Ben Bogin, Francis Bond, Georgeta Bordea, Logan Born, Emanuela Boros, Elizabeth Boschee, Cristina Bosco, Zied Bouraoui, Tom Bourgeade, Laurestine Bradford, Stephanie Brandl, Ana Brassard, Jonathan Brophy, Caroline Brun, Christian Buck, Sven Buechel, Paul Buitelaar, Razvan Bunescu, Laurie Burchell, Miriam Butt, Jan Buys, Lisa Bylinina, Bill Byrne, Laura Cabello Piqueras, Elena Cabrio, Samuel Cahyawijaya, Agostina Calabrese, Nitay Calderon, Eduardo Calò, Jose Camacho-collados, Ricardo Campos, Marie Candito, Shuyang Cao, Ziqiang Cao, Fabio Carrella, Xavier Carreras, Jorge Carrillo-de-albornoz, Lucien Carroll, Fabio Casati, Tommaso Caselli, Pierluigi Cassotti, Francesco Cazzaro, Amanda Cercas Curry, Dumitru-clementin Cercel, Christophe Cerisara, Alessandra Cervone, Rahma Chaabouni, Haixia Chai, Tuhin Chakrabarty, Yllias Chali, Ilias Chalkidis, Hou Pong Chan, Zhangming Chan, Anshuma Chandak, Senthil Chandramohan, Buru Chang, Ernie Chang, Yung-chun Chang, Guan-lin Chao, Emile Chapuis, Shubham Chatterjee, Rochana Chaturvedi, Kushal Chawla, Ciprian Chelba, Canyu Chen, Chacha Chen, Chung-chi Chen, Derek Chen, Fuxiang Chen, Hsin-hsi Chen, Jie Chen, Lei Chen, Lei Chen, Meng Chen, Mingda Chen, Qian Chen, Qianglong Chen, Qibin Chen, Qingcai Chen, Sanxing Chen, Shizhe Chen, Tongfei Chen, Xiaoli Chen, Xiuying Chen, Yan-ying Chen, Yi-pei Chen, Yunmo Chen, Zhiyu Chen, Fei Cheng, Shanbo Cheng, Emmanuele Chersoni, Ethan A. Chi, Jenny Chim, Hyundong Cho, Key-sun Choi, Alexandra Chronopoulou, George Chrysostomou, Alessandra Teresa Cignarella, Philipp Cimiano, Elizabeth Clark, Chloé Clavel, Simon Clematide, Ann Clifton, Miruna Clinciu, Oana Cocarascu, Davide Colla, Andrei Coman, Simone Conia, John Conroy, Paul Cook, Gonçalo Correia, Israel Cuevas, Peng Cui, Shaobo Cui, Tonya Custis, Arthur Câmara, Thenmozhi D., Jeff Da, Giovanni Da San Martino, Raj Dabre, Gautier Dagan, Deborah Dahl, Wenliang Dai, Xiang Dai, Rumen Dangovski, Falavigna Daniele, Verna Dankers, Aswarth Abhilash Dara, Franck Dary, Mithun Das Gupta, Saurabh Dash, Brian Davis, Heidar Davoudi, Michiel De Jong, Loic De Langhe, Budhaditya Deb, Alok Debnath, Thierry Declerck, Mathieu Dehouck, Luciano Del Corro, Sebastien Delecraz, Vera Demberg, David Demeter, Steve Deneefe, Yuntian Deng, Pascal Denis, Nina Dethlefs, Daniel Deutsch, Murthy Devarakonda, Hannah Devinney, Prajit Dhar, Shehzaad Dhuliawala, Luigi Di Caro, Mona Diab, Shizhe Diao, Gaël Dias, Caiwen Ding, Chenchen Ding, Liang Ding, Nemanja Djuric, Giovanna Maria Dora Dore, Bonaventure F. P. Dossou, Jad Doughman, Doug Downey, Gabriel Doyle, Mauro Dragoni, Rotem Dror, Jinhua Du, Yupei Du, Xiangyu Duan, Pablo Duboue, Philipp Dufter, Kevin Duh, Ewan Dunbar, Jonathan Dunn, Gerard Dupont, Nadir Durrani, Ritam Dutt, Oliver Eberle, Sauleh Eetemadi, Steffen Eger, Annerose Eichel, Bryan Eikema, Julian Eisenschlos, Heba Elfardy, Micha Elsner, Saman Enayati, Aykut Erdem, Akiko Eriguchi, Katrin Erk, Ramy Eskander, Alex Fabbri, Marzieh Fadaee, Fahim Faisal, Neele Falk, Federico Fancellu, Qixiang Fang, Hossein Fani, Stefano Faralli, Oladimeji Farri, Nawshad Farruque, Manaal Faruqui, Mehwish Fatima, Adam Faulkner, Pedro Faustini, Marc Feger, Nils Feldhus, Anna Feldman, Ghazi Felhi, Mariano Felice, Weixi Feng, Yue Feng, Manos Fergadiotis, Patrick Fernandes, Daniel Fernández-gonzález, Elisabetta Fersini, George Filandrianos, Elena Filatova, Mark Fishel, Lucie Flek, Michael Flor, Negar Foroutan Eghlidi, Jennifer Foster, Stella Frank, Jesse Freitas, Simona Frenda, Annemarie Friedrich, Lisheng Fu, Fumiyo Fukumoto, Kotaro Funakoshi, David Gaddy, Andrea Galassi, Leilei Gan, Yujian Gan, William Gantt, Junbin Gao, Qiaozi Gao, Shen Gao, Muskan Garg, Guillermo Garrido, Susan Gauch, Gregor Geigle, Zorik Gekhman, Alborz Geramifard, Felix Gervits, Mozhdeh Gheini, Reshmi Ghosh, Sucheta Ghosh, Voula Giouli, Dimitris Gkoumas, Serge Gladkoff, Catalina Goanta, Jonas Golde, Seraphina Goldfarb-tarrant, Sujatha Das Gollapalli, Jose Manuel Gomez-perez, Jeff Good, Philip John Gorinski, Koustava Goswami, Isao Goto, Christan Grant, Thomas Green, Derek Greene, Milan Gritta, Paul Groth, Julian Grove, Adam Grycner, Jiasheng Gu, Jiuxiang Gu, Xiaodong Gu, Yi Guan, Marco Guerini, Nuno M. Guerreiro, Xiaoyu Guo, Yanzhu Guo, Zhihui Guo, Abhinav Gupta, Ankit Gupta, Ankita Gupta, Ashim Gupta, Pranjal Gupta, Izzeddin Gur, Suchin Gururangan, Ximena Gutierrez-vasques, Jeremy Gwinnup, Tunga Güngör, Le An Ha, Katharina Haemmerl, Gholamreza Haffari, Joonghyuk Hahn, Michael Hahn, Udo Hahn, Eva Hajicova, Dilek Hakkani-tur, Kishaloy Halder, Karina Halevy, Jiuzhou Han, Lifeng Han, Ting Han, Xudong Han, Yo-sub Han, Viktor Hangya, Sanda Harabagiu, Mareike Hartmann, Sadid A. Hasan, Sabit Hassan, Nabil Hathout, Amartya Hatua, Annette Hautli-janisz, Adi Haviv, Yoshihiko Hayashi, Shirley Anugrah Hayati, T. J. Hazen, Rishi Hazra, Han He, Wanwei He, Wei He, Xiaoting He, Xuanli He, Xuehai He, Yun He, Behnam Hedayatnia, Kevin Heffernan, Benjamin Heinzerling, Jindřich Helcl, William Held, Leonhard Hennig, Christian Herold, Jonathan Herzig, Gerhard Heyer, Derrick Higgins, Anthony Hills, Tatsuya Hiraoka, Vinh Thinh Ho, Cuong Hoang, Eben Holderness, Takeshi Homma, Ales Horak, Andrea Horbach, Sho Hoshino, Md Azam Hossain, Feng Hou, Yifan Hou, Yufang Hou, Shu-kai Hsieh, I-hung Hsu, Han Hu, Po Hu, Xinyu Hua, Chieh-yang Huang, Fei Huang, Hen-hsen Huang, Jie Huang, Junbo Huang, Kuan-hao Huang, Quzhe Huang, Zhiqi Huang, Vojtěch Hudeček, Pere-Lluís Huguet Cabot, Kai Hui, Chia-chien Hung, Julie Hunter, Nikolai Ilinykh, Dmitry Ilvovsky, Michimasa Inaba, Diana Inkpen, Koji Inoue, Hayate Iso, Takumi Ito, Maor Ivgi, Kenichi Iwatsuki, Vivek Iyer, Peter Izsak, Cassandra L. Jacobs, Sarthak Jain, Masoud Jalili Sabet, Sepehr Janghorbani, Adam Jatowt, Inigo Jauregi Unanue, Ganesh Jawahar, Harsh Jhamtani, Shaoxiong Ji, Yangfeng Ji, Chengyue Jiang, Junfeng Jiang, Longquan Jiang, Ming Jiang, Yuchen Eleanor Jiang, Ziyan Jiang, Baoyu Jing, Unso Jo, Richard Johansson, Aditya Joshi, Rishabh Joshi, Taehee Jung, Besim Kabashi, Sylvain Kahane, Mihir Kale, Laura Kallmeyer, Ehsan Kamalloo, Hidetaka Kamigaito, Jaap Kamps, Lis Kanashiro Pereira, Hiroshi Kanayama, Yoshinobu Kano, Diptesh Kanojia, Sudipta Kar, Georgi Karadzhov, Elena Karagjosova, Mladen Karan, Sarvnaz Karimi, Börje Karlsson, Sanjeev Kumar Karn, Constantinos Karouzos, Pradeep Karturi, Zdeněk Kasner, Yoshihide Kato, Uri Katz, Yoav Katz, Divyansh Kaushik, Pride Kavumba, Daisuke Kawahara, Gary Kazantsev, Ashkan Kazemi, Yova Kementchedjhieva, Muhammad Khalifa, Abdul Khan, Sopan Khosla, Halil Kilicoglu, Gyuwan Kim, Hyunwoo Kim, Jonggu Kim, Joo-kyung Kim, Miyoung Kim, Seungone Kim, Sungdong Kim, Young Jin Kim, Youngbin Kim, David King, Tracy Holloway King, Svetlana Kiritchenko, Jan-christoph Klie, Julien Kloetzer, René Knaebel, Sang-ki Ko, Thomas Kober, Elena Kochkina, Konstantinos Kogkalidis, Mare Koit, Thomas Kollar, Alexander Koller, Mamoru Komachi, Rik Koncel-kedziorski, Grzegorz Kondrak, Sai Koneru, Deguang Kong, Miloslav Konopík, Yannis Korkontzelos, Katerina Korre, Fajri Koto, Alexander Kotov, Mahnaz Koupaee, Venelin Kovatchev, Pavel Kral, Lea Krause, Kalpesh Krishna, Mateusz Krubiński, Canasai Kruengkrai, Jaap Kruijt, Ruben Kruiper, Sicong Kuang, Mayank Kulkarni, Deepak Kumar, Sachin Kumar, Shankar Kumar, Olli Kuparinen, Robin Kurtz, Andrey Kutuzov, Haewoon Kwak, Gorka Labaka, Sofie Labat, Faisal Ladhak, Cheng-i Lai, Tuan Lai, Wen Lai, Vasileios Lampos, Gerasimos Lampouras, Lukas Lange, Ekaterina Lapshinova-koltunski, Stefan Larson, Mark Last, Alexandra Lavrentovich, Hoang-quynh Le, Hung Le, Phong Le, Joseph Le Roux, Kevin Leach, Dong-ho Lee, Grandee Lee, Ji-ung Lee, John Lee, Lung-hao Lee, Nayeon Lee, Roy Kawei Lee, Els Lefever, Wenqiang Lei, Jochen Leidner, Heather Lent, Ran Levy, Bei Li, Bryan Li, Changmao Li, Cheng Li, Dingcheng Li, Dongfang Li, Jiacheng Li, Jialu Li, Jiazhao Li, Jing Li, Jiyi Li, Juan Li, Lei Li, Liunian Harold Li, Maoxi Li, Miao Li, Peifeng Li, Sheng Li, Shiyang Li, Shuyang Li, Siheng Li, Wei Li, Wei Li, Weikang Li, Wenyan Li, Xiangju Li, Xiaodi Li, Xue Li, Yanran Li, Yanzeng Li, Yaoyiran Li, Yizhi Li, Yongbin Li, Yue Li, Yuncong Li, Zhuang Li, Zichao Li, Chao-chun Liang, Xinnian Liang, Yueqing Liang, Baohao Liao, Jindřich Libovický, Constantine Lignos, Gilbert Lim, Kwan Hui Lim, Tomasz Limisiewicz, Lucy Lin, Weizhe Lin, Zhenxi Lin, Nedim Lipka, Pierre Lison, Shir Lissak, Danni Liu, Fangyu Liu, Fenglin Liu, Hui Liu, Jiangming Liu, Kang Liu, Lei Liu, Nayu Liu, Nelson F. Liu, Tianyu Liu, Tianyu Liu, Ting Liu, Yang Janet Liu, Yiyi Liu, Yonghui Liu, Yongkang Liu, Yue Liu, Zihan Liu, Zitao Liu, Zoey Liu, Nikola Ljubešić, Sharid Loaiciga, Colin Lockard, Pintu Lohar, Yunfei Long, Oier Lopez De Lacalle, Jaime Lorenzo-trueba, Daniel Loureiro, Junru Lu, Keming Lu, Xing Han Lu, Yanbin Lu, Yao Lu, Yujie Lu, Nurul Lubis, Jiaming Luo, Man Luo, Alex Luu, Haoran Lv, Shangwen Lv, Teresa Lynn, Meryem M'hamdi, Jie Ma, Jing Ma, Long-long Ma, Mingyu Derek Ma, Xiaofei Ma, Andrew Mackey, Aman Madaan, Avinash Madasu, Mounica Maddela, Manuel Mager, Bernardo Magnini, Adyasha Maharana, Quan Mai, Frederic Mailhot, Jean Maillard, Peter Makarov, Aaron Maladry, Ankur Mali, Anton Malko, Jonathan Mallinson, Eric Malmi, Valentin Malykh, Ramesh Manuvinakurike, Vladislav Maraev, Ana Marasovic, David Mareček, Katerina Margatina, Katja Markert, Edison Marrese-taylor, Federico Martelli, Louis Martin, Héctor Martínez Alonso, Claudia Marzi, Sarah Masud, Sandeep Mathias, Prashant Mathur, Diana Maynard, Sahisnu Mazumder, Alessandro Mazzei, R. Thomas Mccoy, John P. Mccrae, Bridget Mcinnes, Nick Mckenna, Nikhil Mehta, Fanchao Meng, Yan Meng, Zhao Meng, Orfeas Menis Mastromichalakis, Elena Merdjanovska, Eleni Metheniti, Ivan Vladimir Meza Ruiz, Paul Michel, Timothee Mickus, Stuart Middleton, Aristides Milios, Tristan Miller, David Mimno, Erxue Min, Seyedabolghasem Mirroshandel, Paramita Mirza, Abhijit Mishra, Kanishka Misra, Yusuke Miyao, Ashutosh Modi, Alireza Mohammadshahi, Hosein Mohebbi, Afroz Mohiuddin, Diego Molla, Manuel Montes, Mehrad Moradshahi, Roser Morante, Jose G. Moreno, Alejandro Moreo, Marius Mosbach, Pablo Mosteiro, Lili Mou, Diego Moussallem, Maximilian Mozes, Emir Munoz, Dragos Munteanu, Rudra Murthy, Alberto Muñoz-ortiz, Mathias Müller, Dawn Nafus, Masaaki Nagata, Saeed Najafi, Tetsuji Nakagawa, Yuta Nakashima, Diane Napolitano, Jason Naradowsky, Vivi Nastase, Anmol Nayak, Ambreen Nazir, Ani Nenkova, Mariana Neves, Jun-ping Ng, Raymond Ng, Vincent Ng, Axel-cyrille Ngonga Ngomo, Dat Quoc Nguyen, Kiet Nguyen, Nhung Nguyen, Quoc-an Nguyen, Trung Hieu Nguyen, Vincent Nguyen, Xuanfan Ni, Garrett Nicolai, Massimo Nicosia, Feng Nie, Yixin Nie, Jan Niehues, Mitja Nikolaus, Giannis Nikolentzos, Takashi Ninomiya, Kosuke Nishida, Sergiu Nisioi, Gibson Nkhata, Tadashi Nomoto, Aurélie Névéol, Alexander O'connor, Tim Oates, Kemal Oflazer, Shu Okabe, Naoaki Okazaki, Tsuyoshi Okita, Oleg Okun, Eda Okur, Antoni Oliver, Mattia Opper, Abigail Oppong, Brian Ore, Hadas Orgad, Maite Oronoz, Petya Osenova, Jessica Ouyang, Teresa Paccosi, Ankur Padia, Aishwarya Padmakumar, Shramay Palta, Tuğba Pamay Arslan, Mugdha Pandya, Wei Pang, Pinelopi Papalampidi, Nikos Papasarantopoulos, Sara Papi, Emerson Paraiso, Ashwin Paranjape, Letitia Parcalabescu, Thiago Pardo, Antonio Pareja-lora, Chanjun Park, Jong Park, Sungkyu Park, Alicia Parrish, Tommaso Pasini, Clemente Pasti, Braja Gopal Patra, Viviana Patti, Debjit Paul, Indraneil Paul, Sachin Pawar, Sarah Payne, Pavel Pecina, Jiaxin Pei, Weiping Pei, Stephan Peitz, Baolin Peng, Bo Peng, Hao Peng, Qiyao Peng, Wei Peng, Juan Antonio Perez-ortiz, Charith Peris, Ben Peters, Matthew Peters, Eva Pettersson, Thang Pham, Scott Piao, Maciej Piasecki, Massimo Piccardi, Matúš Pikuliak, Nisha Pillai, Telmo Pires, Flammie Pirinen, Benjamin Piwowarski, Flor Miriam Plaza-del-arco, Brian Plüss, Massimo Poesio, Simone Paolo Ponzetto, Octavian Popescu, Amir Pouran Ben Veyseh, Karan Praharaj, Piotr Przybyła, Stephen Pulman, Juan Manuel Pérez, Ehsan Qasemi, Hongjin Qian, Kun Qian, Kechen Qin, Jielin Qiu, Ariadna Quattoni, Ella Rabinovich, Muhammad Rahman, Sunny Rai, Vyas Raina, Sara Rajaee, Ori Ram, Taraka Rama, Giulia Rambelli, Abhinav Ramesh Kashyap, Rita Ramos, Alan Ramponi, Leonardo Ranaldi, Tharindu Ranasinghe, Surangika Ranathunga, Priya Rani, Ahmad Rashid, Pushpendre Rastogi, David Rau, Vikas Raunak, Eran Raveh, Shauli Ravfogel, Soumya Ray, Evgeniia Razumovskaia, Hanumant Redkar, Georg Rehm, Ricardo Rei, Machel Reid, Navid Rekabsaz, Ricardo Ribeiro, Giuseppe Riccardi, German Rigau, Matīss Rikters, Tharathorn Rimchala, Laura Rimell, Fabio Rinaldi, Ruty Rinott, Anthony Rios, Lina M. Rojas Barahona, Subendhu Rongali, Michael Rosner, Michael Roth, Guy Rotman, Bryan Routledge, Marco Rovera, Soumyadeep Roy, Yu-ping Ruan, Koustav Rudra, Federico Ruggeri, Irene Russo, Phillip Rust, Max Ryabinin, Maria Ryskina, Egil Rønningstad, Susanna Rücker, Malliga S, Kogilavani S V, Kenji Sagae, Keisuke Sakaguchi, Ander Salaberria, Shailaja Keyur Sampat, David Samuel, Ramon Sanabria, George Sanchez, Hugo Sanjurjo-gonzález, Sonal Sannigrahi, Rodrigo Santos, Naomi Saphra, Ruhi Sarikaya, Anoop Sarkar, Felix Sasaki, Ryohei Sasano, Nishanth Sastry, Danielle Saunders, Thiusius Savarimuthu, Beatrice Savoldi, Apoorv Saxena, Federico Scafoglieri, Andreas Scherbakov, Dominik Schlechtweg, Jonathan Schler, Michael Sejr Schlichtkrull, Robin Schmidt, Nathan Schneider, Stephanie Schoch, Annika Marie Schoene, Merel Scholman, Sabine Schulte Im Walde, Philip Schulz, Stefan Schweter, Anastasiia Sedova, Elad Segal, Cory Shain, Guokan Shang, Yutong Shao, Ori Shapira, Matthew Shardlow, Shuaijie She, Artem Shelmanov, Aili Shen, Lingfeng Shen, Xiaoyu Shen, Yuming Shen, Michael Sheng, Qiang Sheng, Tom Sherborne, Freda Shi, Zhan Shi, Zhengxiang Shi, Tomohide Shibata, Yutaro Shigeto, Takahiro Shinozaki, Kumar Shridhar, Akshat Shrivastava, Kai Shu, Raphael Shu, Anna Shvets, Anthony Sicilia, Alejandro Sierra-múnera, João Ricardo Silva, Danilo Silva De Carvalho, Patrick Simianer, Edwin Simpson, Mayank Singh, Pranaydeep Singh, Koustuv Sinha, Sunayana Sitaram, Milena Slavcheva, Kevin Small, Marco Antonio Sobrevilla Cabezudo, Swapna Somasundaran, Kai Song, Linfeng Song, Wei Song, Yan Song, Alexey Sorokin, Xabier Soto, Sajad Sotudeh, Andreas Spitz, Ivan Srba, Makesh Narsimhan Sreedhar, Hiranmai Sri Adibhatla, Balaji Vasan Srinivasan, Miloš Stanojević, Gabriel Stanovsky, Katherine Stasaski, Dario Stojanovski, Alessandro Stolfo, Tomek Strzalkowski, Dan Su, Katsuhito Sudoh, Yoshi Suhara, Alane Suhr, Changzhi Sun, Chenkai Sun, Jian Sun, Ming Sun, Qingfeng Sun, Zewei Sun, Megha Sundriyal, Hanna Suominen, Colin Swaelens, Sandesh Swamy, Vinitra Swamy, Piotr Szymański, Danae Sánchez Villegas, Víctor M. Sánchez-cartagena, Felipe Sánchez-martínez, Santosh T.y.s.s, Sho Takase, Zeerak Talat, George Tambouratzis, Fabio Tamburini, Akihiro Tamura, Chenhao Tan, Fei Tan, Xingwei Tan, Liyan Tang, Raphael Tang, Shuai Tang, Xuting Tang, Yuka Tateisi, Marta Tatu, Selma Tekir, Serra Sinem Tekiroğlu, Irina Temnikova, Daniela Teodorescu, Urmish Thakker, Mokanarangan Thayaparan, Anton Thielmann, Brian Thompson, Craig Thomson, Camilo Thorne, Tristan Thrush, Jörg Tiedemann, Refael Tikochinski, Erik Tjong Kim Sang, Evgeniia Tokarchuk, Takenobu Tokunaga, Nadi Tomeh, Marc Tomlinson, Atnafu Lambebo Tonja, Samia Touileb, Marcos Treviso, Chen-tse Tsai, Adam Tsakalidis, Yu-hsiang Tseng, Yuen-hsien Tseng, Eleftheria Tsipidi, Don Tuggener, Martin Tutek, Kiyotaka Uchimoto, Dennis Ulmer, Kanimozhi Uma, Prajna Upadhyay, Masao Utiyama, Sowmya Vajjala, Marco Valentino, Antal Van Den Bosch, Daan Van Esch, Carel Van Niekerk, Vincent Vandeghinste, Keith Vanderlinden, Lindsey Vanderlyn, Natalia Vanetik, Rossella Varvara, Shikhar Vashishth, Eva Maria Vecchi, Giulia Venturi, Rakesh Verma, Rohil Verma, Giorgos Vernikos, David Vilar, Serena Villata, Esau Villatoro-tello, Juraj Vladika, Piek Vossen, Thuy Vu, Xuan-son Vu, Ekaterina Vylomova, Tomasz Walkowiak, Yu Wan, Chuan-ju Wang, Fei Wang, Hai Wang, Haoyu Wang, Hong Wang, Jianzong Wang, Jiayi Wang, Jin Wang, Jing Wang, Kaifu Wang, Liang Wang, Lingzhi Wang, Longshaokan Wang, Longyue Wang, Miaosen Wang, Ping Wang, Qingyun Wang, Shun Wang, Wei Wang, Weichao Wang, Xin Wang, Xing Wang, Xinyi Wang, Xu Wang, Yasheng Wang, Yining Wang, Zhaowei Wang, Zhilin Wang, Zhiruo Wang, Prashan Wanigasekara, Moshe Wasserblat, Shinji Watanabe, Lucas Weber, Anna Wegmann, Jerry Wei, Wei Wei, Benjamin Weiss, Gail Weiss, Leonie Weissweiler, Charles Welch, Rongxiang Weng, Aaron White, John Wieting, Gijs Wijnholds, Adina Williams, Miles Williams, Steven Wilson, Genta Winata, Guillaume Wisniewski, Seungpil Won, Ka Ho Wong, Alina Wróblewska, Di Wu, Fangzhao Wu, Minghao Wu, Stephen Wu, Winston Wu, Xianchao Wu, Xiaofeng Wu, Xixin Wu, Yuxiang Wu, Joern Wuebker, Amelie Wührl, Min Xiao, Yuqing Xie, Zhenchang Xing, Chao Xiong, Ying Xiong, Lv Xiucheng, Dongkuan Xu, Fangyuan Xu, Hanzi Xu, Haotian Xu, Hongfei Xu, Jia Xu, Jinan Xu, Qiongkai Xu, Ruifeng Xu, Silei Xu, Xinnuo Xu, Yueshen Xu, Zhen Xu, Huiyin Xue, Linting Xue, Christos Xypolopoulos, Ivan Yamshchikov, An Yan, Ming Yan, Xi Yan, Xifeng Yan, Bohao Yang, Hao Yang, Hsiu-yu Yang, Linyi Yang, Longfei Yang, Shiquan Yang, Tao Yang, Xianjun Yang, Ze Yang, Roman Yangarber, Ken Yano, Tae Yano, Wenlin Yao, Fanghua Ye, Asaf Yehudai, Wen-wai Yim, Seid Muhie Yimam, Congchi Yin, Seunghyun Yoon, Soyoung Yoon, Ori Yoran, Naoki Yoshinaga, Chenyu You, Steve Young, Bei Yu, Juntao Yu, Kai Yu, Pengfei Yu, Shoubin Yu, Tiezheng Yu, Xiaodong Yu, Xinyan Yu, Yanchao Yu, Jianhua Yuan, Shuzhou Yuan, Frances Yung, Olga Zamaraeva, Daoguang Zan, Fabio Massimo Zanzotto, Alessandra Zarcone, Xingshan Zeng, Torsten Zesch, Shuang (sophie) Zhai, Haolan Zhan, Biao Zhang, Bowen Zhang, Ge Zhang, Haodi Zhang, Haopeng Zhang, Jason Zhang, Jianguo Zhang, Lei Zhang, Michael Zhang, Ruiyi Zhang, Sheng Zhang, Shiyue Zhang, Tianchi Zhang, Yanzhe Zhang, Yichi Zhang, Yu Zhang, Yuan Zhang, Yuhui Zhang, Zhirui Zhang, Zhisong Zhang, Hai Zhao, Jinming Zhao, Lin Zhao, Mengjie Zhao, Qinghua Zhao, Tiancheng Zhao, Xiaoyan Zhao, Yilun Zhao, Chujie Zheng, Yinhe Zheng, Alisa Zhila, Yang Zhong, Ben Zhou, Guangyou Zhou, Junpei Zhou, Kaitlyn Zhou, Wangchunshu Zhou, Xiang Zhou, Yichu Zhou, Yue Zhou, Zhengyu Zhou, Su Zhu, Wanrong Zhu, Wanzheng Zhu, Xuan Zhu, Yuan Zhuang, Claus Zinn, Yftah Ziser, Michael Zock, Bowei Zou, Wei Zou, Vilém Zouhar

**Outstanding Reviewers**

Gavin Abercrombie, Sallam Abualhaija, Yamen Ajjour, Emily Allaway, Milad Alshomary, Talita Anthonio, Lauriane Aufrant, Gorka Azkune, Lisa Beinborn, Valeriia Bolotova-baranova, Michele Cafagna, Deng Cai, Giovanni Cassani, Hanjie Chen, Cheng-han Chiang, Trevor Cohn, Karel D'oosterlinck, Jay Deyoung, Frank Drewes, Markus Dreyer, Tobias Falke, Yimai Fang, Xiaocheng Feng, Olivier Ferret, Antske Fokkens, Saadia Gabriel, Atticus Geiger, Tomas Goldsack, Konstantin Golobokov, Colin Gordon, Liane Guillou, Meiqi Guo, Nitish Gupta, William Havard, Michael Heck, Sophie Henning, Nora Hollenstein, Radu Tudor Ionescu, Tatsuya Ishigaki, Robin Jia, Min-yen Kan, Graham Katz, Christo Kirov, Ioannis Konstas, Michael Kranzlein, Udo Kruschwitz, Roland Kuhn, Yi-an Lai, Young-suk Lee, Yves Lepage, Piyawat Lertvittayakumjorn, Matthias Lindemann, Zhengyuan Liu, Henrique Lopes Cardoso, Brielen Madureira, Yuval Marton, Jonathan May, Kathleen Mckeown, Clara Meister, Zaiqiao Meng, Filip Miletic, Kata Naszadi, Yasumasa Onoe, Juri Opitz, Tiago Pimentel, Barbara Plank, Traian Rebedea, Ehud Reiter, Mathieu Roche, Rudolf Rosa, Candace Ross, Sumegh Roychowdhury, Sebastian Ruder, Elizabeth Salesky, David Schlangen, Hendrik Schuff, Sebastian Schuster, Djamé Seddah, Mattia Setzu, Kyle Shaffer, Vered Shwartz, Olivier Siohan, Matthew Stone, Alessandro Suglia, Benjamin Van Durme, Neeraj Varshney, Jake Vasilakes, Dirk Väth, Henning Wachsmuth, Michael Wiegand, Tomer Wolfson, Hanqi Yan, Eugene Yang, Marcely Zanon Boito, Amir Zeldes

# 2

# Anti-Harassment Policy

EACL 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://www.aclweb.org/adminwiki/index.php?title=Anti-Harassment_Policy`

# 3

# Meal Info

**Breakfast**: Breakfast is included for all hotel delegates who reserved their hotel room through the EACL room block, Please check in with the hotel for the location and hours of service. If you booked your room outside the block we can not guarantee breakfast and you will need to check your reservation if breakfast is included, the EACL will not be providing breakfast for delegates booked outside the block.

**Break**: Tea, coffee, and light snacks will be provided early morning, mid-morning and mid-afternoon and will be found in the Regency Ballroom on level 7

**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.

**Welcome Reception**: The EACL is pleased to invite all Main Conference attendees to the Welcome Monday May 1, 2023 held on the reception level patio. Lite Hors d'oeuvres will be served. You'll receive your admission ticket at check in as well as a drink ticket. No admission without this entry ticket.

**Dinner**: Dinner will only be provided on Wednesday Evening May 3, 2023 at the Social Event held at the Valamar Presidential Hotel Seaside Landing. All other Dinners are not provided.

# 4

# Social Events

**Social Event - May 3rd, 2023**
Venue: **Valamar President Hotel**
Time: **18:30–22.00**

We have planned the following social events for EACL 2023. Please follow ACL's code of conduct when you are attending these events. Social Event - Wednesday, May 3rd, 2023 Venue: Valamar President Hotel Located at the Seaside Landing: 18.30 - 22.00 One Entrance ticket will be included with each main conference registration so make sure to place it in your wallet, your ticket is your entrance in, no ticket no entrance. You can purchase additional tickets by modifying your registration and adding a second ticket.

**Location**: Valamar President Hotel: Ul. Iva Dulčića 142, 20000, Dubrovnik, Croatia

**Getting there**: Walking distance from the main conference there will be directional signs to get you there.

**Schedule**: 18:30–22:00 Buffet Dinner & Cash Bar each attendee will receive a drink ticket when they check in.

**Entertainment** (19:00–9:30): Kolo Dancers (traditional folk dance) continues to be the centre of village social life. The Kolo as a dance became a tool for social gathering, and was often the main occasion in which young men and women could get to know each other. DJ and Dancing to follow.

# 5

# Keynotes

## Keynote Talk: Going beyond the benefits of scale by reasoning about data

**Edward Grefenstette**
Cohere

**Tuesday, May 2, 2023** - Room: **Elafiti 1, 2, 3 & 4** - Time: **09:30-10:30**

**Abstract:** Transformer-based Large Language Models (LLMs) have taken NLP—and the world—by storm. This inflection point in our field marks a shift from focussing on domain-specific neural architecture design and the development of novel optimization techniques and objectives to a renewed focus on the scaling of model size and of the amount of data ingested during training. This paradigm shift yields surprising and delightful applications of LLMs, such as open-ended conversation, code understanding and synthesis, some degree of tool-use, and some zero-shot instruction-following capabilities. In this talk, I outline and lightly speculate on the mechanisms and properties which enable these diverse applications, and posit that the training regimen which enables these capabilities points to a further shift, namely one where we go from focussing on scale, to focussing on reasoning about what data to train on. I will briefly

discuss recent advances in open-ended learning in Reinforcement Learning, and how some of the concepts at play in that work may inspire or directly apply to the development of novel ways of reasoning about data in supervised learning, in particular in areas pertaining to LLMs.

**Bio:** Ed Grefenstette is the Head of Machine Learning at Cohere, a provider of cutting-edge NLP models that's solving all kinds of language problems; including text summarization, composition, classification and more. In addition, Ed is an Honorary Professor at UCL. Ed's previous industry experience comprises Facebook AI Research (FAIR), DeepMind, and Dark Blue Labs, where he was the CTO (acquired by Google in 2014). Prior to this, Ed worked at the University of Oxford's Department of Computer Science, and was a Fulford Junior Research Fellow at Somerville College, whilst also lecturing students at Hertford College taking Oxford's new computer science and philosophy course. Ed's research interests span several topics, including natural language and generation, machine reasoning, open ended learning, and meta-learning.

# Keynote Talk: Chatbots for Good and Evil

**Kevin Munger**
Penn State University

**Wednesday, May 3, 2023** - Room: **Elafiti 1, 2, 3 & 4** - Time: **15:45-16:45**

**Abstract:** The capacities of LLM-powered chatbots have been progressing on the order of months and have recently passed into mainstream public awareness and adoption. These tools have been used for a variety of scientific and policy interventions, but these advances call for a significant re-thinking of their place in society. Psychological research suggests that "intentionality" is a key factor in persuasion and social norm enforcement, and the proliferation of LLMs represents a significant shock to the "intentionality" contained in text and particularly in immediate, personalized chat. I argue that we are in a period of "informational disequilibrium," where different actors have different levels of awareness of this technological shock. This period may thus represent a golden age for actors aiming to use these technologies at scale, for any number of normative ends; this includes social scientists and computational linguists. More broadly, I argue that the "ethical" frameworks for evaluating research practices using LLM-powered chatbots are insufficient to the scale of the current challenge. This is a potentially revolutionary technology that requires thinking in moral and political terms: given the power imbalances involved, it is of paramount importance that chatbots for good do not inadvertently become chatbots for evil.

**Bio:** Kevin Munger is the Jeffrey L. Hyde and Sharon D. Hyde and Political Science Board of Visitors Early Career Professor of Political Science and Assistant Professor of Political Science and Social Data Analytics at Penn State University.Kevin’s research focuses on the implications of the internet and social media for the communication of political information. His speciality is the investigation of the economics of online media; current research models "Clickbait Media" and uses digital experiments to test the implications of these models on consumers of political information.

# Keynote Talk: Language Use in Embodied AI

**Joyce Chai**
University of Michigan

**Thursday, May 4, 2023** - Room: **Elafiti 1, 2, 3 & 4** - Time: **14:15-15:15**

**Abstract:** With the emergence of a new generation of embodied AI agents, it becomes increasingly important to enable language communication between humans and agents. Language plays many important roles in embodied AI. In this talk, I will share some of the experiences in my lab that study the pragmatics of language, for example, in mediating perceptual differences, learning from language instructions, and planning for joint tasks. I will talk about how the embodied context shapes language use and influences computational models for language grounding to perception and action. I will show the importance of collaborative effort and theory of mind in language communication and how they affect common ground for situated tasks. I will discuss key challenges as well as new perspectives on these problems brought by recent advances in LLM and generative AI.

**Bio:** Joyce Chai is a Professor in the Department of Electrical Engineering and Computer Science at the University of Michigan. Before joining UM in 2019, she was a Professor of Computer Science and Engineering at Michigan State University. She holds a Ph.D. in Computer Science from Duke University. Her research interests span from natural language processing and embodied AI to human-AI collaboration. She is fascinated by how experience with the world and how social pragmatics shape language learning and language use; and is excited about developing language technology that is sensorimotor grounded, pragmatically rich, and cognitively motivated. Her current work explores the intersection between language, perception, and action to enable situated communication with embodied agents. She served on the executive board of NAACL and as Program Co-Chair for multiple conferences – most recently ACL 2020. She is a recipient of the National Science Foundation Career Award and has received several paper awards with her students (e.g., the Best Long Paper Award at ACL 2010 and an Outstanding Paper Award at EMNLP 2021). She is a Fellow of ACL.

# Panel: Low-Resource Languages in NLP Products

**Mariana Romanyshyn, Antonios Anastasopoulos, Mona Diab, Julia Makogon, Ivan Vulic**

**Wednesday, May 3, 2023** - Room: **Elafiti 1, 2, 3 & 4** - Time: **16:30–18:00**

**Abstract:** The panel discussion will bring together experts from industry and academia to share their experience building solutions for low-resource languages. We anticipate a lively discussion about the advantages and limitations of multilingual solutions and language-specific models, the challenges of evaluating models for low-resource languages, and the level of language awareness needed in the development process. In addition, the panelists will explore ways to increase the acceptance rate of papers that target low-resource languages at *ACL conferences. We hope that the panel discussion will increase the visibility of research for low-resource languages and emphasize its relevance.

**Moderator: Mariana Romanyshyn, Grammarly**
**Bio:** Mariana Romanyshyn is an Area Tech Lead for Computational Linguistics at Grammarly, Ukraine. She has professional experience in syntactic parsing, sentiment analysis, named entity recognition, fact extraction, and text anonymization. For the last eight years, Mariana has been working on error correction and text improvement algorithms at Grammarly. Mariana is an active speaker at AI conferences, co-organizer of the yearly Grammarly CompLing Summer School, co-organizer of the UNLP workshop, struggling reformer of Ukrainian university syllabuses, and active contributor of the Lang-uk group, focused on advancements in Ukrainian NLP.

**Panelists:**

**Antonios Anastasopoulos, George Mason University**
**Bio:** Antonios Anastasopoulos is an Assistant Professor in Computer Science at George Mason University. He received his PhD in Computer Science from the University of Notre Dame and then did a postdoc at Language Technologies Institute at Carnegie Mellon University. He also holds a BSc-MSc in Electrical and Computer Engineering from the National Technical University of Athens, Greece. His research is on natural language processing with a focus on multilinguality, low-resource settings, cross-lingual learning, and endangered languages, with the ultimate goal of building language technologies for under-served communities around the world. He is currently funded by the NSF, the NEH, the US DoD, Google, Amazon, and Meta.

**Mona Diab, Meta**
**Bio:** Mona Diab is the Lead Responsible AI Research Scientist with Meta. She is also a full Professor of Computer Science at the George Washington University (on leave) where she directs the CARE4Lang NLP

Lab. Before joining Meta, she led the Lex Conversational AI project within Amazon AWS AI. Her current focus is on Responsible AI and how to operationalize it for NLP technologies. Her interests span building robust technologies for low-resource scenarios with a special interest in Arabic technologies, (mis) information propagation, computational socio-pragmatics, computational psycholinguistics, NLG evaluation metrics, language modeling, and resource creation.

**Julia Makogon, Semantrum**
**Bio:** Julia Makogon is a Lead ML/NLP Engineer at Semantrum, a Ukrainian AI company that specializes in media analytics and reputation management. She studied Applied Mathematics at DSTU, Kamyanske, Ukraine, before pursuing a career in NLP. Julia developed multiple NLP applications for media monitoring, sentiment analysis, and legal document analysis for Ukrainian and other European languages. Her expertise lies in building industry solutions with limited resources. Julia serves at the Program Committee of the Ukrainian NLP workshop and is passionate about advancing solutions for the Ukrainian language.

**Ivan Vulic, University of Cambridge**
**Bio:** Ivan Vulić is a Principal Research Associate and a Royal Society University Research Fellow in the Language Technology Lab, University of Cambridge. He is also a Senior Scientist at PolyAI. He is a member of the Steering Committee of the Centre for Human Inspired Artificial Intelligence (CHIA) at Cambridge. Ivan holds a PhD in Computer Science from KU Leuven awarded summa cum laude. In 2021 he was awarded the annual Karen Spärck Jones Award from the British Computing Society for his research contributions to NLP and Information Retrieval. His core expertise is in representation learning, cross-lingual learning, conversational AI, human language understanding, distributional, lexical, multi-modal, and knowledge-enhanced semantics in monolingual and multilingual contexts, transfer learning for enabling cross-lingual NLP applications such as conversational AI in low-resource languages, and machine learning for (cross-lingual and multilingual) NLP. He has published numerous papers at top-tier NLP and Information Retrieval conferences and journals, and his research work also resulted in several best paper awards. He serves as an area chair and regularly reviews for all major NLP and Machine Learning conferences and journals. Ivan has given numerous invited talks at academia and industry and co-organised a number of NLP conferences and workshops.

# 6

# Tutorials: Friday, May 5, 2023

## Overview

| | | |
|---|---|---|
| 08:30 - 16:30 | ***Day 1 Registration*** | |
| 09:00 - 18:00 | ***Day 1 Full Day Tutorial*** | |
| | *Tutorial 1 – Mining, Assessing, and Improving Arguments in NLP and the Social Sciences*<br>Gabriella Lapesa, Eva Maria Vecchi, Serena Villata, Henning Wachsmuth | *Elafiti 1* |
| 09:00 - 12:45 | ***Day 1 Morning Tutorial*** | |
| | *Tutorial 2 – Emotion Analysis from Texts*<br>Sanja Stajner, Roman Klinger | *Elafiti 2* |
| 14:15 - 18:00 | ***Day 1 Afternoon Tutorial*** | |
| | *Tutorial 3 – Summarization of Dialogues and Conversations At Scale*<br>Diyi Yang, Chenguang Zhu | *Elafiti 2* |

# 7

# Tutorials: Saturday, May 6, 2023

## Overview

| | | |
|---|---|---|
| 08:30 - 16:30 | ***Day 2 Registration*** | |
| 09:00 - 12:45 | ***Day 2 Morning Tutorial*** | |
| | *Tutorial 4 – Understanding Ethics in NLP Authoring and Reviewing*<br>Luciana Benotti, Karën Fort, Min-Yen Kan, Yulia Tsvetkov | *Elafiti 2* |
| 14:15 - 18:00 | ***Day 2 Afternoon Tutorials*** | |
| | *Tutorial 5 – AutoML for NLP*<br>Kevin Duh, Xuan Zhang | *Elafiti 1* |
| | *Tutorial 6 – Tutorial on Privacy-Preserving Natural Language Processing*<br>Ivan Habernal, Fatemehsadat Mireshghallah, Patricia Thaine, Sepideh Ghanavati, Oluwaseyi Feyisetan | *Elafiti 2* |

# Message from the Tutorial Chairs

Welcome to the Tutorials Session of EACL 2022.

NLP is a rapidly-changing field, which has undergone different periods, and the knowledge needed to be at pace is changing rapidly. The EACL tutorial session is organized to give conference attendees an 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: ACL, EACL, and EMNLP.
We would like to thank the tutorial authors for their contributions and flexibility while organizing the conference in the hybrid mode.

We hope you enjoy the tutorials while Understanding Ethics, Preserving Privacy, and Analyzing Emotions in NLP and while Summarizing Dialogues, Mining Arguments, and Learning AutoML.

Fabio Massimo Zanzotto
Sameer Pradhan
EACL 2022 Tutorial Co-chairs

# T1 - Mining, Assessing, and Improving Arguments in NLP and the Social Sciences

**Gabriella Lapesa, Eva Maria Vecchi, Serena Villata, Henning Wachsmuth**

Friday, May 5, 2023 - 9:00-18:00 (Elafiti 1)

`https://sites.google.com/view/argmintutorialeacl2023/home-page`

Computational argumentation is an interdisciplinary research field, connecting Natural Language Processing (NLP) to other disciplines such as the social sciences. This tutorial will focus on a task that recently got into the center of attention in the community: argument quality assessment, that is, what makes an argument good or bad? We structure the tutorial along three main coordinates: (1) the notions of argument quality across disciplines (how do we recognize good and bad arguments?), (2) the modeling of subjectivity (who argues to whom; what are their beliefs?), and (3) the generation of improved arguments (what makes an argument better?). The tutorial highlights interdisciplinary aspects of the field, ranging from the collaboration of theory and practice (e.g., in NLP and social sciences), to approaching different types of linguistic structures (e.g., social media versus parliamentary texts), and facing the ethical issues involved (e.g., how to build applications for the social good). A key feature of this tutorial is its interactive nature: We will involve the participants in two annotation studies on the assessment and the improvement of quality, and we will encourage them to reflect on the challenges and potential of these tasks.

---

**Gabriella Lapesa**, University of Stuttgart, Institute for Natural Language Processing
Gabriella Lapesa leads the research group E-DELIB (Powering-up E-DELIBeration: towards AI-supported moderation) at the Institute for Natural Language Processing, University of Stuttgart. Her group works at the intersection between NLP (AM) and social science (Deliberative Theory) to develop methods and tools to support moderators in deliberative discussion. As a research associate in the project MARDY (Modeling ARgumentation Dynamics in Political Discourse, University of Stuttgart and Bremen), she works on NLP methods to scale-up the analysis policy debates in multiple textual sources (i.e., who claims what in the debate on immigration or Covid-19?). Gabriella has co-chaired the 9th Argument Mining workshop (2022). With Eva Maria Vecchi, she co-taught a course on interdisciplinary AM at ESSLLI 2022.

**Eva Maria Vecchi**, University of Stuttgart, Institute for Natural Language Processing
Eva Maria Vecchi has a background in linguistics and mathematics and holds a Ph.D. degree in cognitive and neurosciences. She is a postdoctoral researcher at the Institute for Natural Language Processing at IMS Stuttgart, working on the E-DELIB project. Her focus is on the interdisciplinary effort between NLP techniques for argument mining (AM) and theories in the social sciences with the goal of a more collaborative, productive, and ethical endeavor for e-Deliberation. She has taught courses and tutorials on AM and other topics, most recently with Gabriella Lapesa at ESSLLI 2022. Her current research aims at a better understanding of the role bias has in computational argumentation and e-Deliberation, particularly the impact it has on the models, implementation, and social aspects of computational argumentation.

**Serena Villata**, Université Côte d’Azur, Inria, CNRS, I3S
Serena Villata is a research director in computer science at CNRS, and she pursues her research at the I3S laboratory in Sophia Antipolis (France). Her research area is computational argumentation, with a focus

on legal and medical texts, political debates and social network harmful content (abusive language, disinformation). Her work conjugates argument-based reasoning frameworks with natural language arguments extracted from text. She is the author of over 150 scientific publications on the topic. She holds a Chair of the Interdisciplinary Institute for AI 3IA Côte d'Azur on "Artificial Argumentation for Humans". Serena has co-chaired the 7th Workshop on Argument Mining at COLING 2020. She has also given tutorials on Argument Mining at ESSLLI 2017 and IJCAI 2016.

**Henning Wachsmuth**, Leibniz University Hannover, Institute of Artificial Intelligences
Henning Wachsmuth is the head of the Natural Language Processing Group at Leibniz University Hannover. He is an internationally leading researcher on computational argumentation with more than 60 publications on the topic, many at major NLP and AI venues. Other interests include social bias mitigation, computational reframing, and explainable NLP. Henning has co-chaired the 6th Workshop on Argument Mining at ACL 2019, and has given tutorials on argumentation at ASIRF 2018 (Cole and Achilles, 2019), EuroCSS 2018, KI 2019 (Benzmüller and Stuckenschmidt, 2019), and KI 2020 (Schmid et al., 2020). He is an initiator of the CLEF shared task series Touché on argument retrieval (Bondarenko et al., 2022), and co-chaired SemEval tasks on argument reasoning comprehension (Habernal et al., 2018), propaganda technique detection (Da San Martino et al., 2020), and identifying human values in arguments.

# T2 - Emotion Analysis from Texts

**Sanja Stajner, Roman Klinger**

Friday, May 5, 2023 - 9:00-12:45 (Elafiti 2)

`https://eacl2023tutorial.github.io/`

Emotion analysis in text is an area of research that encompasses a set of various natural language processing (NLP) tasks, including classification and regression settings, as well as structured prediction tasks like role labelling or stimulus detection. In this tutorial, we provide an overview of research from emotion psychology which sets the ground for choosing adequate NLP methodology, and present existing resources and classification methods used for emotion analysis in texts. We further discuss appraisal theories and how events can be interpreted regarding their presumably caused emotion and briefly introduce emotion role labelling. In addition to these technical topics, we discuss the use cases of emotion analysis in text, their societal impact, ethical considerations, as well as the main challenges in the field.

---

**Sanja Štajner**, Karlsruhe, Germany
Sanja Štajner has over 14 years of research experience across academia and industry on various psycholinguistic topics in NLP. The last four years, she has led and participated in industry-oriented projects that combined psychology and NLP focusing on sentiment analysis, emotion detection, personality modelling, and mental health assessment. Sanja served as a COLING 2018 area chair for psycholinguistics and cognitive modelling track, and an ACL 2022 demo chair. She has experience as tutorial presenter (COLING 2018, AIST 2018, RANLP 2017) for international audiences and as a lecturer at Masters and PhD levels.

**Roman Klinger**, Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
Roman Klinger is senior lecturer at Stuttgart University, where he teaches courses on Emotion Analysis since 2016 (see `https://www.emotionanalysis.de/`). He has been principal investigator on several externally funded projects with focus on emotion analysis. Roman served as senior area chair for sentiment analysis and argumentation mining at ACL 2022 and EACL 2021 and for evaluation and resources at EACL 2023. He was organizer of the WASSA workshop (on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis) in 2018, 2019, 2022, and 2023.

# T3 - Summarization of Dialogues and Conversations At Scale

**Diyi Yang, Chenguang Zhu**

Friday, May 5, 2023 - 14:15-18:00 (Elafiti 1)

Conversations are the natural communication format for people. This fact has motivated the large body of question answering and chatbot research as a seamless way for people to interact with machines. The conversations between people however, captured as video, audio or private or public written conversations, largely remain untapped as a source of compelling starting point for developing language technology. Summarizing such conversations can be enormously beneficial: automatic minutes for meetings or meeting highlights sent to relevant people can optimize communication in various groups while minimizing demands on people's time; similarly analysis of conversations in online support groups can provide valuable information to doctors about the patient concerns.

Summarizing written and spoken conversation poses unique research challenges—text reformulation, discourse and meaning analysis beyond the sentence, collecting data, and proper evaluation metrics. All these have been revisited by researchers since the emergence of neural approaches as the dominant approach for solving language processing problems. In this tutorial, we will survey the cutting-edge methods for summarization of conversations, covering key sub-areas whose combination is needed for a successful solution.

**Diyi Yang**, Computer Science Department, Stanford University
Diyi Yang is an assistant professor in the Computer Science Department at Stanford University. Her research focuses on dialogue summarization, learning with limited and noisy text data, user-centric language generation, and computational social science. Diyi has organized four workshops at NLP conferences: Widening NLP Workshops at NAACL 2018 and ACL 2019, Casual Inference workshop at EMNLP 2021, and NLG Evaluation workshop at EMNLP 2021. She also gave a tutorial at ACL 2022 on Learning with Limited Data.

**Chenguang Zhu**, Microsoft Azure Cognitive, Services Research
Chenguang Zhu is a Principal Research Manager in Microsoft Azure Cognitive Services Research Group, where he leads the Knowledge & Language Team. His research in NLP covers text summarization, task-oriented dialogue and knowledge graph,. Dr. Zhu has led teams to achieve first places in multiple NLP competitions. He holds a Ph.D. degree in Computer Science from Stanford University. Dr. Zhu has given talks at Stanford University, Carnegie Mellon University and UC Berkeley. He has given tutorials on Knowledge- Augmented Methods for Natural Language Processing at ACL 2022 and WSDM 2023. He is also the main organizer of The Workshop on Knowledge Augmented Methods for NLP at AAAI 2023.

# T4 - Understanding Ethics in NLP Authoring and Reviewing

**Luciana Benotti, Karën Fort, Min-Yen Kan, Yulia Tsvetkov**

Saturday, May 6, 2023 - 9:00-12:45 (Elafiti 2)

With NLP research now quickly being transferred into real-world applications, it is important to be aware of and think through the consequences of our scientific investigation. Such ethical considerations are important in both authoring and reviewing. This tutorial will equip participants with basic guidelines for thinking deeply about ethical issues and review common considerations that recur in NLP research. The methodology is interactive and participatory, including case studies and working in groups. Importantly, the participants will be co-building the tutorial outcomes and will be working to create further tutorial materials to share as public outcomes.

**Luciana Benotti**, Universidad Nacional de Córdoba, Argentina
Luciana Benotti (`luciana.benotti@unc.edu.ar`, she/her) is an Associate Professor at the Universidad Nacional de Córdoba, in Argentina. Her research interests cover many aspects of situated and grounded language, including the study of misunderstandings, bias, stereotypes, and clarification requests. She is the elected chair of the NAACL executive board and is also serving as a member at large of the ACL Ethics committee.

**Karën Fort**, Sorbonne Université
Karën Fort (`karen.fort@sorbonne-universite.fr`, she/her) is an Associate Professor at Sorbonne Université and does her research at LORIA in Nancy, France. She has been working on ethics in NLP since 2014. She was co-chair of the first two ethics committees in the field (EMNLP 2020 and NAACL 2021) and is co-chair of the ACL ethics committee. She has been a member of the Sorbonne IRB between 2019 and 2022 and she teaches ethics at undergraduate and graduate level in Paris, Nancy, and the University of Malta.

**Min-Yen Kan**, National University of Singapore
Min-Yen Kan (`kanmy@comp.nus.edu.sg`, he/him): Associate Professor at the National University of Singapore and a co-chair of the ACL Ethics Committee. He has taught over 5,000 graduate and undergraduate students on his research interests in digital libraries, information retrieval and natural language processing.

**Yulia Tsvetkov**, Paul G. Allen School of Computer Science and Engineering, University of Washington, USA
Yulia Tsvetkov (`yuliats@cs.washington.edu`, she/her) is an Assistant Professor at the Paul G. Allen School of Computer Science and Engineering at the University of Washington, USA. Her research focuses on computational ethics, multilingual NLP, and machine learning for NLP. She developed a course on Computational Ethics in NLP and is teaching it at both undergraduate and graduate levels since 2017, and she is a co-chair of the ACL Ethics Committee.

# T5 - AutoML for NLP

**Kevin Duh, Xuan Zhang**

Saturday, May 6, 2023 - 14:15-18:00 (Elafiti 1)

Automated Machine Learning (AutoML) is an emerging field that has potential to impact how we build models in NLP. As an umbrella term that includes topics like hyperparameter optimization and neural architecture search, AutoML has recently become mainstream at major conferences such as NeurIPS, ICML, and ICLR.
What does this mean to NLP? Currently, models are often built in an ad hoc process: we might borrow default hyperparameters from previous work and try a few variant architectures, but it is never guaranteed that final trained model is optimal. Automation can introduce rigor in this model-building process. This tutorial will summarize the main AutoML techniques and illustrate how to apply them to improve the NLP model-building process.

**Kevin Duh**, Johns Hopkins University Baltimore, USA
Kevin Duh is a senior research scientist at the Johns Hopkins University Human Language Technology Center of Excellence (HLTCOE) and an assistant research professor in the Department of Computer Science. His research interests lie at the intersection of NLP and Machine Learning. He has given several conference tutorials on the topics of machine learning and machine translation at, e.g., AMTA 2022, SLTU 2018, IJCNN 2017, DL4MT Winter School 2015.

**Xuan Zhang**, Johns Hopkins University Baltimore, USA
Xuan Zhang is a Ph.D. student in the Department of Computer Science at Johns Hopkins University (JHU). She performs research in Machine Translation, with specific interests in Sign Language Translation, Hyperparameter Optimization, Curriculum Learning, and Domain Adaptation. She co-presented the AMTA 2022 tutorial on AutoML.

# T6 - Tutorial on Privacy-Preserving Natural Language Processing

**Ivan Habernal, Fatemehsadat Mireshghallah, Patricia Thaine, Sepideh Ghanavati, Oluwaseyi Feyisetan**

Saturday, May 6, 2023 - 14:15-18:00 (Elafiti 2)

This cutting-edge tutorial will help the NLP community to get familiar with current research in privacy-preserving methods. We will cover topics as diverse as membership inference, differential privacy, homomorphic encryption, or federated learning, all with typical applications to NLP. The goal is not only to draw the interest of the broader community, but also to present some typical use-cases and potential pitfalls in applying privacy-preserving methods to human language technologies.

**Ivan Habernal**, Trustworthy Human Language Technologies, Technical University of Darmstadt
Ivan Habernal is currently leading a junior independent research group at the Technical University of Darmstadt, Germany, funded ad-personam by the state of Hessen. His group entitled "Trustworthy Human Language Technologies" focuses on privacy-preserving NLP and legal argument mining, among others. He has a track of top NLP publications (h-index 19), chairing workshops and tutorials, area chairing, organizing SemEval competition, giving invited talks, and also some recent industrial experience in areas where privacy matters a lot but the tools are not ready yet (healthcare and online personalization).

**Fatemeh Mireshghallah**, Computer Science and Engineering Department, University of California San Diego
Fatemehsadat Mireshghallah is a Ph.D. student at the CSE department of UC San Diego. Her research interests are Trustworthy Machine Learning and Natural Language Processing. She received her B.S. from Sharif university of technology in Iran. She is a recipient of the National Center for Women & IT (NCWIT) Collegiate award in 2020 for her work on privacy-preserving inference, and a finalist of the Qualcomm Innovation Fellowship in 2021. She has interned twice at Microsoft Research's Language and Intelligent Assistance group, where she worked on private training of large language models. She is also serving as a NAACL 2022 D&I co-chair and WiNLP committee member.

**Patricia Thaine**, Private AI, Canada
Patricia Thaine is the Co-Founder and CEO of Private AI, a Computer Science PhD Candidate at the University of Toronto and a Postgraduate Affiliate at the Vector Institute doing research on privacy-preserving natural language processing, with a focus on applied cryptography. She also does research on computational methods for lost language decipherment. Patricia is a recipient of the NSERC Postgraduate Scholarship, the RBC Graduate Fellowship, the Beatrice 'Trixie' Worsley Graduate Scholarship in Computer Science, and the Ontario Graduate Scholarship. She has eight years of research and software development experience, including at the McGill Language Development Lab, the University of Toronto's Computational Linguistics Lab, the University of Toronto's Department of Linguistics, and the Public Health Agency of Canada. She is the Co-Founder and CEO of Private AI, the former President of the Computer Science Graduate Student Union at the University of Toronto, and a member of the Board of Directors of Equity Showcase, one of Canada's oldest not-for-profit charitable organizations.

**Sepideh Ghanavati**, School of Computing and Information Science, University of Maine
Sepideh Ghanavati is an assistant professor in Computer Science at the University of Maine. She is the director of Privacy Engineering - Regulatory Compliance Lab (PERC_Lab). Her research interests are in the areas of information privacy and security, software engineering, machine learning and the Internet of Things (IoT). Previously, she worked as an assistant professor at Texas Tech University, visiting assistant

professor at Radboud University, the Netherlands and as a visiting faculty at Carnegie Mellon University. She is the recipient of Google Faculty Research award in 2018. She has more than 10 years of academic and industry experience in the area of privacy and regulatory compliance and has published more than 30 peer-reviewed publications. She was a co-organizer of the ‘Privacy and Language Technologies’ at the 2019 AAAI Spring Symposium and has been part of the organizing committee of several workshops and conferences in the past.

**Oluwaseyi Feyisetan**, Meta, USA
Seyi is a Staff Research Scientist at Facebook. Prior to Facebook, he was a Senior Applied Scientist at Amazon where he worked on Differential Privacy in the context of NLP. He holds 4 pending patents with Amazon on preserving privacy in NLP systems. He completed his PhD at the University of Southampton in the UK and has published in top tier conferences and journals on crowdsourcing, homomorphic encryption, and privacy. He has served as a reviewer at top NLP conferences including ACL and EMNLP. Prior to Amazon, he spent 7 years in the UK where he worked at different startups and institutions focusing on regulatory compliance, machine learning and NLP within the finance sector. He also sits on the research advisory board of the IAPP.

# 8

# Main Conference

## Main Conference Program (Overview)

### Main Conference Program (Overview): Day 0

17:00-21:30 Registration (Foyer)

18:30-21:30 **Welcome Reception** (Lobby Terrace)

# Main Conference Program (Overview): Day 1

7:30 Registration (Foyer)

9:00-9:30 **Session 1: Welcome Address** (Elafiti 1-4)

9:30-10:30 **Session 2 Keynote Talk:** Edward Grefenstette (Elafiti 1-4)

10:30-11:15 Morning Break (Nocturno located in the Exhibit Center)

11:15-12:45 **Session 3:**

| | |
|---|---|
| **NLP Applications**<br>*Elafiti 2* | **Computational Social Science and Social Media**<br>*Elafiti 3* |
| **Dialogue and Interactive Systems**<br>*Elafiti 4* | **In Person Poster Session**<br>*Exhibit/Business Center* |

12:45-14:15 Lunch Break

14:15-15:45 **Session 4: Virtual Poster Sessions** (Elafiti 1-4)

15:45-16:30 Afternoon Break (Nocturno located in the Exhibit Center)

16:30-18:00 **Session 5:**

| | |
|---|---|
| **Text Classification, Sentiment Analysis, Argument Mining**<br>*Elafiti 2* | **Ethical and Sustainable NLP**<br>*Elafiti 3* |
| **Summarization and Medical Applications**<br>*Elafiti 4* | **In Person Poster Session**<br>*Exhibit/Business Center* |

## Main Conference Program (Overview): Day 2

| | | | |
|---|---|---|---|
| 8:30 | Registration (Lobby) | | |
| 9:00-10:30 | **Session 6:** | **Information Extraction**<br>*Elafiti 2* | **Generation**<br>*Elafiti 3* |
| | | **Interpretability and Model Analysis**<br>*Elafiti 4* | **In Person Poster Session**<br>*Exhibit/Business Center* |
| 10:30-11:15 | Morning Break (Nocturno located in the Exhibit Center) | | |
| 11:15-12:45 | **Session 7:** | **Language Resources and Evaluation 1**<br>*Elafiti 2* | **Machine Learning for NLP**<br>*Elafiti 3* |
| | | **Language Grounding and Multi-Modality**<br>*Elafiti 4* | **In Person Poster Session**<br>*Exhibit/Business Center* |
| 12:45-13:30 | Lunch Break | | |
| 14:00-14:45 | **Business Meeting** (Elafiti 1-4) | | |
| 14:45-15:45 | **Session 8 Keynote Talk:** Kevin Munger (Plenary Elafiti 1-4) | | |
| 15:45-16:30 | Afternoon Break (Nocturno located in the Exhibit Center) | | |
| 16:30-18:00 | **Session 9 Panel: Low-resource Languages in NLP Products** (Plenary Elafiti 1-4) | | |
| 18:30-22:00 | **Social Event:** Located at Valamar President | | |

# Main Conference Program (Overview): Day 3

8:30 Registration (Level 3 Foyer)

9:00-10:30 **Session 10:**

| | |
|---|---|
| **Machine Translation and Multilinguality**<br>*Elafiti 2* | **Lexical Semantics, Discourse and Anaphora**<br>*Elafiti 3* |
| **Language Resources and Evaluation 2**<br>*Elafiti 4* | **In Person Poster Session**<br>*Exhibit/Business Center* |

10:30-11:15 Morning Break (Nocturno located in the Exhibit Center)

11:15-12:45 **Session 11:**

| | |
|---|---|
| **Question Generation and Answering**<br>*Elafiti 2* | **Semantics: Sentence level and Other areas**<br>*Elafiti 3* |
| **Large Language Models**<br>*Elafiti 4* | **In Person Poster Session**<br>*Exhibit/Business Center* |

12:45-14:15 Lunch Break

14:15-15:45 **Session 12 Keynote Talk:** Joyce Chai (Plenary Elafiti 1-4)

15:15-16:00 Afternoon Break (Nocturno located in the Exhibit Center)

16:30-17:30 **Best Paper Awards & Closing Session** (Plenary Elafiti 1-4)

# Main Conference: Tuesday, May 2, 2023

## Parallel Session 3 - 11:15-12:45

### Session 3 Orals – Computational Social Science and Social Media – Room B

11:15-12:45 (Elafiti 3)

**Creation and evaluation of timelines for longitudinal user posts**
*Anthony Hills, Adam Tsakalidis, Federico Nanni, Ioannis Zachos and Maria Liakata* 11:15-11:30 (Elafiti 3)
There is increasing interest to work with user generated content in social media, especially textual posts over time. Currently there is no consistent way of segmenting user posts into timelines in a meaningful way that improves the quality and cost of manual annotation. Here we propose a set of methods for segmenting longitudinal user posts into timelines likely to contain interesting moments of change in a user's behaviour, based on their online posting activity. We also propose a novel framework for evaluating timelines and show its applicability in the context of two different social media datasets. Finally, we present a discussion of the linguistic content of highly ranked timelines.

**Extracting Victim Counts from Text**
*Mian Zhong, Shehzaad Dhuliawala and Niklas Stoehr* 11:30-11:45 (Elafiti 3)
Decision-makers in the humanitarian sector rely on timely and exact information during crisis events. Knowing how many civilians were injured during an earthquake is vital to allocate aids properly. Information about such victim counts are however often only available within full-text event descriptions from newspapers and other reports. Extracting numbers from text is challenging: numbers have different formats and may require numeric reasoning. This renders purely tagging approaches insufficient. As a consequence, fine-grained counts of injured, displaced, or abused victims beyond fatalities are often not extracted and remain unseen. We cast victim count extraction as a question answering (QA) task with a regression or classification objective. We compare tagging approaches: regex, dependency parsing, semantic role labeling, and advanced text-to-text models. Beyond model accuracy, we analyze extraction reliability and robustness which are key for this sensitive task. In particular, we discuss model calibration and investigate out-of-distribution and few-shot performance. Ultimately, we make a comprehensive recommendation on which model to select for different desiderata and data domains. Our work is among the first to apply numeracy-focused large language models in a real-world use case with a positive impact.

**How people talk about each other: Modeling Generalized Intergroup Bias and Emotion**
*Venkata Subrahmanyan Govindarajan, Katherine Atwell, Barea Sinno, Malihe Alikhani, David Beaver and Junyi Jessy Li* 11:45-12:00 (Elafiti 3)
Current studies of bias in NLP rely mainly on identifying (unwanted or negative) bias towards a specific demographic group. While this has led to progress recognizing and mitigating negative bias, and having a clear notion of the targeted group is necessary, it is not always practical. In this work we extrapolate to a broader notion of bias, rooted in social science and psychology literature. We move towards predicting interpersonal group relationship (IGR) - modeling the relationship between the speaker and the target in an utterance - using fine-grained interpersonal emotions as an anchor. We build and release a dataset of English tweets by US Congress members annotated for interpersonal emotion - the first of its kind, and 'found supervision' for IGR labels; our analyses show that subtle emotional signals are indicative of different biases. While humans can perform better than chance at identifying IGR given an utterance, we show that neural models perform much better; furthermore, a shared encoding between IGR and interpersonal perceived emotion enabled performance gains in both tasks.

**Multilingual Content Moderation: A Case Study on Reddit**
*Meng Ye, Karan Sikka, Katherine Atwell, Sabit Hassan, Ajay Divakaran and Malihe Alikhani* 12:00-12:15 (Elafiti 3)
Content moderation is the process of flagging content based on pre-defined platform rules. There has been a growing need for AI moderators to safeguard users as well as protect the mental health of human moderators from traumatic content. While prior works have focused on identifying hateful/offensive language, they are not adequate for meeting the challenges of content moderation since 1) moderation decisions are based on violation of rules, which subsumes detection of offensive speech, and 2) such rules often differ across communities which entails an adaptive solution. We propose to study the challenges of content moderation by introducing a multilingual dataset of 1.8 Million Reddit comments spanning 56 subreddits in English, German, Spanish and French1. We perform extensive experimental analysis to highlight the underlying challenges and suggest related research problems such as cross-lingual transfer, learning under label noise (human biases), transfer of moderation models, and predicting the violated rule. Our dataset and analysis can help better prepare for the challenges and opportunities of auto moderation.

**Lessons Learned from a Citizen Science Project for Natural Language Processing**
*Jan-christoph Klie, Ji-ung Lee, Kevin Stowe, Gözde Şahin, Nafise Sadat Moosavi, Luke Bates, Dominic Petrak, Richard Eckart De Castilho and Iryna Gurevych* 12:15-12:30 (Elafiti 3)
Many Natural Language Processing (NLP) systems use annotated corpora for training and evaluation. However, labeled data is often costly to obtain and scaling annotation projects is difficult, which is why annotation tasks are often outsourced to paid crowdworkers. Citizen Science is an alternative to crowdsourcing that is relatively unexplored in the context of NLP. To investigate whether and how well Citizen Science can be applied in this setting, we conduct an exploratory study into engaging different groups of volunteers in Citizen Science for NLP by re-annotating parts of a pre-existing crowdsourced dataset. Our results show that this can yield high-quality annotations and at- tract motivated volunteers, but also requires considering factors such as scalability, participation over time, and legal and ethical issues. We summarize lessons learned in the form of guidelines and provide our code and data to aid future work on Citizen Science.

**Generation-Based Data Augmentation for Offensive Language Detection: Is It Worth It?**
*Camilla Casula and Sara Tonelli* 12:30-12:45 (Elafiti 3)
Generation-based data augmentation (DA) has been presented in several works as a way to improve offensive language detection. However, the effectiveness of generative DA has been shown only in limited scenarios, and the potential injection of biases when using generated data to classify offensive language has not been investigated. Our aim is that of analyzing the feasibility of generative data augmentation more in-depth with two main focuses. First, we investigate the robustness of models trained on generated data in a variety of data augmentation setups, both novel and already presented in previous work, and compare their performance on four widely-used English offensive language

datasets that present inherent differences in terms of content and complexity. In addition to this, we analyze models using the HateCheck suite, a series of functional tests created to challenge hate speech detection systems. Second, we investigate potential lexical bias issues through a qualitative analysis on the generated data. We find that the potential positive impact of generative data augmentation on model performance is unreliable, and generative DA can also have unpredictable effects on lexical bias.

## Session 3 Orals – Dialogue and Interactive Systems – Room C

11:15-12:45 (Elafiti 4)

**CLICK: Contrastive Learning for Injecting Contextual Knowledge to Conversational Recommender System**
*Hyeongjun Yang, Heesoo Won, Youbin Ahn and Kyong-ho Lee* 11:15-11:30 (Elafiti 4)
Conversational recommender systems (CRSs) capture a user preference through a conversation. However, the existing CRSs lack capturing comprehensive user preferences. This is because the items mentioned in a conversation are mainly regarded as a user preference. Thus, they have limitations in identifying a user preference from a dialogue context expressed without preferred items. Inspired by the characteristic of an online recommendation community where participants identify a context of a recommendation request and then comment with appropriate items, we exploit the Reddit data. Specifically, we propose a Contrastive Learning approach for Injecting Contextual Knowledge (CLICK) from the Reddit data to the CRS task, which facilitates the capture of a context-level user preference from a dialogue context, regardless of the existence of preferred item-entities. Moreover, we devise a relevance-enhanced contrastive learning loss to consider the fine-grained reflection of multiple recommendable items. We further develop a response generation module to generate a persuasive rationale for a recommendation. Extensive experiments on the benchmark CRS dataset show the effectiveness of CLICK, achieving significant improvements over state-of-the-art methods.

**Fiction-Writing Mode: An Effective Control for Human-Machine Collaborative Writing**
*Wenjie Zhong, Jason Naradowsky, Hiroya Takamura, Ichiro Kobayashi and Yusuke Miyao* 11:30-11:45 (Elafiti 4)
We explore the idea of incorporating concepts from writing skills curricula into human-machine collaborative writing scenarios, focusing on adding writing modes as a control for text generation models. Using crowd-sourced workers, we annotate a corpus of narrative text paragraphs with writing mode labels. Classifiers trained on this data achieve an average accuracy of ~87% on held-out data. We fine-tune a set of large language models to condition on writing mode labels, and show that the generated text is recognized as belonging to the specified mode with high accuracy.
To study the ability of writing modes to provide fine-grained control over generated text, we devise a novel turn-based text reconstruction game to evaluate the difference between the generated text and the author's intention. We show that authors prefer text suggestions made by writing mode-controlled models on average 61.1% of the time, with satisfaction scores 0.5 higher on a 5-point ordinal scale. When evaluated by humans, stories generated via collaboration with writing mode-controlled models achieve high similarity with the professionally written target story. We conclude by identifying the most common mistakes found in the generated stories.

**Instruction Clarification Requests in Multimodal Collaborative Dialogue Games: Tasks, and an Analysis of the CoDraw Dataset**
*Brielen Madureira and David Schlangen* 11:45-12:00 (Elafiti 4)
In visual instruction-following dialogue games, players can engage in repair mechanisms in face of an ambiguous or underspecified instruction that cannot be fully mapped to actions in the world. In this work, we annotate Instruction Clarification Requests (iCRs) in CoDraw, an existing dataset of interactions in a multimodal collaborative dialogue game. We show that it contains lexically and semantically diverse iCRs being produced self-motivatedly by players deciding to clarify in order to solve the task successfully. With 8.8k iCRs found in 9.9k dialogues, CoDraw-iCR (v1) is a large spontaneous iCR corpus, making it a valuable resource for data-driven research on clarification in dialogue. We then formalise and provide baseline models for two tasks: Determining when to make an iCR and how to recognise them, in order to investigate to what extent these tasks are learnable from data.

**Opportunities and Challenges in Neural Dialog Tutoring**
*Jakub Macina, Nico Daheim, Lingzhi Wang, Tanmay Sinha, Manu Kapur, Iryna Gurevych and Mrinmaya Sachan* 12:00-12:15 (Elafiti 4)
Designing dialog tutors has been challenging as it involves modeling the diverse and complex pedagogical strategies employed by human tutors. Although there have been significant recent advances in neural conversational systems using large language models and growth in available dialog corpora, dialog tutoring has largely remained unaffected by these advances. In this paper, we rigorously analyze various generative language models on two dialog tutoring datasets for language learning using automatic and human evaluations to understand the new opportunities brought by these advances as well as the challenges we must overcome to build models that would be usable in real educational settings. We find that although current approaches can model tutoring in constrained learning scenarios when the number of concepts to be taught and possible teacher strategies are small, they perform poorly in less constrained scenarios. Our human quality evaluation shows that both models and ground-truth annotations exhibit low performance in terms of equitable tutoring, which measures learning opportunities for students and how engaging the dialog is. To understand the behavior of our models in a real tutoring setting, we conduct a user study using expert annotators and find a significantly large number of model reasoning errors in 45% of conversations. Finally, we connect our findings to outline future work.

**Zero and Few-Shot Localization of Task-Oriented Dialogue Agents with a Distilled Representation**
*Mehrad Moradshahi, Sina Semnani and Monica Lam* 12:15-12:30 (Elafiti 4)
Task-oriented Dialogue (ToD) agents are mostly limited to a few widely-spoken languages, mainly due to the high cost of acquiring training data for each language. Existing low-cost approaches that rely on cross-lingual embeddings or naive machine translation sacrifice a lot of accuracy for data efficiency, and largely fail in creating a usable dialogue agent. We propose automatic methods that use ToD training data in a source language to build a high-quality functioning dialogue agent in another target language that has no training data (i.e. zero-shot) or a small training set (i.e. few-shot). Unlike most prior work in cross-lingual ToD that only focuses on Dialogue State Tracking (DST), we build an end-to-end agent.
We show that our approach closes the accuracy gap between few-shot and existing full-shot methods for ToD agents. We achieve this by (1) improving the dialogue data representation, (2) improving entity-aware machine translation, and (3) automatic filtering of noisy translations. We evaluate our approach on the recent bilingual dialogue dataset BiToD. In Chinese to English transfer, in the zero-shot setting, our method achieves 46.7% and 22.0% in Task Success Rate (TSR) and Dialogue Success Rate (DSR) respectively. In the few-shot setting where 10% of the data in the target language is used, we improve the state-of-the-art by 15.2% and 14.0%, coming within 5% of full-shot training.

**The StatCan Dialogue Dataset: Retrieving Data Tables through Conversations with Genuine Intents**
*Xing Han Lu, Siva Reddy and Harm De Vries* 12:30-12:45 (Elafiti 4)
We introduce the StatCan Dialogue Dataset consisting of 19,379 conversation turns between agents working at Statistics Canada and online

users looking for published data tables. The conversations stem from genuine intents, are held in English or French, and lead to agents retrieving one of over 5000 complex data tables. Based on this dataset, we propose two tasks: (1) automatic retrieval of relevant tables based on a on-going conversation, and (2) automatic generation of appropriate agent responses at each turn. We investigate the difficulty of each task by establishing strong baselines. Our experiments on a temporal data split reveal that all models struggle to generalize to future conversations, as we observe a significant drop in performance across both tasks when we move from the validation to the test set. In addition, we find that response generation models struggle to decide when to return a table. Considering that the tasks pose significant challenges to existing models, we encourage the community to develop models for our task, which can be directly used to help knowledge workers find relevant tables for live chat users.

## Session 3 Orals – NLP Applications – Room A

11:15-12:45 (Elafiti 2)

**Don't Mess with Mister-in-Between: Improved Negative Search for Knowledge Graph Completion**
*Fan Jiang, Tom Drummond and Trevor Cohn* 11:15-11:30 (Elafiti 2)
The best methods for knowledge graph completion use a 'dual-encoding' framework, a form of neural model with a bottleneck that facilitates fast approximate search over a vast collection of candidates. These approaches are trained using contrastive learning to differentiate between known positive examples and sampled negative instances. The mechanism for sampling negatives to date has been very simple, driven by pragmatic engineering considerations (e.g., using mismatched instances from the same batch). We propose several novel means of finding more informative negatives, based on searching for candidates with high lexical overlaps, from the dual-encoder model and according to knowledge graph structures. Experimental results on four benchmarks show that our best single model improves consistently over previous methods and obtains new state-of-the-art performance, including the challenging large-scale Wikidata5M dataset. Combing different kinds of strategies through model ensembling results in a further performance boost.

**Modelling Temporal Document Sequences for Clinical ICD Coding**
*Boon Liang Clarence Ng, Diogo Santos and Marek Rei* 11:30-11:45 (Elafiti 2)
Past studies on the ICD coding problem focus on predicting clinical codes primarily based on the discharge summary. This covers only a small fraction of the notes generated during each hospital stay and leaves potential for improving performance by analysing all the available clinical notes. We propose a hierarchical transformer architecture that uses text across the entire sequence of clinical notes in each hospital stay for ICD coding, and incorporates embeddings for text metadata such as their position, time, and type of note. While using all clinical notes increases the quantity of data substantially, superconvergence can be used to reduce training costs. We evaluate the model on the MIMIC-III dataset. Our model exceeds the prior state-of-the-art when using only discharge summaries as input, and achieves further performance improvements when all clinical notes are used as input.

**Assistive Recipe Editing through Critiquing**
*Diego Antognini, Shuyang Li, Boi Faltings and Julian Mcauley* 11:45-11:55 (Elafiti 2)
There has recently been growing interest in the automatic generation of cooking recipes that satisfy some form of dietary restrictions, thanks in part to the availability of online recipe data. Prior studies have used pre-trained language models, or relied on small paired recipe data (e.g., a recipe paired with a similar one that satisfies a dietary constraint). However, pre-trained language models generate inconsistent or incoherent recipes, and paired datasets are not available at scale. We address these deficiencies with RecipeCrit, a hierarchical denoising auto-encoder that edits recipes given ingredient-level critiques. The model is trained for recipe completion to learn semantic relationships within recipes. Our work's main innovation is our unsupervised critiquing module that allows users to edit recipes by interacting with the predicted ingredients; the system iteratively rewrites recipes to satisfy users' feedback. Experiments on the Recipe1M recipe dataset show that our model can more effectively edit recipes compared to strong language-modeling baselines, creating recipes that satisfy user constraints and are more correct, serendipitous, coherent, and relevant as measured by human judges.

**Metaphor Detection with Effective Context Denoising**
*Shun Wang, Yucheng Li, Chenghua Lin, Loic Barrault and Frank Guerin* 11:55-12:05 (Elafiti 2)
We propose a novel RoBERTa-based model, RoPPT, which introduces a target-oriented parse tree structure in metaphor detection. Compared to existing models, RoPPT focuses on semantically relevant information and achieves the state-of-the-art on several main metaphor datasets. We also compare our approach against several popular denoising and pruning methods, demonstrating the effectiveness of our approach in context denoising. Our code and dataset can be found at https://github.com/MajiBear000/RoPPT.

**Friend-training: Learning from Models of Different but Related Tasks**
*Mian Zhang, Lifeng Jin, Linfeng Song, Haitao Mi, Xiabing Zhou and Dong Yu* 12:05-12:20 (Elafiti 2)
Current self-training methods such as standard self-training, co-training, tri-training, and others often focus on improving model performance on a single task, utilizing differences in input features, model architectures, and training processes. However, many tasks in natural language processing are about different but related aspects of language, and models trained for one task can be great teachers for other related tasks. In this work, we propose friend-training, a cross-task self-training framework, where models trained to do different tasks are used in an iterative training, pseudo-labeling, and retraining process to help each other for better selection of pseudo-labels. With two dialogue understanding tasks, conversational semantic role labeling and dialogue rewriting, chosen for a case study, we show that the models trained with the friend-training framework achieve the best performance compared to strong baselines.

**MTEB: Massive Text Embedding Benchmark**
*Niklas Muennighoff, Nouamane Tazi, Loic Magne and Nils Reimers* 12:20-12:35 (Elafiti 2)
Text embeddings are commonly evaluated on a small set of datasets from a single task not covering their possible applications to other tasks. It is unclear whether state-of-the-art embeddings on semantic textual similarity (STS) can be equally well applied to other tasks like clustering or reranking. This makes progress in the field difficult to track, as various models are constantly being proposed without proper evaluation. To solve this problem, we introduce the Massive Text Embedding Benchmark (MTEB). MTEB spans 8 embedding tasks covering a total of 58 datasets and 112 languages. Through the benchmarking of 33 models on MTEB, we establish the most comprehensive benchmark of text embeddings todate. We find that no particular text embedding method dominates across all tasks. This suggests that the field has yet to converge on a universal text embedding method and scale it up sufficiently to provide state-of-theart results on all embedding tasks. MTEB comes with open-source code and a public leaderboard at https://github.com/embeddings-benchmark/mteb.

## Session 3 Posters

11:15-12:45 (Exhibit Hall)

**Improving Sign Recognition with Phonology**
*Lee Kezar, Jesse Thomason and Zed Sehyr* 11:15-12:45 (Exhibit Hall)
We use insights from research on American Sign Language (ASL) phonology to train models for isolated sign language recognition (ISLR), a step towards automatic sign language understanding. Our key insight is to explicitly recognize the role of phonology in sign production to achieve more accurate ISLR than existing work which does not consider sign language phonology. We train ISLR models that take in pose estimations of a signer producing a single sign to predict not only the sign but additionally its phonological characteristics, such as the handshape. These auxiliary predictions lead to a nearly 9% absolute gain in sign recognition accuracy on the WLASL benchmark, with consistent improvements in ISLR regardless of the underlying prediction model architecture. This work has the potential to accelerate linguistic research in the domain of signed languages and reduce communication barriers between deaf and hearing people.

**Behavior Cloned Transformers are Neurosymbolic Reasoners**
*Ruoyao Wang, Peter Jansen, Marc-alexandre Cote and Prithviraj Ammanabrolu* 11:15-12:45 (Exhibit Hall)
In this work, we explore techniques for augmenting interactive agents with information from symbolic modules, much like humans use tools like calculators and GPS systems to assist with arithmetic and navigation. We test our agent's abilities in text games – challenging benchmarks for evaluating the multi-step reasoning abilities of game agents in grounded, language-based environments. Our experimental study indicates that injecting the actions from these symbolic modules into the action space of a behavior cloned transformer agent increases performance on four text game benchmarks that test arithmetic, navigation, sorting, and common sense reasoning by an average of 22%, allowing an agent to reach the highest possible performance on unseen games. This action injection technique is easily extended to new agents, environments, and symbolic modules.

**AutoTriggER: Label-Efficient and Robust Named Entity Recognition with Auxiliary Trigger Extraction**
*Dong-ho Lee, Ravi Kiran Selvam, Sheikh Muhammad Sarwar, Bill Yuchen Lin, Fred Morstatter, Jay Pujara, Elizabeth Boschee, James Allan and Xiang Ren* 11:15-12:45 (Exhibit Hall)
Deep neural models for named entity recognition (NER) have shown impressive results in overcoming label scarcity and generalizing to unseen entities by leveraging distant supervision and auxiliary information such as explanations. However, the costs of acquiring such additional information are generally prohibitive. In this paper, we present a novel two-stage framework (AutoTriggER) to improve NER performance by automatically generating and leveraging "entity triggers" which are human-readable cues in the text that help guide the model to make better decisions. Our framework leverages post-hoc explanation to generate rationales and strengthens a model's prior knowledge using an embedding interpolation technique. This approach allows models to exploit triggers to infer entity boundaries and types instead of solely memorizing the entity words themselves. Through experiments on three well-studied NER datasets, AutoTriggER shows strong label-efficiency, is capable of generalizing to unseen entities, and outperforms the RoBERTa-CRF baseline by nearly 0.5 F1 points on average.

**Incorporating Task-Specific Concept Knowledge into Script Learning**
*Chenkai Sun, Tie Xu, Chengxiang Zhai and Heng Ji* 11:15-12:45 (Exhibit Hall)
In this paper, we present Tetris, a new task of Goal-Oriented Script Completion. Unlike previous work, it considers a more realistic and general setting, where the input includes not only the goal but also additional user context, including preferences and history. To address this problem, we propose a novel approach, which uses two techniques to improve performance: (1) concept prompting, and (2) script-oriented contrastive learning that addresses step repetition and hallucination problems. On our WikiHow-based dataset, we find that both methods improve performance.

**Salient Span Masking for Temporal Understanding**
*Jeremy Cole, Aditi Chaudhary, Bhuwan Dhingra and Partha Talukdar* 11:15-12:45 (Exhibit Hall)
Salient Span Masking (SSM) has shown itself to be an effective strategy to improve closed-book question answering performance. SSM extends general masked language model pretraining by creating additional unsupervised training sentences that mask a single entity or date span, thus oversampling factual information. Despite the success of this paradigm, the span types and sampling strategies are relatively arbitrary and not widely studied for other tasks. Thus, we investigate SSM from the perspective of temporal tasks, where learning a good representation of various temporal expressions is important. To that end, we introduce Temporal Span Masking (TSM) intermediate training. First, we find that SSM alone improves the downstream performance on three temporal tasks by an avg. +5.8 points. Further, we are able to achieve additional improvements (avg. +0.29 points) by adding the TSM task. These comprise the new best reported results on the targeted tasks. Our analysis suggests that the effectiveness of SSM stems from the sentences chosen in the training data rather than the mask choice: sentences with entities frequently also contain temporal expressions. Nonetheless, the additional targeted spans of TSM can still improve performance, especially in a zero-shot context.

**BERT Shows Garden Path Effects**
*Tovah Irwin, Kyra Wilson and Alec Marantz* 11:15-12:45 (Exhibit Hall)
Garden path sentences (i.e. "the horse raced past the barn fell") are sentences that readers initially incorrectly parse, requiring partial or total re-analysis of the sentence structure. Given human difficulty in parsing garden paths, we aim to compare transformer language models' performance on these sentences. We assess a selection of models from the BERT family which have been fine-tuned on the question-answering task, and evaluate each model's performance on comprehension questions based on garden path and control sentences. We then further investigate the semantic roles assigned to arguments of verbs in garden path and control sentences by utilizing a probe task to directly assess which semantic role(s) the model assigns. We find that the models have relatively low performance in certain instances of question answering based on garden path contexts, and the model incorrectly assigns semantic roles, aligning for the most part with human performance.

**A Federated Approach for Hate Speech Detection**
*Jay Gala, Deep Gandhi, Jash Mehta and Zeerak Talat* 11:15-12:45 (Exhibit Hall)
Hate speech detection has been the subject of high research attention, due to the scale of content created on social media. In spite of the attention and the sensitive nature of the task, privacy preservation in hate speech detection has remained under-studied. The majority of research has focused on centralised machine learning infrastructures which risk leaking data. In this paper, we show that using federated machine learning can help address privacy the concerns that are inherent to hate speech detection while obtaining up to 6.81% improvement in terms of F1-score.

**A Discerning Several Thousand Judgments: GPT-3 Rates the Article + Adjective + Numeral + Noun Construction**
*Kyle Mahowald* 11:15-12:45 (Exhibit Hall)
Knowledge of syntax includes knowledge of rare, idiosyncratic constructions. LLMs must overcome frequency biases in order to master such constructions. In this study, I prompt GPT-3 to give acceptability judgments on the English-language Article + Adjective + Numeral + Noun

construction (e.g., "a lovely five days"). I validate the prompt using the CoLA corpus of acceptability judgments and then zero in on the AANN construction. I compare GPT- 3's judgments to crowdsourced human judgments on a subset of sentences. GPT-3's judgments are broadly similar to human judgments and generally align with proposed constraints in the literature but, in some cases, GPT-3's judgments and human judgments diverge from the literature and from each other.

**TraVLR: Now You See It, Now You Don't! A Bimodal Dataset for Evaluating Visio-Linguistic Reasoning**
*Keng Ji Chow, Samson Tan and Min-yen Kan* 11:15-12:45 (Exhibit Hall)
Numerous visio-linguistic (V+L) representation learning methods have been developed, yet existing datasets do not adequately evaluate the extent to which they represent visual and linguistic concepts in a unified space. We propose several novel evaluation settings for V+L models, including cross-modal transfer. Furthermore, existing V+L benchmarks often report global accuracy scores on the entire dataset, making it difficult to pinpoint the specific reasoning tasks that models fail and succeed at. We present TraVLR, a synthetic dataset comprising four V+L reasoning tasks. TraVLR's synthetic nature allows us to constrain its training and testing distributions along task-relevant dimensions, enabling the evaluation of out-of-distribution generalisation. Each example in TraVLR redundantly encodes the scene in two modalities, allowing either to be dropped or added during training or testing without losing relevant information. We compare the performance of four state-of-the-art V+L models, finding that while they perform well on test examples from the same modality, they all fail at cross-modal transfer and have limited success accommodating the addition or deletion of one modality. We release TraVLR as an open challenge for the research community.

**Fair Enough: Standardizing Evaluation and Model Selection for Fairness Research in NLP**
*Xudong Han, Timothy Baldwin and Trevor Cohn* 11:15-12:45 (Exhibit Hall)
Modern NLP systems exhibit a range of biases, which a growing literature on model debiasing attempts to correct. However, current progress is hampered by a plurality of definitions of bias, means of quantification, and oftentimes vague relation between debiasing algorithms and theoretical measures of bias. This paper seeks to clarify the current situation and plot a course for meaningful progress in fair learning, with two key contributions: (1) making clear inter-relations among the current gamut of methods, and their relation to fairness theory; and (2) addressing the practical problem of model selection, which involves a trade-off between fairness and accuracy and has led to systemic issues in fairness research. Putting them together, we make several recommendations to help shape future work.

**RPTCS: A Reinforced Persona-aware Topic-guiding Conversational System**
*Zishan Ahmad, Kshitij Mishra, Asif Ekbal and Pushpak Bhattacharyya* 11:15-12:45 (Exhibit Hall)
Although there has been a plethora of work on open-domain conversational systems, most of the systems lack the mechanism of controlling the concept transitions in a dialogue. For activities like switching from casual chit-chat to task-oriented conversation, an agent with the ability to manage the flow of concepts in a conversation might be helpful. The user would find the dialogue more engaging and be more receptive to such transitions if these concept transitions were made while taking into account the user's persona. Focusing on persona-aware concept transitions, we propose a Reinforced Persona-aware Topic-guiding Conversational System (RPTCS). Due to the lack of a persona-aware topic transition dataset, we propose a novel conversation dataset creation mechanism in which the conversational agent leads the discourse to drift to a set of target concepts depending on the persona of the speaker and the context of the conversation. To avoid scarcely available expensive human resource, the entire data-creation process is mostly automatic with human-in-loop only for quality checks. This created conversational dataset named PTCD is used to develop the RPTCS in two steps. First, a maximum likelihood estimation loss-based conversational model is trained on PTCD. Then this trained model is fine-tuned in a Reinforcement Learning (RL) framework by employing novel reward functions to assure persona, topic, and context consistency with non-repetitiveness in generated responses. Our experimental results demonstrate the strength of the proposed system with respect to strong baselines.

**Exploring Segmentation Approaches for Neural Machine Translation of Code-Switched Egyptian Arabic-English Text**
*Marwa Gaser, Manuel Mager, Injy Hamed, Nizar Habash, Slim Abdennadher and Ngoc Thang Vu* 11:15-12:45 (Exhibit Hall)
Data sparsity is one of the main challenges posed by code-switching (CS), which is further exacerbated in the case of morphologically rich languages. For the task of machine translation (MT), morphological segmentation has proven successful in alleviating data sparsity in monolingual contexts; however, it has not been investigated for CS settings. In this paper, we study the effectiveness of different segmentation approaches on MT performance, covering morphology-based and frequency-based segmentation techniques. We experiment on MT from code-switched Arabic-English to English. We provide detailed analysis, examining a variety of conditions, such as data size and sentences with different degrees of CS. Empirical results show that morphology-aware segmenters perform the best in segmentation tasks but underperform in MT. Nevertheless, we find that the choice of the segmentation setup to use for MT is highly dependent on the data size. For extreme low-resource scenarios, a combination of frequency and morphology-based segmentations is shown to perform the best. For more resourced settings, such a combination does not bring significant improvements over the use of frequency-based segmentation.

**BERT Is Not The Count: Learning to Match Mathematical Statements with Proofs**
*Weixian Li, Yftah Ziser, Maximin Coavoux and Shay B. Cohen* 11:15-12:45 (Exhibit Hall)
We introduce a task consisting in matching a proof to a given mathematical statement. The task fits well within current research on Mathematical Information Retrieval and, more generally, mathematical article analysis (Mathematical Sciences, 2014). We present a dataset for the task (the MATcH dataset) consisting of over 180k statement-proof pairs extracted from modern mathematical research articles.We find this dataset highly representative of our task, as it consists of relatively new findings useful to mathematicians. We propose a bilinear similarity model and two decoding methods to match statements to proofs effectively. While the first decoding method matches a proof to a statement without being aware of other statements or proofs, the second method treats the task as a global matching problem. Through a symbol replacement procedure, we analyze the "insights" that pre-trained language models have in such mathematical article analysis and show that while these models perform well on this task with the best performing mean reciprocal rank of 73.7, they follow a relatively shallow symbolic analysis and matching to achieve that performance.

**Contrastive Learning with Keyword-based Data Augmentation for Code Search and Code Question Answering**
*Shinwoo Park, Youngwook Kim and Yo-sub Han* 11:15-12:45 (Exhibit Hall)
The semantic code search is to find code snippets from the collection of candidate code snippets with respect to a user query that describes functionality. Recent work on code search proposes data augmentation of queries for contrastive learning. This data augmentation approach modifies random words in queries. When a user web query for searching code snippet is too brief, the important word that represents the search intent of the query could be undesirably modified. A code snippet has informative components such as function name and documentation that describe its functionality. We propose to utilize these code components to identify important words and preserve them in the data augmentation step. We present KeyDAC (Keyword-based Data Augmentation for Contrastive learning) that identifies important words for code search from queries and code components based on term matching. KeyDAC augments query-code pairs while preserving keywords, and then leverages generated training instances for contrastive learning. We use KeyDAC to fine-tune various pre-trained language models and evaluate the performance of code search and code question answering via CoSQA and WebQueryTest. The experimental results confirm that KeyDAC substantially outperforms the current state-of-the-art performance, and achieves the new state-of-the-arts for both tasks.

**Large Scale Multi-Lingual Multi-Modal Summarization Dataset**
*Yash Verma, Anubhav Jangra, Raghvendra Verma and Sriparna Saha* 11:15-12:45 (Exhibit Hall)
Significant developments in techniques such as encoder-decoder models have enabled us to represent information comprising multiple modalities. This information can further enhance many downstream tasks in the field of information retrieval and natural language processing; however, improvements in multi-modal techniques and their performance evaluation require large-scale multi-modal data which offers sufficient diversity. Multi-lingual modeling for a variety of tasks like multi-modal summarization, text generation, and translation leverages information derived from high-quality multi-lingual annotated data. In this work, we present the current largest multi-lingual multi-modal summarization dataset (M3LS), and it consists of over a million instances of document-image pairs along with a professionally annotated multi-modal summary for each pair. It is derived from news articles published by British Broadcasting Corporation(BBC) over a decade and spans 20 languages, targeting diversity across five language roots, it is also the largest summarization dataset for 13 languages and consists of cross-lingual summarization data for 2 languages. We formally define the multi-lingual multi-modal summarization task utilizing our dataset and report baseline scores from various state-of-the-art summarization techniques in a multi-lingual setting. We also compare it with many similar datasets to analyze the uniqueness and difficulty of M3LS. The dataset and code used in this work are made available at "https://github.com/anubhav-jangra/M3LS".

**UScore: An Effective Approach to Fully Unsupervised Evaluation Metrics for Machine Translation**
*Jonas Belouadi and Steffen Eger* 11:15-12:45 (Exhibit Hall)
The vast majority of evaluation metrics for machine translation are supervised, i.e., (i) are trained on human scores, (ii) assume the existence of reference translations, or (iii) leverage parallel data. This hinders their applicability to cases where such supervision signals are not available. In this work, we develop fully unsupervised evaluation metrics. To do so, we leverage similarities and synergies between evaluation metric induction, parallel corpus mining, and MT systems. In particular, we use an unsupervised evaluation metric to mine pseudo-parallel data, which we use to remap deficient underlying vector spaces (in an iterative manner) and to induce an unsupervised MT system, which then provides pseudo-references as an additional component in the metric. Finally, we also induce unsupervised multilingual sentence embeddings from pseudo-parallel data. We show that our fully unsupervised metrics are effective, i.e., they beat supervised competitors on 4 out of our 5 evaluation datasets. We make our code publicly available.

**Social Commonsense for Explanation and Cultural Bias Discovery**
*Lisa Bauer, Hanna Tischer and Mohit Bansal* 11:15-12:45 (Exhibit Hall)
Social commonsense contains many human biases due to social and cultural influence (Sap et al., 2020; Emelin et al., 2020). We focus on identifying cultural biases in data, specifically causal assumptions and commonsense implications, that strongly influence model decisions for a variety of tasks designed for social impact. This enables us to examine data for bias by making explicit the causal (if-then, inferential) relations in social commonsense knowledge used for decision making, furthering interpretable commonsense reasoning from a dataset perspective. We apply our methods on 2 social tasks: emotion detection and perceived value detection. We identify influential social commonsense knowledge to explain model behavior in the following ways. First, we augment large-scale language models with social knowledge and show improvements for the tasks, indicating the implicit assumptions a model requires to be successful on each dataset. Second, we identify influential events in the datasets by using social knowledge to cluster data and demonstrate the influence that these events have on model behavior via leave-K-out experiments. This allows us to gain a dataset-level understanding of the events and causal commonsense relationships that strongly influence predictions. We then analyze these relationships to detect influential cultural bias in each dataset. Finally, we use our influential event identification for detecting mislabeled examples and improve training and performance through their removal. We support our findings with manual analysis.

**Retrieve-and-Fill for Scenario-based Task-Oriented Semantic Parsing**
*Akshat Shrivastava, Shrey Desai, Anchit Gupta, Ali Elkahky, Aleksandr Livshits, Alexander Zotov and Ahmed Aly* 11:15-12:45 (Exhibit Hall)
Task-oriented semantic parsing models have achieved strong results in recent years, but unfortunately do not strike an appealing balance between model size, runtime latency, and cross-domain generalizability. We tackle this problem by introducing scenario-based semantic parsing: a variant of the original task which first requires disambiguating an utterance's "scenario" (an intent-slot template with variable leaf spans) before generating its frame, complete with ontology and utterance tokens. This formulation enables us to isolate coarse-grained and fine-grained aspects of the task, each of which we solve with off-the-shelf neural modules, also optimizing for the axes outlined above. Concretely, we create a Retrieve-and-Fill (RAF) architecture comprised of (1) a retrieval module which ranks the best scenario given an utterance and (2) a filling module which imputes spans into the scenario to create the frame. Our model is modular, differentiable, interpretable, and allows us to garner extra supervision from scenarios. RAF achieves strong results in high-resource, low-resource, and multilingual settings, outperforming recent approaches by wide margins despite, using base pre-trained encoders, small sequence lengths, and parallel decoding.

**Document Flattening: Beyond Concatenating Context for Document-Level Neural Machine Translation**
*Minghao Wu, George Foster, Lizhen Qu and Gholamreza Haffari* 11:15-12:45 (Exhibit Hall)
Existing work in document-level neural machine translation commonly concatenates several consecutive sentences as a pseudo-document, and then learns inter-sentential dependencies. This strategy limits the model's ability to leverage information from distant context. We overcome this limitation with a novel Document Flattening (DocFlat) technique that integrates Flat-Batch Attention (FBA) and Neural Context Gate (NCG) into Transformer model to utilizes information beyond the pseudo-document boundaries. FBA allows the model to attend to all the positions in the batch and model the relationships between positions explicitly and NCG identifies the useful information from the distant context. We conduct comprehensive experiments and analyses on three benchmark datasets for English-German translation, and validate the effectiveness of two variants of DocFlat. Empirical results show that our approach outperforms strong baselines with statistical significance on BLEU, COMET and accuracy on the contrastive test set. The analyses highlight that DocFlat is highly effective in capturing the long-range information.

**Vote'n'Rank: Revision of Benchmarking with Social Choice Theory**
*Mark Rofin, Vladislav Mikhailov, Mikhail Florinsky, Andrey Kravchenko, Tatiana Shavrina, Elena Tutubalina, Daniel Karabekyan and Ekaterina Artemova* 11:15-12:45 (Exhibit Hall)
The development of state-of-the-art systems in different applied areas of machine learning (ML) is driven by benchmarks, which have shaped the paradigm of evaluating generalisation capabilities from multiple perspectives. Although the paradigm is shifting towards more fine-grained evaluation across diverse tasks, the delicate question of how to aggregate the performances has received particular interest in the community. In general, benchmarks follow the unspoken utilitarian principles, where the systems are ranked based on their mean average score over task-specific metrics. Such aggregation procedure has been viewed as a sub-optimal evaluation protocol, which may have created the illusion of progress. This paper proposes Vote'n'Rank, a framework for ranking systems in multi-task benchmarks under the principles of the social choice theory. We demonstrate that our approach can be efficiently utilised to draw new insights on benchmarking in several ML sub-fields and identify the best-performing systems in research and development case studies. The Vote'n'Rank's procedures are more robust than the mean average while being able to handle missing performance scores and determine conditions under which the system becomes the winner.

**Investigating the Effect of Relative Positional Embeddings on AMR-to-Text Generation with Structural Adapters**

*Sebastien Montella, Alexis Nasr, Johannes Heinecke, Frederic Bechet and Lina M. Rojas Barahona* 11:15-12:45 (Exhibit Hall)
Text generation from Abstract Meaning Representation (AMR) has substantially benefited from the popularized Pretrained Language Models (PLMs). Myriad approaches have linearized the input graph as a sequence of tokens to fit the PLM tokenization requirements. Nevertheless, this transformation jeopardizes the structural integrity of the graph and is therefore detrimental to its resulting representation. To overcome this issue, Ribeiro et al. (2021b) have recently proposed StructAdapt, a structure-aware adapter which injects the input graph connectivity within PLMs using Graph Neural Networks (GNNs). In this paper, we investigate the influence of Relative Position Embeddings (RPE) on AMR-to-Text, and, in parallel, we examine the robustness of StructAdapt. Through ablation studies, graph attack and link prediction, we reveal that RPE might be partially encoding input graphs. We suggest further research regarding the role of RPE will provide valuable insights for Graph-to-Text generation.

**On the Intersection of Context-Free and Regular Languages**
*Clemente Pasti, Andreas Opedal, Tiago Pimentel, Tim Vieira, Jason Eisner and Ryan Cotterell* 11:15-12:45 (Exhibit Hall)
The Bar-Hillel construction is a classic result in formal language theory. It shows, by a simple construction, that the intersection of a context-free language and a regular language is itself context-free. In the construction, the regular language is specified by a finite-state automaton. However, neither the original construction (Bar-Hillel et al., 1961) nor its weighted extension (Nederhof and Satta, 2003) can handle finite-state automata with epsilon arcs. While it is possible to remove epsilon arcs from a finite-state automaton efficiently without modifying the language, such an operation modifies the automaton's set of paths. We give a construction that generalizes the Bar- Hillel in the case the desired automaton has epsilon arcs, and further prove that our generalized construction leads to a grammar that encodes the structure of both the input automaton and grammar while retaining the asymptotic size of the original construction.

**Aggregating Crowdsourced and Automatic Judgments to Scale Up a Corpus of Anaphoric Reference for Fiction and Wikipedia Texts**
*Juntao Yu, Silviu Paun, Maris Camilleri, Paloma Garcia, Jon Chamberlain, Udo Kruschwitz and Massimo Poesio* 11:15-12:45 (Exhibit Hall)
Although several datasets annotated for anaphoric reference / coreference exist, even the largest such datasets have limitations in term of size, range of domains, coverage of anaphoric phenomena, and size of documents included. Yet, the approaches proposed to scale up anaphoric annotation haven't so far resulted in datasets overcoming these limitations. In this paper, we introduce a new release of a corpus for anaphoric reference labelled via a game-with-a-purpose. This new release is comparable in size to the largest existing corpora for anaphoric reference due in part to substantial activity by the players, in part thanks to the use of a new resolve-and-aggregate paradigm to 'complete' markable annotations through the combination of an anaphoric resolver and an aggregation method for anaphoric reference. The proposed method could be adopted to greatly speed up annotation time in other projects involving games-with-a-purpose. In addition, the corpus covers genres for which no comparable size datasets exist (Fiction and Wikipedia); it covers singletons and non-referring expressions; and it includes a substantial number of long documents (> 2K in length).

**Teacher Intervention: Improving Convergence of Quantization Aware Training for Ultra-Low Precision Transformers**
*Minsoo Kim, Kyuhong Shim, Seongmin Park, Wonyong Sung and Jungwook Choi* 11:15-12:45 (Exhibit Hall)
Pre-trained Transformer models such as BERT have shown great success in a wide range of applications, but at the cost of substantial increases in model complexity. Quantization-aware training (QAT) is a promising method to lower the implementation cost and energy consumption. However, aggressive quantization below 2-bit causes considerable accuracy degradation due to unstable convergence, especially when the downstream dataset is not abundant. This work proposes a proactive knowledge distillation method called Teacher Intervention (TI) for fast converging QAT of ultra-low precision pre-trained Transformers. TI intervenes layer-wise signal propagation with the intact signal from the teacher to remove the interference of propagated quantization errors, smoothing loss surface of QAT and expediting the convergence. Furthermore, we propose a gradual intervention mechanism to stabilize the recovery of subsections of Transformer layers from quantization. The proposed schemes enable fast convergence of QAT and improve the model accuracy regardless of the diverse characteristics of downstream fine-tuning tasks. We demonstrate that TI consistently achieves superior accuracy with significantly lower fine-tuning iterations on well-known Transformers of natural language processing as well as computer vision compared to the state-of-the-art QAT methods.

**Generative Replay Inspired by Hippocampal Memory Indexing for Continual Language Learning**
*Aru Maekawa, Hidetaka Kamigaito, Kotaro Funakoshi and Manabu Okumura* 11:15-12:45 (Exhibit Hall)
Continual learning aims to accumulate knowledge to solve new tasks without catastrophic forgetting for previously learned tasks. Research on continual learning has led to the development of generative replay, which prevents catastrophic forgetting by generating pseudo-samples for previous tasks and learning them together with new tasks. Inspired by the biological brain, we propose the hippocampal memory indexing to enhance the generative replay by controlling sample generation using compressed features of previous training samples. It enables the generation of a specific training sample from previous tasks, thus improving the balance and quality of generated replay samples. Experimental results indicate that our method effectively controls the sample generation and consistently outperforms the performance of current generative replay methods.

**A Survey of Text Games for Reinforcement Learning Informed by Natural Language**
*Philip Osborne, Heido Nõmm and André Freitas* 11:15-12:45 (Exhibit Hall)
Reinforcement Learning has shown success in a number of complex virtual environments. However, many challenges still exist towards solving problems with natural language as a core component. Interactive Fiction Games (or Text Games) are one such problem type that offer a set of safe, partially observable environments where natural language is required as part of the Reinforcement Learning solution. Therefore, this survey's aim is to assist in the development of new Text Game problem settings and solutions for Reinforcement Learning informed by natural language. Specifically, this survey: 1) introduces the challenges in Text Game Reinforcement Learning problems, 2) outlines the generation tools for rendering Text Games and the subsequent environments generated, and 3) compares the agent architectures currently applied to provide a systematic review of benchmark methodologies and opportunities for future researchers.

**What's New? Summarizing Contributions in Scientific Literature**
*Hiroaki Hayashi, Wojciech Kryscinski, Bryan Mccann, Nazneen Rajani and Caiming Xiong* 11:15-12:45 (Exhibit Hall)
With thousands of academic articles shared on a daily basis, it has become increasingly difficult to keep up with the latest scientific findings. To overcome this problem, we introduce a new task of disentangled paper summarization, which seeks to generate separate summaries for the paper contributions and the context of the work, making it easier to identify the key findings shared in articles. For this purpose, we extend the S2ORC corpus of academic articles, which spans a diverse set of domains ranging from economics to psychology, by adding disentangled "contribution" and "context" reference labels. Together with the dataset, we introduce and analyze three baseline approaches: 1) a unified model controlled by input code prefixes, 2) a model with separate generation heads specialized in generating the disentangled outputs, and 3) a training strategy that guides the model using additional supervision coming from inbound and outbound citations. We also propose a comprehensive automatic evaluation protocol which reports the relevance, novelty, and disentanglement of generated outputs. Through a human study involving expert annotators, we show that in 79%, of cases our new task is considered more helpful than traditional scientific paper summarization.

### What's New? Summarizing Contributions in Scientific Literature

*Hiroaki Hayashi, Wojciech Kryscinski, Bryan Mccann, Nazneen Rajani and Caiming Xiong* 11:15-12:45 (Exhibit Hall)

With thousands of academic articles shared on a daily basis, it has become increasingly difficult to keep up with the latest scientific findings. To overcome this problem, we introduce a new task of disentangled paper summarization, which seeks to generate separate summaries for the paper contributions and the context of the work, making it easier to identify the key findings shared in articles. For this purpose, we extend the S2ORC corpus of academic articles, which spans a diverse set of domains ranging from economics to psychology, by adding disentangled "contribution" and "context" reference labels. Together with the dataset, we introduce and analyze three baseline approaches: 1) a unified model controlled by input code prefixes, 2) a model with separate generation heads specialized in generating the disentangled outputs, and 3) a training strategy that guides the model using additional supervision coming from inbound and outbound citations. We also propose a comprehensive automatic evaluation protocol which reports the relevance, novelty, and disentanglement of generated outputs. Through a human study involving expert annotators, we show that in 79%, of cases our new task is considered more helpful than traditional scientific paper summarization.

### Meta Self-Refinement for Robust Learning with Weak Supervision

*Dawei Zhu, Xiaoyu Shen, Michael Hedderich and Dietrich Klakow* 11:15-12:45 (Exhibit Hall)

Training deep neural networks (DNNs) under weak supervision has attracted increasing research attention as it can significantly reduce the annotation cost. However, labels from weak supervision can be noisy, and the high capacity of DNNs enables them to easily overfit the label noise, resulting in poor generalization. Recent methods leverage self-training to build noise-resistant models, in which a teacher trained under weak supervision is used to provide highly confident labels for teaching the students. Nevertheless, the teacher derived from such frameworks may have fitted a substantial amount of noise and therefore produce incorrect pseudo-labels with high confidence, leading to severe error propagation. In this work, we propose Meta Self-Refinement (MSR), a noise-resistant learning framework, to effectively combat label noise from weak supervision. Instead of relying on a fixed teacher trained with noisy labels, we encourage the teacher to refine its pseudo-labels. At each training step, MSR performs a meta gradient descent on the current mini-batch to maximize the student performance on a clean validation set. Extensive experimentation on eight NLP benchmarks demonstrates that MSR is robust against label noise in all settings and outperforms state-of-the-art methods by up to 11.4% in accuracy and 9.26% in F1 score.

### Persona Expansion with Commonsense Knowledge for Diverse and Consistent Response Generation

*Donghyun Kim, Youbin Ahn, Wongyu Kim, Chanhee Lee, Kyungchan Lee, Kyong-ho Lee, Jeonguk Kim, Donghoon Shin and Yeonsoo Lee* 11:15-12:45 (Exhibit Hall)

Generating diverse and consistent responses is the ultimate goal of a persona-based dialogue. Although many studies have been conducted, the generated responses tend to be generic and bland due to the personas' limited descriptiveness. Therefore, it is necessary to expand the given personas for more attractive responses. However, indiscriminate expansion of personas threaten the consistency of responses and therefore reduce the interlocutor's interest in conversation. To alleviate this issue, we propose a consistent persona expansion framework that improves not only the diversity but also the consistency of persona-based responses. To do so, we define consistency criteria to avoid possible contradictions among personas as follows: 1) Intra-Consistency and 2) Inter-Consistency. Then, we construct a silver profile dataset to deliver the ability to conform with the consistency criteria to the expansion model. Finally, we propose a persona expansion model with an encoder-decoder structure, which considers the relatedness and consistency among personas. Our experiments on the Persona-Chat dataset demonstrate the superiority of the proposed framework.

### Persona Expansion with Commonsense Knowledge for Diverse and Consistent Response Generation

*Donghyun Kim, Youbin Ahn, Wongyu Kim, Chanhee Lee, Kyungchan Lee, Kyong-ho Lee, Jeonguk Kim, Donghoon Shin and Yeonsoo Lee* 11:15-12:45 (Exhibit Hall)

Generating diverse and consistent responses is the ultimate goal of a persona-based dialogue. Although many studies have been conducted, the generated responses tend to be generic and bland due to the personas' limited descriptiveness. Therefore, it is necessary to expand the given personas for more attractive responses. However, indiscriminate expansion of personas threaten the consistency of responses and therefore reduce the interlocutor's interest in conversation. To alleviate this issue, we propose a consistent persona expansion framework that improves not only the diversity but also the consistency of persona-based responses. To do so, we define consistency criteria to avoid possible contradictions among personas as follows: 1) Intra-Consistency and 2) Inter-Consistency. Then, we construct a silver profile dataset to deliver the ability to conform with the consistency criteria to the expansion model. Finally, we propose a persona expansion model with an encoder-decoder structure, which considers the relatedness and consistency among personas. Our experiments on the Persona-Chat dataset demonstrate the superiority of the proposed framework.

### UnifEE: Unified Evidence Extraction for Fact Verification

*Nan Hu, Zirui Wu, Yuxuan Lai, Chen Zhang and Yansong Feng* 11:15-12:45 (Exhibit Hall)

FEVEROUS is a fact extraction and verification task that requires systems to extract evidence of both sentences and table cells from a Wikipedia dump, then predict the veracity of the given claim accordingly. Existing works extract evidence in the two formats separately, ignoring potential connections between them. In this paper, we propose a Unified Evidence Extraction model (UnifEE), which uses a mixed evidence graph to extract the evidence in both formats. With the carefully-designed unified evidence graph, UnifEE allows evidence interactions among all candidates in both formats at similar granularity. Experiments show that, with information aggregated from related evidence candidates in the fusion graph, UnifEE can make better decisions about which evidence should be kept, especially for claims requiring multi-hop reasoning or a combination of tables and texts. Thus it outperforms all previous evidence extraction methods and brings significant improvement in the subsequent claim verification step.

### K-hop neighbourhood regularization for few-shot learning on graphs: A case study of text classification

*Niels Van Der Heijden, Ekaterina Shutova and Helen Yannakoudakis* 11:15-12:45 (Exhibit Hall)

We present FewShotTextGCN, a novel method designed to effectively utilize the properties of word-document graphs for improved learning in low-resource settings. We introduce K-hop Neighbourhood Regularization, a regularizer for heterogeneous graphs, and show that it stabilizes and improves learning when only a few training samples are available. We furthermore propose a simplification in the graph-construction method, which results in a graph that is 7 times less dense and yields better performance in little-resource settings while performing on par with the state of the art in high-resource settings. Finally, we introduce a new variant of Adaptive Pseudo-Labeling tailored for word-document graphs. When using as little as 20 samples for training, we outperform a strong TextGCN baseline with 17% in absolute accuracy on average over eight languages. We demonstrate that our method can be applied to document classification without any language model pretraining on a wide range of typologically diverse languages while performing on par with large pretrained language models.

### Improving Visual-Semantic Embedding with Adaptive Pooling and Optimization Objective

*Zijian Zhang, Chang Shu, Ya Xiao, Yuan Shen, Di Zhu, Youxin Chen, Jing Xiao, Jey Han Lau, Qian Zhang and Zheng Lu* 11:15-12:45 (Exhibit Hall)

Visual-Semantic Embedding (VSE) aims to learn an embedding space where related visual and semantic instances are close to each other. Recent VSE models tend to design complex structures to pool visual and semantic features into fixed-length vectors and use hard triplet loss for optimization. However, we find that: (1) combining simple pooling methods is no worse than these sophisticated methods; and (2) only considering the most difficult-to-distinguish negative sample leads to slow convergence and poor Recall@K improvement. To this end, we propose an adaptive pooling strategy that allows the model to learn how to aggregate features through a combination of simple pooling methods. We also introduce a strategy to dynamically select a group of negative samples to make the optimization converge faster and perform better. Experimental results on Flickr30K and MS-COCO demonstrate that a standard VSE using our pooling and optimization strategies outperforms current state-of-the-art systems (at least 1.0 extbackslash% on the metrics of recall) in image-to-text and text-to-image retrieval.

**Modelling Emotion Dynamics in Song Lyrics with State Space Models**
*Yingjin Song and Daniel Beck* 11:15-12:45 (Exhibit Hall)
Most previous work in music emotion recognition assumes a single or a few song-level labels for the whole song. While it is known that different emotions can vary in intensity within a song, annotated data for this setup is scarce and difficult to obtain. In this work, we propose a method to predict emotion dynamics in song lyrics without song-level supervision. We frame each song as a time series and employ a State Space Model (SSM), combining a sentence-level emotion predictor with an Expectation-Maximization (EM) procedure to generate the full emotion dynamics. Our experiments show that applying our method consistently improves the performance of sentence-level baselines without requiring any annotated songs, making it ideal for limited training data scenarios. Further analysis through case studies shows the benefits of our method while also indicating the limitations and pointing to future directions.

**Policy-based Reinforcement Learning for Generalisation in Interactive Text-based Environments**
*Edan Toledo, Jan Buys and Jonathan Shock* 11:15-12:45 (Exhibit Hall)
Text-based environments enable RL agents to learn to converse and perform interactive tasks through natural language. However, previous RL approaches applied to text-based environments show poor performance when evaluated on unseen games. This paper investigates the improvement of generalisation performance through the simple switch from a value-based update method to a policy-based one, within text-based environments. We show that by replacing commonly used value-based methods with REINFORCE with baseline, a far more general agent is produced. The policy-based agent is evaluated on Coin Collector and Question Answering with interactive text (QAit), two text-based environments designed to test zero-shot performance. We see substantial improvements on a variety of zero-shot evaluation experiments, including tripling accuracy on various QAit benchmark configurations. The results indicate that policy-based RL has significantly better generalisation capabilities than value-based methods within such text-based environments, suggesting that RL agents could be applied to more complex natural language environments.

**Do Deep Neural Networks Capture Compositionality in Arithmetic Reasoning?**
*Keito Kudo, Yoichi Aoki, Tatsuki Kuribayashi, Ana Brassard, Masashi Yoshikawa, Keisuke Sakaguchi and Kentaro Inui* 11:15-12:45 (Exhibit Hall)
Compositionality is a pivotal property of symbolic reasoning. However, how well recent neural models capture compositionality remains underexplored in the symbolic reasoning tasks. This study empirically addresses this question by systematically examining recently published pre-trained seq2seq models with a carefully controlled dataset of multi-hop arithmetic symbolic reasoning. We introduce a skill tree on compositionality in arithmetic symbolic reasoning that defines the hierarchical levels of complexity along with three compositionality dimensions: systematicity, productivity, and substitutivity. Our experiments revealed that among the three types of composition, the models struggled most with systematicity, performing poorly even with relatively simple compositions. That difficulty was not resolved even after training the models with intermediate reasoning steps.

**BLM-AgrF: A New French Benchmark to Investigate Generalization of Agreement in Neural Networks**
*Aixiu An, Chunyang Jiang, Maria A. Rodriguez, Vivi Nastase and Paola Merlo* 11:15-12:45 (Exhibit Hall)
Successful machine learning systems currently rely on massive amounts of data, which are very effective in hiding some of the shallowness of the learned models. To help train models with more complex and compositional skills, we need challenging data, on which a system is successful only if it detects structure and regularities, that will allow it to generalize. In this paper, we describe a French dataset (BLM-AgrF) for learning the underlying rules of subject-verb agreement in sentences, developed in the BLM framework, a new task inspired by visual IQ tests known as Raven's Progressive Matrices. In this task, an instance consists of sequences of sentences with specific attributes. To predict the correct answer as the next element of the sequence, a model must correctly detect the generative model used to produce the dataset. We provide details and share a dataset built following this methodology. Two exploratory baselines based on commonly used architectures show that despite the simplicity of the phenomenon, it is a complex problem for deep learning systems.

**Self-Adapted Utterance Selection for Suicidal Ideation Detection in Lifeline Conversations**
*Zhong-ling Wang, Po-hsien Huang, Wen-yau Hsu and Hen-hsen Huang* 11:15-12:45 (Exhibit Hall)
This paper investigates a crucial aspect of mental health by exploring the detection of suicidal ideation in spoken phone conversations between callers and counselors at a suicide prevention hotline. These conversations can be lengthy, noisy, and cover a broad range of topics, making it challenging for NLP models to accurately identify the caller's suicidal ideation. To address these difficulties, we introduce a novel, self-adaptive approach that identifies the most critical utterances that the NLP model can more easily distinguish. The experiments use real-world Lifeline transcriptions, expertly labeled, and show that our approach outperforms the baseline models in overall performance with an F-score of 66.01%. In detecting the most dangerous cases, our approach achieves a significantly higher F-score of 65.94% compared to the baseline models, an improvement of 8.9%. The selected utterances can also provide valuable insights for suicide prevention research. Furthermore, our approach demonstrates its versatility by showing its effectiveness in sentiment analysis, making it a valuable tool for NLP applications beyond the healthcare domain.

**LongEval: Guidelines for Human Evaluation of Faithfulness in Long-form Summarization**
*Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan and Kyle Lo* 11:15-12:45 (Exhibit Hall)
While human evaluation remains best practice for accurately judging the faithfulness of automatically-generated summaries, few solutions exist to address the increased difficulty and workload when evaluating long-form summaries. Through a survey of 162 papers on long-form summarization, we first shed light on current human evaluation practices surrounding long-form summaries. We find that 73% of these papers do not perform any human evaluation on model-generated summaries, while other works face new difficulties that manifest when dealing with long documents (e.g., low inter-annotator agreement). Motivated by our survey, we present LongEval, a set of guidelines for human evaluation of faithfulness in long-form summaries that addresses the following challenges: (1) How can we achieve high inter-annotator agreement on faithfulness scores? (2) How can we minimize annotator workload while maintaining accurate faithfulness scores? and (3) Do humans benefit from automated alignment between summary and source snippets? We deploy LongEval in annotation studies on two long-form summarization datasets in different domains (SQuALITY and PubMed), and we find that switching to a finer granularity of judgment (e.g., clause-level) reduces inter-annotator variance in faithfulness scores (e.g., std-dev from 18.5 to 6.8). We also show that scores from a partial annotation of fine-grained units highly correlates with scores from a full annotation workload (0.89 Kendall"s tau using 50% judgements). We release our human judgments, annotation templates, and software as a Python library for future research.

**Empathy Identification Systems are not Accurately Accounting for Context**

*Andrew Lee, Jonathan Kummerfeld, Larry An and Rada Mihalcea* 11:15-12:45 (Exhibit Hall)
Understanding empathy in text dialogue data is a difficult, yet critical, skill for effective human-machine interaction. In this work, we ask whether systems are making meaningful progress on this challenge. We consider a simple model that checks if an input utterance is similar to a small set of empathetic examples. Crucially, the model does not look at what the utterance is a response to, i.e., the dialogue context. This model performs comparably to other work on standard benchmarks and even outperforms state-of-the-art models for empathetic rationale extraction by 16.7 points on T-F1 and 4.3 on IOU-F1. This indicates that current systems rely on the surface form of the response, rather than whether it is suitable in context. To confirm this, we create examples with dialogue contexts that change the interpretation of the response and show that current systems continue to label utterances as empathetic. We discuss the implications of our findings, including improvements for empathetic benchmarks and how our model can be an informative baseline.

**Enhancing Multi-Document Summarization with Cross-Document Graph-based Information Extraction**
*Zixuan Zhang, Heba Elfardy, Markus Dreyer, Kevin Small, Heng Ji and Mohit Bansal* 11:15-12:45 (Exhibit Hall)
Information extraction (IE) and summarization are closely related, both tasked with presenting a subset of the information contained in a natural language text. However, while IE extracts structural representations, summarization aims to abstract the most salient information into a generated text summary – thus potentially encountering the technical limitations of current text generation methods (e.g., hallucination). To mitigate this risk, this work uses structured IE graphs to enhance the abstractive summarization task. Specifically, we focus on improving Multi-Document Summarization (MDS) performance by using cross-document IE output, incorporating two novel components: (1) the use of auxiliary entity and event recognition systems to focus the summary generation model; (2) incorporating an alignment loss between IE nodes and their text spans to reduce inconsistencies between the IE graphs and text representations. Operationally, both the IE nodes and corresponding text spans are projected into the same embedding space and pairwise distance is minimized. Experimental results on multiple MDS benchmarks show that summaries generated by our model are more factually consistent with the source documents than baseline models while maintaining the same level of abstractiveness.

**Multi-Modal Bias: Introducing a Framework for Stereotypical Bias Assessment beyond Gender and Race in Vision–Language Models**
*Sepehr Janghorbani and Gerard De Melo* 11:15-12:45 (Exhibit Hall)
Recent breakthroughs in self-supervised training have led to a new class of pretrained vision–language models. While there have been investigations of bias in multimodal models, they have mostly focused on gender and racial bias, giving much less attention to other relevant groups, such as minorities with regard to religion, nationality, sexual orientation, or disabilities. This is mainly due to lack of suitable benchmarks for such groups. We seek to address this gap by providing a visual and textual bias benchmark called MMBias, consisting of around 3,800 images and phrases covering 14 population subgroups. We utilize this dataset to assess bias in several prominent self-supervised multimodal models, including CLIP, ALBEF, and ViLT. Our results show that these models demonstrate meaningful bias favoring certain groups. Finally, we introduce a debiasing method designed specifically for such large pretrained models that can be applied as a post-processing step to mitigate bias, while preserving the remaining accuracy of the model.

**Performance Prediction via Bayesian Matrix Factorisation for Multilingual Natural Language Processing Tasks**
*Viktoria Schram, Daniel Beck and Trevor Cohn* 11:15-12:45 (Exhibit Hall)
Performance prediction for Natural Language Processing (NLP) seeks to reduce the experimental burden resulting from the myriad of different evaluation scenarios, e.g., the combination of languages used in multilingual transfer. In this work, we explore the framework of Bayesian matrix factorisation for performance prediction, as many experimental settings in NLP can be naturally represented in matrix format. Our approach outperforms the state-of-the-art in several NLP benchmarks, including machine translation and cross-lingual entity linking. Furthermore, it also avoids hyperparameter tuning and is able to provide uncertainty estimates over predictions.

**Unified Neural Topic Model via Contrastive Learning and Term Weighting**
*Sungwon Han, Mingi Shin, Sungkyu Park, Changwook Jung and Meeyoung Cha* 11:15-12:45 (Exhibit Hall)
Two types of topic modeling predominate: generative methods that employ probabilistic latent models and clustering methods that identify semantically coherent groups. This paper newly presents UTopic (Unified neural Topic model via contrastive learning and term weighting) that combines the advantages of these two types. UTopic uses contrastive learning and term weighting to learn knowledge from a pretrained language model and discover influential terms from semantically coherent clusters. Experiments show that the generated topics have a high-quality topic-word distribution in terms of topic coherence, outperforming existing baselines across multiple topic coherence measures. We demonstrate how our model can be used as an add-on to existing topic models and improve their performance.

**Memory-efficient Temporal Moment Localization in Long Videos**
*Cristian Rodriguez, Edison Marrese-taylor, Basura Fernando, Hiroya Takamura and Qi Wu* 11:15-12:45 (Exhibit Hall)
Temporal Moment Localization is a challenging multi-modal task which aims to identify the start and end timestamps of a moment of interest in an input untrimmed video, given a query in natural language. Solving this task correctly requires understanding the temporal relationships in the entire input video, but processing such long inputs and reasoning about them is memory and computationally expensive. In light of this issue, we propose Stochastic Bucket-wise Feature Sampling (SBFS), a stochastic sampling module that allows methods to process long videos at a constant memory footprint. We further combine SBFS with a new consistency loss to propose Locformer, a Transformer-based model that can process videos as long as 18 minutes. We test our proposals on relevant benchmark datasets, showing that not only can Locformer achieve excellent results, but also that our sampling is more effective than competing counterparts. Concretely, SBFS consistently improves the performance of prior work, by up to 3.13\% in the mean temporal IoU, leading to a new state-of-the-art performance on Charades-STA and YouCookII, while also obtaining up to 12.8x speed-up at testing time and reducing memory requirements by up to 5x.

**DREEAM: Guiding Attention with Evidence for Improving Document-Level Relation Extraction**
*Youmi Ma, An Wang and Naoaki Okazaki* 11:15-12:45 (Exhibit Hall)
Document-level relation extraction (DocRE) is the task of identifying all relations between each entity pair in a document. Evidence, defined as sentences containing clues for the relationship between an entity pair, has been shown to help DocRE systems focus on relevant texts, thus improving relation extraction. However, evidence retrieval (ER) in DocRE faces two major issues: high memory consumption and limited availability of annotations. This work aims at addressing these issues to improve the usage of ER in DocRE. First, we propose DREEAM, a memory-efficient approach that adopts evidence information as the supervisory signal, thereby guiding the attention modules of the DocRE system to assign high weights to evidence. Second, we propose a self-training strategy for DREEAM to learn ER from automatically-generated evidence on massive data without evidence annotations. Experimental results reveal that our approach exhibits state-of-the-art performance on the DocRED benchmark for both DocRE and ER. To the best of our knowledge, DREEAM is the first approach to employ ER self-training.

**Probing Cross-Lingual Lexical Knowledge from Multilingual Sentence Encoders**
*Ivan Vulić, Goran Glavaš, Fangyu Liu, Nigel Collier, Edoardo Maria Ponti and Anna Korhonen* 11:15-12:45 (Exhibit Hall)
Pretrained multilingual language models (LMs) can be successfully transformed into multilingual sentence encoders (SEs; e.g., LaBSE, xMP-Net) via additional fine-tuning or model distillation with parallel data. However, it remains unclear how to best leverage them to represent

sub-sentence lexical items (i.e., words and phrases) in cross-lingual lexical tasks. In this work, we probe SEs for the amount of cross-lingual lexical knowledge stored in their parameters, and compare them against the original multilingual LMs. We also devise a simple yet efficient method for exposing the cross-lingual lexical knowledge by means of additional fine-tuning through inexpensive contrastive learning that requires only a small amount of word translation pairs. Using bilingual lexical induction (BLI), cross-lingual lexical semantic similarity, and cross-lingual entity linking as lexical probing tasks, we report substantial gains on standard benchmarks (e.g., +10 Precision@1 points in BLI). The results indicate that the SEs such as LaBSE can be 'rewired' into effective cross-lingual lexical encoders via the contrastive learning procedure, and that it is possible to expose more cross-lingual lexical knowledge compared to using them as off-the-shelf SEs. This way, we also provide an effective tool for harnessing 'covert' multilingual lexical knowledge hidden in multilingual sentence encoders.

**Socratic Question Generation: A Novel Dataset, Models, and Evaluation**
*Beng Heng Ang, Sujatha Das Gollapalli and See-kiong Ng* 11:15-12:45 (Exhibit Hall)
Socratic questioning is a form of reflective inquiry often employed in education to encourage critical thinking in students, and to elicit awareness of beliefs and perspectives in a subject during therapeutic counseling. Specific types of Socratic questions are employed for enabling reasoning and alternate views against the context of individual personal opinions on a topic. Socratic contexts are different from traditional question generation contexts where "answer-seeking" questions are generated against a given formal passage on a topic, narrative stories or conversations.

We present SocratiQ, the first large dataset of 110K (question, context) pairs for enabling studies on Socratic Question Generation (SoQG). We provide an in-depth study on the various types of Socratic questions and present models for generating Socratic questions against a given context through prompt tuning. Our automated and human evaluation results demonstrate that our SoQG models can produce realistic, type-sensitive, human-like Socratic questions enabling potential applications in counseling and coaching.

**COVID-VTS: Fact Extraction and Verification on Short Video Platforms**
*Fuxiao Liu, Yaser Yacoob and Abhinav Shrivastava* 11:15-12:45 (Exhibit Hall)
We introduce a new benchmark, COVID-VTS, for fact-checking multi-modal information involving short-duration videos with COVID19-focused information from both the real world and machine generation. We propose, TwtrDetective, an effective model incorporating cross-media consistency checking to detect token-level malicious tampering in different modalities, and generate explanations. Due to the scarcity of training data, we also develop an efficient and scalable approach to automatically generate misleading video posts by event manipulation or adversarial matching. We investigate several state-of-the-art models and demonstrate the superiority of TwtrDetective.

# Parallel Session 4 - 14:15-15:45

## Session 4 Orals – Ethical and Sustainable NLP – Room B

14:15-15:45 (Elafiti 3)

**A Two-Sided Discussion of Preregistration of NLP Research**
*Anders Søgaard, Daniel Hershcovich and Miryam De Lhoneux* 16:30-16:45 (Elafiti 3)
Van Miltenburg et al. (2021) suggest NLP research should adopt preregistration to prevent fishing expeditions and to promote publication of negative results. At face value, this is a very reasonable suggestion, seemingly solving many methodological problems with NLP research. We discuss pros and cons - some old, some new: a) Preregistration is challenged by the practice of retrieving hypotheses after the results are known; b) preregistration may bias NLP toward confirmatory research; c) preregistration must allow for reclassification of research as exploratory; d) preregistration may increase publication bias; e) preregistration may increase flag-planting; f) preregistration may increase p-hacking; and finally, g) preregistration may make us less risk tolerant. We cast our discussion as a dialogue, presenting both sides of the debate.

**Counter-GAP: Counterfactual Bias Evaluation through Gendered Ambiguous Pronouns**
*Zhongbin Xie, Vid Kocijan, Thomas Lukasiewicz and Oana-maria Camburu* 16:45-17:00 (Elafiti 3)
Bias-measuring datasets play a critical role in detecting biased behavior of language models and in evaluating progress of bias mitigation methods. In this work, we focus on evaluating gender bias through coreference resolution, where previous datasets are either hand-crafted or fail to reliably measure an explicitly defined bias. To overcome these shortcomings, we propose a novel method to collect diverse, natural, and minimally distant text pairs via counterfactual generation, and construct Counter-GAP, an annotated dataset consisting of 4008 instances grouped into 1002 quadruples. We further identify a bias cancellation problem in previous group-level metrics on Counter-GAP, and propose to use the difference between inconsistency across genders and within genders to measure bias at a quadruple level. Our results show that four pre-trained language models are significantly more inconsistent across different gender groups than within each group, and that a name-based counterfactual data augmentation method is more effective to mitigate such bias than an anonymization-based method.

**In-Depth Look at Word Filling Societal Bias Measures**
*Matúš Pikuliak, Ivana Beňová and Viktor Bachratý* 17:00-17:15 (Elafiti 3)
Many measures of societal bias in language models have been proposed in recent years. A popular approach is to use a set of word filling prompts to evaluate the behavior of the language models. In this work, we analyze the validity of two such measures – StereoSet and CrowS-Pairs. We show that these measures produce unexpected and illogical results when appropriate control group samples are constructed. Based on this, we believe that they are problematic and using them in the future should be reconsidered. We propose a way forward with an improved testing protocol. Finally, we also introduce a new gender bias dataset for Slovak.

**Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey**
*Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos and Yulia Tsvetkov* 17:15-17:30 (Elafiti 3)
Recent advances in the capacity of large language models to generate human-like text have resulted in their increased adoption in user-facing settings. In parallel, these improvements have prompted a heated discourse around the risks of societal harms they introduce, whether inadvertent or malicious. Several studies have explored these harms and called for their mitigation via development of safer, fairer models. Going beyond enumerating the risks of harms, this work provides a survey of practical methods for addressing potential threats and societal harms from language generation models. We draw on several prior works' taxonomies of language model risks to present a structured overview of strategies for detecting and ameliorating different kinds of risks/harms of language generators. Bridging diverse strands of research, this survey aims to serve as a practical guide for both LM researchers and practitioners, with explanations of different strategies' motivations, their limitations, and open problems for future research.

**Multi-Modal Bias: Introducing a Framework for Stereotypical Bias Assessment beyond Gender and Race in Vision–Language Mod-**

**els**

*Sepehr Janghorbani and Gerard De Melo* 17:30-17:45 (Elafiti 3)

Recent breakthroughs in self-supervised training have led to a new class of pretrained vision–language models. While there have been investigations of bias in multimodal models, they have mostly focused on gender and racial bias, giving much less attention to other relevant groups, such as minorities with regard to religion, nationality, sexual orientation, or disabilities. This is mainly due to lack of suitable benchmarks for such groups. We seek to address this gap by providing a visual and textual bias benchmark called MMBias, consisting of around 3,800 images and phrases covering 14 population subgroups. We utilize this dataset to assess bias in several prominent self-supervised multimodal models, including CLIP, ALBEF, and ViLT. Our results show that these models demonstrate meaningful bias favoring certain groups. Finally, we introduce a debiasing method designed specifically for such large pretrained models that can be applied as a post-processing step to mitigate bias, while preserving the remaining accuracy of the model.

**SODAPOP: Open-Ended Discovery of Social Biases in Social Commonsense Reasoning Models**

*Haozhe An, Zongxia Li, Jieyu Zhao and Rachel Rudinger* 17:45-18:00 (Elafiti 3)

A common limitation of diagnostic tests for detecting social biases in NLP models is that they may only detect stereotypic associations that are pre-specified by the designer of the test. Since enumerating all possible problematic associations is infeasible, it is likely these tests fail to detect biases that are present in a model but not pre-specified by the designer. To address this limitation, we propose SODAPOP (SOcial bias Discovery from Answers about PeOPle), an approach for automatic social bias discovery in social commonsense question-answering. The SODAPOP pipeline generates modified instances from the Social IQa dataset (Sap et al., 2019b) by (1) substituting names associated with different demographic groups, and (2) generating many distractor answers from a masked language model. By using a social commonsense model to score the generated distractors, we are able to uncover the model's stereotypic associations between demographic groups and an open set of words. We also test SODAPOP on debiased models and show the limitations of multiple state-of-the-art debiasing algorithms.

## Session 4 Orals – Summarization and Medical Applications – Room C

14:15-15:45 (Elafiti 4)

**An Empirical Study of Clinical Note Generation from Doctor-Patient Encounters**

*Asma Ben Abacha, Wen-wai Yim, Yadan Fan and Thomas Lin* 16:30-16:45 (Elafiti 4)

Medical doctors spend on average 52 to 102 minutes per day writing clinical notes from their patient encounters (Hripcsak et al., 2011). Reducing this workload calls for relevant and efficient summarization methods. In this paper, we introduce new resources and empirical investigations for the automatic summarization of doctor-patient conversations in a clinical setting. In particular, we introduce the MTS-Dialog dataset; a new collection of 1,700 doctor-patient dialogues and corresponding clinical notes. We use this new dataset to investigate the feasibility of this task and the relevance of existing language models, data augmentation, and guided summarization techniques. We compare standard evaluation metrics based on n-gram matching, contextual embeddings, and Fact Extraction to assess the accuracy and the factual consistency of the generated summaries. To ground these results, we perform an expert-based evaluation using relevant natural language generation criteria and task-specific criteria such as critical omissions, and study the correlation between the automatic metrics and expert judgments. To the best of our knowledge, this study is the first attempt to introduce an open dataset of doctor-patient conversations and clinical notes, with detailed automated and manual evaluations of clinical note generation.

**CHARD: Clinical Health-Aware Reasoning Across Dimensions for Text Generation Models**

*Steven Y. Feng, Vivek Khetan, Bogdan Sacaleanu, Anatole Gershman and Eduard Hovy* 16:45-17:00 (Elafiti 4)

We motivate and introduce CHARD: Clinical Health-Aware Reasoning across Dimensions, to investigate the capability of text generation models to act as implicit clinical knowledge bases and generate free-flow textual explanations about various health-related conditions across several dimensions. We collect and present an associated dataset, CHARDat, consisting of explanations about 52 health conditions across three clinical dimensions. We conduct extensive experiments using BART and T5 along with data augmentation, and perform automatic, human, and qualitative analyses. We show that while our models can perform decently, CHARD is very challenging with strong potential for further exploration.

**Incorporating Question Answering-Based Signals into Abstractive Summarization via Salient Span Selection**

*Daniel Deutsch and Dan Roth* 17:00-17:15 (Elafiti 4)

In this work, we propose a method for incorporating question-answering (QA) signals into a summarization model. Our method identifies salient noun phrases (NPs) in the input document by automatically generating wh-questions that are answered by the NPs and automatically determining whether those questions are answered in the gold summaries. This QA-based signal is incorporated into a two-stage summarization model which first marks salient NPs in the input document using a classification model, then conditionally generates a summary. Our experiments demonstrate that the models trained using QA-based supervision generate higher-quality summaries than baseline methods of identifying salient spans on benchmark summarization datasets. Further, we show that the content of the generated summaries can be controlled based on which NPs are marked in the input document. Finally, we propose a method of augmenting the training data so the gold summaries are more consistent with the marked input spans used during training and show how this results in models which learn to better exclude unmarked document content.

**KGVL-BART: Knowledge Graph Augmented Visual Language BART for Radiology Report Generation**

*Kaveri Kale, Pushpak Bhattacharyya, Milind Gune, Aditya Shetty and Rustom Lawyer* 17:15-17:30 (Elafiti 4)

Timely generation of radiology reports and diagnoses is a challenge worldwide due to the enormous number of cases and shortage of radiology specialists. In this paper, we propose a Knowledge Graph Augmented Vision Language BART (KGVL-BART) model that takes as input two chest X-ray images- one frontal and the other lateral- along with tags which are diagnostic keywords, and outputs a report with the patient-specific findings. Our system development effort is divided into 3 stages: i) construction of the Chest X-ray KG (referred to as chestX-KG), ii) image feature extraction, and iii) training a KGVL-BART model using the visual, text, and KG data. The dataset we use is the well-known Indiana University Chest X-ray reports with the train, validation, and test split of 3025 instances, 300 instances, and 500 instances respectively. We construct a Chest X-Ray knowledge graph from these reports by extracting entity1-relation-entity2 triples; the triples get extracted by a rule-based tool of our own. Constructed KG is verified by two experienced radiologists (with experience of 30 years and 8 years, respectively). We demonstrate that our model- KGVL-BART- outperforms State-of-the-Art transformer-based models on standard NLG scoring metrics. We also include a qualitative evaluation of our system by experienced radiologist (with experience of 30 years) on the test data, which showed that 73% of the reports generated were fully correct, only 5.5% are completely wrong and 21.5% have important missing details though overall correct. To the best of our knowledge, ours is the first system to make use of multi-modality and domain knowledge to generate X-ray reports automatically.

**LongEval: Guidelines for Human Evaluation of Faithfulness in Long-form Summarization**
*Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan and Kyle Lo* 17:30-17:45 (Elafiti 4)
While human evaluation remains best practice for accurately judging the faithfulness of automatically-generated summaries, few solutions exist to address the increased difficulty and workload when evaluating long-form summaries. Through a survey of 162 papers on long-form summarization, we first shed light on current human evaluation practices surrounding long-form summaries. We find that 73% of these papers do not perform any human evaluation on model-generated summaries, while other works face new difficulties that manifest when dealing with long documents (e.g., low inter-annotator agreement). Motivated by our survey, we present LongEval, a set of guidelines for human evaluation of faithfulness in long-form summaries that addresses the following challenges: (1) How can we achieve high inter-annotator agreement on faithfulness scores? (2) How can we minimize annotator workload while maintaining accurate faithfulness scores? and (3) Do humans benefit from automated alignment between summary and source snippets? We deploy LongEval in annotation studies on two long-form summarization datasets in different domains (SQuALITY and PubMed), and we find that switching to a finer granularity of judgment (e.g., clause-level) reduces inter-annotator variance in faithfulness scores (e.g., std-dev from 18.5 to 6.8). We also show that scores from a partial annotation of fine-grained units highly correlates with scores from a full annotation workload (0.89 Kendall"s tau using 50% judgements). We release our human judgments, annotation templates, and software as a Python library for future research.

**When Do Pre-Training Biases Propagate to Downstream Tasks? A Case Study in Text Summarization**
*Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen Mckeown and Tatsunori Hashimoto* 17:45-18:00 (Elafiti 4)
Large language models (LLMs) are subject to sociocultural and other biases previously identified using intrinsic evaluations. However, when and how these intrinsic biases in pre-trained LM representations propagate to downstream, fine-tuned NLP tasks like summarization is not well understood. In this work, we investigate one type of bias—name-nationality bias—and trace it from the pre-training stage to a downstream summarization task across multiple summarization modeling choices. We show that these biases manifest themselves as hallucinations in summarization, leading to factually incorrect summaries. We also find that this propagation of biases is algorithm-dependent: more abstractive models allow biases to propagate more directly to downstream tasks as hallucinated facts. Building on these observations, we further analyze how changes to the adaptation method and fine-tuning data set affect name nationality biases and show that while they can reduce the overall rate of hallucinations, they do not change the types of biases that do appear.

## Session 4 Orals – Text Classification, Sentiment Analysis and Argument Mining – Room A

14:15-15:45 (Elafiti 2)

**Modelling Emotion Dynamics in Song Lyrics with State Space Models**
*Yingjin Song and Daniel Beck* 16:30-16:45 (Elafiti 2)
Most previous work in music emotion recognition assumes a single or a few song-level labels for the whole song. While it is known that different emotions can vary in intensity within a song, annotated data for this setup is scarce and difficult to obtain. In this work, we propose a method to predict emotion dynamics in song lyrics without song-level supervision. We frame each song as a time series and employ a State Space Model (SSM), combining a sentence-level emotion predictor with an Expectation-Maximization (EM) procedure to generate the full emotion dynamics. Our experiments show that applying our method consistently improves the performance of sentence-level baselines without requiring any annotated songs, making it ideal for limited training data scenarios. Further analysis through case studies shows the benefits of our method while also indicating the limitations and pointing to future directions.

**ConEntail: An Entailment-based Framework for Universal Zero and Few Shot Classification with Supervised Contrastive Pretraining**
*Haoran Zhang, Aysa Xuemo Fan and Rui Zhang* 16:45-17:00 (Elafiti 2)
A universal classification model aims to generalize to diverse classification tasks in both zero and few shot settings. A promising way toward universal classification is to cast heterogeneous data formats into a dataset-agnostic "meta-task" (e.g., textual entailment, question answering) then pretrain a model on the combined meta dataset. The existing work is either pretrained on specific subsets of classification tasks, or pretrained on both classification and generation data but the model could not fulfill its potential in universality and reliability. These also leave a massive amount of annotated data under-exploited. To fill these gaps, we propose ConEntail, a new framework for universal zero and few shot classification with supervised contrastive pretraining. Our unified meta-task for classification is based on nested entailment. It can be interpreted as "Does sentence a entails [sentence b entails label c]". This formulation enables us to make better use of 57 annotated classification datasets for supervised contrastive pretraining and universal evaluation. In this way, ConEntail helps the model (1) absorb knowledge from different datasets, and (2) gain consistent performance gain with more pretraining data. In experiments, we compare our model with discriminative and generative models pretrained on the same dataset. The results confirm that our framework effectively exploits existing annotated data and consistently outperforms baselines in both zero (9.4% average improvement) and few shot settings (3.5% average improvement). Our code is available in supplementary materials.

**Robustness Challenges in Model Distillation and Pruning for Natural Language Understanding**
*Mengnan Du, Subhabrata Mukherjee, Yu Cheng, Milad Shokouhi, Xia Hu and Ahmed Hassan Awadallah* 17:00-17:15 (Elafiti 2)
Recent work has focused on compressing pre-trained language models (PLMs) like BERT where the major focus has been to improve the in-distribution performance for downstream tasks. However, very few of these studies have analyzed the impact of compression on the generalizability and robustness of compressed models for out-of-distribution (OOD) data. Towards this end, we study two popular model compression techniques including knowledge distillation and pruning and show that the compressed models are significantly less robust than their PLM counterparts on OOD test sets although they obtain similar performance on in-distribution development sets for a task. Further analysis indicates that the compressed models overfit on the shortcut samples and generalize poorly on the hard ones. We further leverage this observation to develop a regularization strategy for robust model compression based on sample uncertainty.

**Unified Neural Topic Model via Contrastive Learning and Term Weighting**
*Sungwon Han, Mingi Shin, Sungkyu Park, Changwook Jung and Meeyoung Cha* 17:15-17:30 (Elafiti 2)
Two types of topic modeling predominate: generative methods that employ probabilistic latent models and clustering methods that identify semantically coherent groups. This paper newly presents UTopic (Unified neural Topic model via contrastive learning and term weighting) that combines the advantages of these two types. UTopic uses contrastive learning and term weighting to learn knowledge from a pretrained language model and discover influential terms from semantically coherent clusters. Experiments show that the generated topics have a high-quality topic-word distribution in terms of topic coherence, outperforming existing baselines across multiple topic coherence measures. We demonstrate how our model can be used as an add-on to existing topic models and improve their performance.

**Reinforced Sequence Training based Subjective Bias Correction**
*Karthic Madanagopal and James Caverlee* 17:30-17:45 (Elafiti 2)
Subjective bias is ubiquitous on news sites, social media, and knowledge resources like Wikipedia. Many existing methods for subjective bias correction have typically focused on making one-word edits and have been trained over a single (often, noisy) domain. In contrast, we propose a novel reinforced sequence training approach for robust subjective bias correction. Three of the unique characteristics of the approach are: (i) it balances bias neutralization with fluency and semantics preservation through reinforcement learning, to broaden the scope to bias beyond a single word; (ii) it is cross-trained over multiple sources of bias to be more robust to new styles of biased writing that are not seen in the training data for a single domain; and (iii) it is used to fine-tune a large pre-trained transformer model to yield state-of-the-art performance in bias text correction task. Extensive experiments show that the proposed approach results in significant improvements in subjective bias correction versus alternatives.

**Uncovering Implicit Inferences for Improved Relational Argument Mining**
*Ameer Saadat-yazdi, Jeff Pan and Nadin Kokciyan* 17:45-18:00 (Elafiti 2)
Argument mining seeks to extract arguments and their structure from unstructured texts. Identifying relations between arguments (such as attack, support, and neutral) is a challenging task because two arguments may be related to each other via implicit inferences. This task often requires external commonsense knowledge to discover how one argument relates to another. State-of-the-art methods, however, rely on pre-defined knowledge graphs, and thus might not cover target argument pairs well. We introduce a new generative neuro-symbolic approach to finding inference chains that connect the argument pairs by making use of the Commonsense Transformer (COMET). We evaluate our approach on three datasets for both the two-label (attack/support) and three-label (attack/support/neutral) tasks. Our approach significantly outperforms the state-of-the-art, by 2-5% in F1 score, on all three datasets.

# Main Conference: Wednesday, May 3, 2023

# Parallel Session 6 - 09:00-10:30

## Session 6 Orals – Generation - Room B

09:00-10:30 (Elafiti 3)

**PANCETTA: Phoneme Aware Neural Completion to Elicit Tongue Twisters Automatically**
*Sedrick Scott Keh, Steven Y. Feng, Varun Gangal, Malihe Alikhani and Eduard Hovy* 09:00-09:15 (Elafiti 3)
Tongue twisters are meaningful sentences that are difficult to pronounce. The process of automatically generating tongue twisters is challenging since the generated utterance must satisfy two conditions at once: phonetic difficulty and semantic meaning. Furthermore, phonetic difficulty is itself hard to characterize and is expressed in natural tongue twisters through a heterogeneous mix of phenomena such as alliteration and homophony. In this paper, we propose PANCETTA: Phoneme Aware Neural Completion to Elicit Tongue Twisters Automatically. We leverage phoneme representations to capture the notion of phonetic difficulty, and we train language models to generate original tongue twisters on two proposed task settings. To do this, we curate a dataset called TT-Corp, consisting of existing English tongue twisters. Through automatic and human evaluation, as well as qualitative analysis, we show that PANCETTA generates novel, phonetically difficult, fluent, and semantically meaningful tongue twisters.

**Investigating the Effect of Relative Positional Embeddings on AMR-to-Text Generation with Structural Adapters**
*Sebastien Montella, Alexis Nasr, Johannes Heinecke, Frederic Bechet and Lina M. Rojas Barahona* 09:15-09:30 (Elafiti 3)
Text generation from Abstract Meaning Representation (AMR) has substantially benefited from the popularized Pretrained Language Models (PLMs). Myriad approaches have linearized the input graph as a sequence of tokens to fit the PLM tokenization requirements. Nevertheless, this transformation jeopardizes the structural integrity of the graph and is therefore detrimental to its resulting representation. To overcome this issue, Ribeiro et al. (2021b) have recently proposed StructAdapt, a structure-aware adapter which injects the input graph connectivity within PLMs using Graph Neural Networks (GNNs). In this paper, we investigate the influence of Relative Position Embeddings (RPE) on AMR-to-Text, and, in parallel, we examine the robustness of StructAdapt. Through ablation studies, graph attack and link prediction, we reveal that RPE might be partially encoding input graphs. We suggest further research regarding the role of RPE will provide valuable insights for Graph-to-Text generation.

**Multimodal Event Transformer for Image-guided Story Ending Generation**
*Yucheng Zhou and Guodong Long* 09:30-09:45 (Elafiti 3)
Image-guided story ending generation (IgSEG) is to generate a story ending based on given story plots and ending image. Existing methods focus on cross-modal feature fusion but overlook reasoning and mining implicit information from story plots and ending image. To tackle this drawback, we propose a multimodal event transformer, an event-based reasoning framework for IgSEG. Specifically, we construct visual and semantic event graphs from story plots and ending image, and leverage event-based reasoning to reason and mine implicit information in a single modality. Next, we connect visual and semantic event graphs and utilize cross-modal fusion to integrate different-modality features. In addition, we propose a multimodal injector to adaptive pass essential information to decoder. Besides, we present an incoherence detection to enhance the understanding context of a story plot and the robustness of graph modeling for our model. Experimental results show that our method achieves state-of-the-art performance for the image-guided story ending generation.

**Adding Instructions during Pretraining: Effective way of Controlling Toxicity in Language Models**
*Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi and Bryan Catanzaro* 09:45-10:00 (Elafiti 3)
Pretrained large language models have become indispensable for solving various natural language processing (NLP) tasks. However, safely deploying them in real world applications is challenging because they generate toxic content. To address this challenge, we propose two novel pretraining data augmentation strategies that significantly reduce model toxicity without compromising its utility. Our two strategies are: (1) MEDA: adds raw toxicity score as meta-data to the pretraining samples, and (2) INST: adds instructions to those samples indicating their toxicity. Our results indicate that our best performing strategy (INST) substantially reduces the toxicity probability up to 61% while preserving the accuracy on five benchmark NLP tasks as well as improving AUC scores on four bias detection tasks by 1.3%. We also demonstrate the generalizability of our techniques by scaling the number of training samples and the number of model parameters.

**PCC: Paraphrasing with Bottom-k Sampling and Cyclic Learning for Curriculum Data Augmentation**
*Hongyuan Lu and Wai Lam* 10:00-10:15 (Elafiti 3)
Curriculum Data Augmentation (CDA) improves neural models by presenting synthetic data with increasing difficulties from easy to hard. However, traditional CDA simply treats the ratio of word perturbation as the difficulty measure and goes through the curriculums only once. This paper presents \textbf{PCC}: \textbf{P}araphrasing with Bottom-k Sampling and \textbf{C}yclic Learning for \textbf{C}urriculum Data Augmentation, a novel CDA framework via paraphrasing, which exploits the textual paraphrase similarity as the curriculum difficulty measure. We propose a curriculum-aware paraphrase generation module composed of three units: a paraphrase candidate generator with bottom-k sampling, a filtering mechanism and a difficulty measure. We also propose a cyclic learning strategy that passes through the curriculums multiple times. The bottom-k sampling is proposed to generate super-hard instances for the later curriculums. Experimental results on few-shot text classification as well as dialogue generation indicate that PCC surpasses competitive baselines. Human evaluation and extensive case studies indicate that bottom-k sampling effectively generates super-hard instances, and PCC significantly improves the baseline dialogue agent.\footnote{Code will be released upon publication.}

**LoFT: Enhancing Faithfulness and Diversity for Table-to-Text Generation via Logic Form Control**
*Yilun Zhao, Zhenting Qi, Linyong Nan, Lorenzo Jaime Flores and Dragomir Radev* 10:15-10:30 (Elafiti 3)
Logical Table-to-Text (LT2T) generation is tasked with generating logically faithful sentences from tables. There currently exists two challenges in the field: 1) Faithfulness: how to generate sentences that are factually correct given the table content; 2) Diversity: how to generate multiple sentences that offer different perspectives on the table. This work proposes LoFT, which utilizes logic forms as fact verifiers and content planners to control LT2T generation. Experimental results on the LogicNLG dataset demonstrate that LoFT is the first model that addresses unfaithfulness and lack of diversity issues simultaneously. Our code is publicly available at https://github.com/Yale-LILY/LoFT.

## Session 6 Orals – Information Extraction – Room A

09:00-10:30 (Elafiti 2)

**AutoTriggER: Label-Efficient and Robust Named Entity Recognition with Auxiliary Trigger Extraction**
*Dong-ho Lee, Ravi Kiran Selvam, Sheikh Muhammad Sarwar, Bill Yuchen Lin, Fred Morstatter, Jay Pujara, Elizabeth Boschee, James Allan and Xiang Ren* 09:00-09:15 (Elafiti 2)
Deep neural models for named entity recognition (NER) have shown impressive results in overcoming label scarcity and generalizing to unseen entities by leveraging distant supervision and auxiliary information such as explanations. However, the costs of acquiring such additional information are generally prohibitive. In this paper, we present a novel two-stage framework (AutoTriggER) to improve NER performance by automatically generating and leveraging “entity triggers” which are human-readable cues in the text that help guide the model to make better decisions. Our framework leverages post-hoc explanation to generate rationales and strengthens a model’s prior knowledge using an embedding interpolation technique. This approach allows models to exploit triggers to infer entity boundaries and types instead of solely memorizing the entity words themselves. Through experiments on three well-studied NER datasets, AutoTriggER shows strong label-efficiency, is capable of generalizing to unseen entities, and outperforms the RoBERTa-CRF baseline by nearly 0.5 F1 points on average.

**Event Temporal Relation Extraction with Bayesian Translational Model**
*Xingwei Tan, Gabriele Pergola and Yulan He* 09:15-09:30 (Elafiti 2)
Existing models to extract temporal relations between events lack a principled method to incorporate external knowledge. In this study, we introduce Bayesian-Trans, a Bayesian learning-based method that models the temporal relation representations as latent variables and infers their values via Bayesian inference and translational functions. Compared to conventional neural approaches, instead of performing point estimation to find the best set parameters, the proposed model infers the parameters’ posterior distribution directly, enhancing the model’s capability to encode and express uncertainty about the predictions. Experimental results on the three widely used datasets show that Bayesian-Trans outperforms existing approaches for event temporal relation extraction. We additionally present detailed analyses on uncertainty quantification, comparison of priors, and ablation studies, illustrating the benefits of the proposed approach.

**GLADIS: A General and Large Acronym Disambiguation Benchmark**
*Lihu Chen, Gael Varoquaux and Fabian Suchanek* 09:30-09:45 (Elafiti 2)
Acronym Disambiguation (AD) is crucial for natural language understanding on various sources, including biomedical reports, scientific papers, and search engine queries. However, existing acronym disambiguation benchmarks and tools are limited to specific domains, and the size of prior benchmarks is rather small. To accelerate the research on acronym disambiguation, we construct a new benchmark with three components: (1) a much larger acronym dictionary with 1.5M acronyms and 6.4M long forms; (2) a pre-training corpus with 160 million sentences; (3) three datasets that cover the general, scientific, and biomedical domains. We then pre-train a language model, \emph{AcroBERT}, on our constructed corpus for general acronym disambiguation, and show the challenges and values of our new benchmark.

**Iterative Document-level Information Extraction via Imitation Learning**
*Yunmo Chen, William Gantt, Weiwei Gu, Tongfei Chen, Aaron White and Benjamin Van Durme* 09:45-10:00 (Elafiti 2)
We present a novel iterative extraction model, IterX, for extracting complex relations, or templates, i.e., N-tuples representing a mapping from named slots to spans of text within a document. Documents may feature zero or more instances of a template of any given type, and the task of template extraction entails identifying the templates in a document and extracting each template’s slot values. Our imitation learning approach casts the problem as a Markov decision process (MDP), and relieves the need to use predefined template orders to train an extractor. It leads to state-of-the-art results on two established benchmarks – 4-ary relation extraction on SciREX and template extraction on MUC-4 – as well as a strong baseline on the new BETTER Granular task.

**Towards Integration of Discriminability and Robustness for Document-Level Relation Extraction**
*Jia Guo, Stanley Kok and Lidong Bing* 10:00-10:15 (Elafiti 2)
Document-level relation extraction (DocRE) predicts relations for entity pairs that rely on long-range context-dependent reasoning in a document. As a typical multi-label classification problem, DocRE faces the challenge of effectively distinguishing a small set of positive relations from the majority of negative ones. This challenge becomes even more difficult to overcome when there exists a significant number of annotation errors in the dataset. In this work, we aim to achieve better integration of both the discriminability and robustness for the DocRE problem. Specifically, we first design an effective loss function to endow high discriminability to both probabilistic outputs and internal representations. We innovatively customize entropy minimization and supervised contrastive learning for the challenging multi-label and long-tailed learning problems. To ameliorate the impact of label errors, we equipped our method with a novel negative label sampling strategy to strengthen the model robustness. In addition, we introduce two new data regimes to mimic more realistic scenarios with annotation errors and evaluate our sampling strategy. Experimental results verify the effectiveness of each component and show that our method achieves new state-of-the-art results on the DocRED dataset, its recently cleaned version, Re-DocRED, and the proposed data regimes.

**Weakly-Supervised Questions for Zero-Shot Relation Extraction**
*Saeed Najafi and Alona Fyshe* 10:15-10:30 (Elafiti 2)
Zero-Shot Relation Extraction (ZRE) is the task of Relation Extraction where the training and test sets have no shared relation types. This very challenging domain is a good test of a model’s ability to generalize. Previous approaches to ZRE reframed relation extraction as Question Answering (QA), allowing for the use of pre-trained QA models. However, this method required manually creating gold question templates for each new relation. Here, we do away with these gold templates and instead learn a model that can generate questions for unseen relations. Our technique can successfully translate relation descriptions into relevant questions, which are then leveraged to generate the correct tail entity. On tail entity extraction, we outperform the previous state-of-the-art by more than 16 F1 points without using gold question templates. On the RE-QA dataset where no previous baseline for relation extraction exists, our proposed algorithm comes within 0.7 F1 points of a system that uses gold question templates. Our model also outperforms the state-of-the-art ZRE baselines on the FewRel and WikiZSL datasets, showing that QA models no longer need template questions to match the performance of models specifically tailored to the ZRE task. Our implementation is available at https://github.com/fyshelab/QA-ZRE.

## Session 6 Orals – Interpretability and Model Analysis – Room C

09:00-10:30 (Elafiti 4)

**Assessing Out-of-Domain Language Model Performance from Few Examples**

*Prasann Singhal, Jarad Forristal, Xi Ye and Greg Durrett* 09:00-09:15 (Elafiti 4)
While pretrained language models have exhibited impressive generalization capabilities, they still behave unpredictably under certain domain shifts. In particular, a model may learn a reasoning process on in-domain training data that does not hold for out-of-domain test data. We address the task of predicting out-of-domain (OOD) performance in a few-shot fashion: given a few target-domain examples and a set of models with similar training performance, can we understand how these models will perform on OOD test data? We benchmark the performance on this task when looking at model accuracy on the few-shot examples, then investigate how to incorporate analysis of the models' behavior using feature attributions to better tackle this problem. Specifically, we explore a set of factors designed to reveal model agreement with certain pathological heuristics that may indicate worse generalization capabilities. On textual entailment, paraphrase recognition, and a synthetic classification task, we show that attribution-based factors can help rank relative model OOD performance. However, accuracy on a few-shot test set is a surprisingly strong baseline, particularly when the system designer does not have in-depth prior knowledge about the domain shift.

**COMPS: Conceptual Minimal Pair Sentences for testing Robust Property Knowledge and its Inheritance in Pre-trained Language Models**
*Kanishka Misra, Julia Rayz and Allyson Ettinger* 09:15-09:30 (Elafiti 4)
A characteristic feature of human semantic cognition is its ability to not only store and retrieve the properties of concepts observed through experience, but to also facilitate the inheritance of properties (can breathe) from superordinate concepts (animal) to their subordinates (dog)—i.e. demonstrate property inheritance. In this paper, we present COMPS, a collection of minimal pair sentences that jointly tests pre-trained language models (PLMs) on their ability to attribute properties to concepts and their ability to demonstrate property inheritance behavior. Analyses of 22 different PLMs on COMPS reveal that they can easily distinguish between concepts on the basis of a property when they are trivially different, but find it relatively difficult when concepts are related on the basis of nuanced knowledge representations. Furthermore, we find that PLMs can show behaviors suggesting successful property inheritance in simple contexts, but fail in the presence of distracting information, which decreases the performance of many models sometimes even below chance. This lack of robustness in demonstrating simple reasoning raises important questions about PLMs' capacity to make correct inferences even when they appear to possess the prerequisite knowledge.

**Mind the Labels: Describing Relations in Knowledge Graphs With Pretrained Models**
*Zdeněk Kasner, Ioannis Konstas and Ondrej Dusek* 09:30-09:45 (Elafiti 4)
Pretrained language models (PLMs) for data-to-text (D2T) generation can use human-readable data labels such as column headings, keys, or relation names to generalize to out-of-domain examples. However, the models are well-known in producing semantically inaccurate outputs if these labels are ambiguous or incomplete, which is often the case in D2T datasets. In this paper, we expose this issue on the task of descibing a relation between two entities. For our experiments, we collect a novel dataset for verbalizing a diverse set of 1,522 unique relations from three large-scale knowledge graphs (Wikidata, DBPedia, YAGO). We find that although PLMs for D2T generation expectedly fail on unclear cases, models trained with a large variety of relation labels are surprisingly robust in verbalizing novel, unseen relations. We argue that using data with a diverse set of clear and meaningful labels is key to training D2T generation systems capable of generalizing to novel domains.

**Step by Step Loss Goes Very Far: Multi-Step Quantization for Adversarial Text Attacks**
*Piotr Gaiński and Klaudia Bałazy* 09:45-09:55 (Elafiti 4)
We propose a novel gradient-based attack against transformer-based language models that searches for an adversarial example in a continuous space of tokens probabilities. Our algorithm mitigates the gap between adversarial loss for continuous and discrete text representations by performing multi-step quantization in a quantization-compensation loop. Experiments show that our method significantly outperforms other approaches on various natural language processing (NLP) tasks.

**Probing Cross-Lingual Lexical Knowledge from Multilingual Sentence Encoders**
*Ivan Vulić, Goran Glavaš, Fangyu Liu, Nigel Collier, Edoardo Maria Ponti and Anna Korhonen* 09:55-10:10 (Elafiti 4)
Pretrained multilingual language models (LMs) can be successfully transformed into multilingual sentence encoders (SEs; e.g., LaBSE, xMPNet) via additional fine-tuning or model distillation with parallel data. However, it remains unclear how to best leverage them to represent sub-sentence lexical items (i.e., words and phrases) in cross-lingual lexical tasks. In this work, we probe SEs for the amount of cross-lingual lexical knowledge stored in their parameters, and compare them against the original multilingual LMs. We also devise a simple yet efficient method for exposing the cross-lingual lexical knowledge by means of additional fine-tuning through inexpensive contrastive learning that requires only a small amount of word translation pairs. Using bilingual lexical induction (BLI), cross-lingual lexical semantic similarity, and cross-lingual entity linking as lexical probing tasks, we report substantial gains on standard benchmarks (e.g., +10 Precision@1 points in BLI). The results indicate that the SEs such as LaBSE can be 'rewired' into effective cross-lingual lexical encoders via the contrastive learning procedure, and that it is possible to expose more cross-lingual lexical knowledge compared to using them as off-the-shelf SEs. This way, we also provide an effective tool for harnessing 'covert' multilingual lexical knowledge hidden in multilingual sentence encoders.

**Understanding Transformer Memorization Recall Through Idioms**
*Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg and Mor Geva* 10:10-10:25 (Elafiti 4)
To produce accurate predictions, language models (LMs) must balance between generalization and memorization. Yet, little is known about the mechanism by which transformer LMs employ their memorization capacity. When does a model decide to output a memorized phrase, and how is this phrase then retrieved from memory? In this work, we offer the first methodological framework for probing and characterizing recall of memorized sequences in transformer LMs. First, we lay out criteria for detecting model inputs that trigger memory recall, and propose idioms as inputs that typically fulfill these criteria. Next, we construct a dataset of English idioms and use it to compare model behavior on memorized vs. non-memorized inputs. Specifically, we analyze the internal prediction construction process by interpreting the model's hidden representations as a gradual refinement of the output probability distribution. We find that across different model sizes and architectures, memorized predictions are a two-step process: early layers promote the predicted token to the top of the output distribution, and upper layers increase model confidence. This suggests that memorized information is stored and retrieved in the early layers of the network. Last, we demonstrate the utility of our methodology beyond idioms in memorized factual statements. Overall, our work makes a first step towards understanding memory recall, and provides a methodological basis for future studies of transformer memorization.

## Session 6 Posters

09:00-10:30 (Exhibit Hall)

**Penguins Don't Fly: Reasoning about Generics through Instantiations and Exceptions**
*Emily Allaway, Jena D. Hwang, Chandra Bhagavatula, Kathleen Mckeown, Doug Downey and Yejin Choi* 09:00-10:30 (Exhibit Hall)

Generics express generalizations about the world (e.g., birds can fly) that are not universally true (e.g., newborn birds and penguins cannot fly). Commonsense knowledge bases, used extensively in NLP, encode some generic knowledge but rarely enumerate such exceptions and knowing when a generic statement holds or does not hold true is crucial for developing a comprehensive understanding of generics. We present a novel framework informed by linguistic theory to generate exemplars—specific cases when a generic holds true or false. We generate ~19k exemplars for ~650 generics and show that our framework outperforms a strong GPT-3 baseline by 12.8 precision points. Our analysis highlights the importance of linguistic theory-based controllability for generating exemplars, the insufficiency of knowledge bases as a source of exemplars, and the challenges exemplars pose for the task of natural language inference.

**Retrieval Enhanced Data Augmentation for Question Answering on Privacy Policies**
*Md Rizwan Parvez, Jianfeng Chi, Wasi Uddin Ahmad, Yuan Tian and Kai-wei Chang* 09:00-10:30 (Exhibit Hall)
Prior studies in privacy policies frame the question answering (QA) task as identifying the most relevant text segment or a list of sentences from a policy document given a user query. Existing labeled datasets are heavily imbalanced (only a few relevant segments), limiting the QA performance in this domain. In this paper, we develop a data augmentation framework based on ensembling retriever models that captures the relevant text segments from unlabeled policy documents and expand the positive examples in the training set. In addition, to improve the diversity and quality of the augmented data, we leverage multiple pre-trained language models (LMs) and cascaded them with noise reduction oracles. Using our augmented data on the PrivacyQA benchmark, we elevate the existing baseline by a large margin (10% F1) and achieve a new state-of-the-art F1 score of 50%. Our ablation studies provide further insights into the effectiveness of our approach.

**Do dialogue representations align with perception? An empirical study**
*Sarenne Wallbridge, Peter Bell and Catherine Lai* 09:00-10:30 (Exhibit Hall)
There has been a surge of interest regarding the alignment of large-scale language models with human language comprehension behaviour. The majority of this research investigates comprehension behaviours from reading isolated, written sentences. We propose studying the perception of dialogue, focusing on an intrinsic form of language use: spoken conversations.

Using the task of predicting upcoming dialogue turns, we ask whether turn plausibility scores produced by state-of-the-art language models correlate with human judgements. We find a strong correlation for some but not all models: masked language models produce stronger correlations than auto-regressive models. In doing so, we quantify human performance on the response selection task for open-domain spoken conversation. To the best of our knowledge, this is the first such quantification.

We find that response selection performance can be used as a coarse proxy for the strength of correlation with human judgements, however humans and models make different response selection mistakes. The model which produces the strongest correlation also outperforms human response selection performance. Through ablation studies, we show that pre-trained language models provide a useful basis for turn representations; however, fine-grained contextualisation, inclusion of dialogue structure information, and fine-tuning towards response selection all boost response selection accuracy by over 30 absolute points.

**StyLEx: Explaining Style Using Human Lexical Annotations**
*Shirley Anugrah Hayati, Kyumin Park, Dheeraj Rajagopal, Lyle Ungar and Dongyeop Kang* 09:00-10:30 (Exhibit Hall)
Large pre-trained language models have achieved impressive results on various style classification tasks, but they often learn spurious domain-specific words to make predictions (Hayati et al., 2021). While human explanation highlights stylistic tokens as important features for this task, we observe that model explanations often do not align with them. To tackle this issue, we introduce StyLEx, a model that learns from human annotated explanations of stylistic features and jointly learns to perform the task and predict these features as model explanations. Our experiments show that StyLEx can provide human like stylistic lexical explanations without sacrificing the performance of sentence-level style prediction on both in-domain and out-of-domain datasets. Explanations from StyLEx show significant improvements in explanation metrics (sufficiency, plausibility) and when evaluated with human annotations. They are also more understandable by human judges compared to the widely-used saliency-based explanation baseline.

**Dynamic Benchmarking of Masked Language Models on Temporal Concept Drift with Multiple Views**
*Katerina Margatina, Shuai Wang, Yogarshi Vyas, Neha Anna John, Yassine Benajiba and Miguel Ballesteros* 09:00-10:30 (Exhibit Hall)
Temporal concept drift refers to the problem of data changing over time. In the field of NLP, that would entail that language (e.g. new expressions, meaning shifts) and factual knowledge (e.g. new concepts, updated facts) evolve over time. Focusing on the latter, we benchmark 11 pretrained masked language models (MLMs) on a series of tests designed to evaluate the effect of temporal concept drift, as it is crucial that widely used language models remain up-to-date with the ever-evolving factual updates of the real world. Specifically, we provide a holistic framework that (1) dynamically creates temporal test sets of any time granularity (e.g. month, quarter, year) of factual data from Wikidata, (2) constructs fine-grained splits of tests (e.g. updated, new, unchanged facts) to ensure comprehensive analysis, and (3) evaluates MLMs in three distinct ways (single-token probing, multi-token generation, MLM scoring). In contrast to prior work, our framework aims to unveil how robust an MLM is over time and thus to provide a signal in case it has become outdated, by leveraging multiple views of evaluation.

**Real-Time Visual Feedback to Guide Benchmark Creation: A Human-and-Metric-in-the-Loop Workflow**
*Anjana Arunkumar, Swaroop Mishra, Bhavdeep Singh Sachdeva, Chitta Baral and Chris Bryan* 09:00-10:30 (Exhibit Hall)
Recent research has shown that language models exploit ‘artifacts’ in benchmarks to solve tasks, rather than truly learning them, leading to inflated model performance. In pursuit of creating better benchmarks, we propose VAIDA, a novel benchmark creation paradigm for NLP, that focuses on guiding crowdworkers, an under-explored facet of addressing benchmark idiosyncrasies. VAIDA facilitates sample correction by providing realtime visual feedback and recommendations to improve sample quality. Our approach is domain, model, task, and metric agnostic, and constitutes a paradigm shift for robust, validated, and dynamic benchmark creation via human-and-metric-in-the-loop workflows. We evaluate via expert review and a user study with NASA TLX. We find that VAIDA decreases effort, frustration, mental, and temporal demands of crowdworkers and analysts, simultaneously increasing the performance of both user groups with a 45.8% decrease in the level of artifacts in created samples. As a by product of our user study, we observe that created samples are adversarial across models, leading to decreases of 31.3% (BERT), 22.5% (RoBERTa), 14.98% (GPT-3 fewshot) in performance.

**Why Can’t Discourse Parsing Generalize? A Thorough Investigation of the Impact of Data Diversity**
*Yang Janet Liu and Amir Zeldes* 09:00-10:30 (Exhibit Hall)
Recent advances in discourse parsing performance create the impression that, as in other NLP tasks, performance for high-resource languages such as English is finally becoming reliable. In this paper we demonstrate that this is not the case, and thoroughly investigate the impact of data diversity on RST parsing stability. We show that state-of-the-art architectures trained on the standard English newswire benchmark do not generalize well, even within the news domain. Using the two largest RST corpora of English with text from multiple genres, we quantify the impact of genre diversity in training data for achieving generalization to text types unseen during training. Our results show that a heterogeneous training regime is critical for stable and generalizable models, across parser architectures. We also provide error analyses of model outputs and out-of-domain performance. To our knowledge, this study is the first to fully evaluate cross-corpus RST parsing generalizability on complete trees, examine between-genre degradation within an RST corpus, and investigate the impact of genre diversity in training data composition.

**Path Spuriousness-aware Reinforcement Learning for Multi-Hop Knowledge Graph Reasoning**
*Chunyang Jiang, Tianchen Zhu, Haoyi Zhou, Chang Liu, Ting Deng, Chunming Hu and Jianxin Li* 09:00-10:30 (Exhibit Hall)
Multi-hop reasoning, a prevalent approach for query answering, aims at inferring new facts along reasonable paths over a knowledge graph. Reinforcement learning methods can be adopted by formulating the problem into a Markov decision process. However, common suffering within RL-based reasoning models is that the agent can be biased to spurious paths which coincidentally lead to the correct answer with poor explanation. In this work, we take a deep dive into this phenomenon and define a metric named Path Spuriousness (PS), to quantitatively estimate to what extent a path is spurious. Guided by the definition of PS, we design a model with a new reward that considers both answer accuracy and path reasonableness. We test our method on four datasets and experiments reveal that our method considerably enhances the agent's capacity to prevent spurious paths while keeping comparable to state-of-the-art performance.

**Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey**
*Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos and Yulia Tsvetkov* 09:00-10:30 (Exhibit Hall)
Recent advances in the capacity of large language models to generate human-like text have resulted in their increased adoption in user-facing settings. In parallel, these improvements have prompted a heated discourse around the risks of societal harms they introduce, whether inadvertent or malicious. Several studies have explored these harms and called for their mitigation via development of safer, fairer models. Going beyond enumerating the risks of harms, this work provides a survey of practical methods for addressing potential threats and societal harms from language generation models. We draw on several prior works' taxonomies of language model risks to present a structured overview of strategies for detecting and ameliorating different kinds of risks/harms of language generators. Bridging diverse strands of research, this survey aims to serve as a practical guide for both LM researchers and practitioners, with explanations of different strategies' motivations, their limitations, and open problems for future research.

**Generation-Based Data Augmentation for Offensive Language Detection: Is It Worth It?**
*Camilla Casula and Sara Tonelli* 09:00-10:30 (Exhibit Hall)
Generation-based data augmentation (DA) has been presented in several works as a way to improve offensive language detection. However, the effectiveness of generative DA has been shown only in limited scenarios, and the potential injection of biases when using generated data to classify offensive language has not been investigated. Our aim is that of analyzing the feasibility of generative data augmentation more in-depth with two main focuses. First, we investigate the robustness of models trained on generated data in a variety of data augmentation setups, both novel and already presented in previous work, and compare their performance on four widely-used English offensive language datasets that present inherent differences in terms of content and complexity. In addition to this, we analyze models using the HateCheck suite, a series of functional tests created to challenge hate speech detection systems. Second, we investigate potential lexical bias issues through a qualitative analysis on the generated data. We find that the potential positive impact of generative data augmentation on model performance is unreliable, and generative DA can also have unpredictable effects on lexical bias.

**Quantifying Context Mixing in Transformers**
*Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała and Afra Alishahi* 09:00-10:30 (Exhibit Hall)
Self-attention weights and their transformed variants have been the main source of information for analyzing token-to-token interactions in Transformer-based models. But despite their ease of interpretation, these weights are not faithful to the models' decisions as they are only one part of an encoder, and other components in the encoder layer can have considerable impact on information mixing in the output representations. In this work, by expanding the scope of analysis to the whole encoder block, we propose Value Zeroing, a novel context mixing score customized for Transformers that provides us with a deeper understanding of how information is mixed at each encoder layer. We demonstrate the superiority of our context mixing score over other analysis methods through a series of complementary evaluations with different viewpoints based on linguistically informed rationales, probing, and faithfulness analysis.

**Semantic Specialization for Knowledge-based Word Sense Disambiguation**
*Sakae Mizuki and Naoaki Okazaki* 09:00-10:30 (Exhibit Hall)
A promising approach for knowledge-based Word Sense Disambiguation (WSD) is to select the sense whose contextualized embeddings computed for its definition sentence are closest to those computed for a target word in a given sentence. This approach relies on the similarity of the sense and context embeddings computed by a pre-trained language model. We propose a semantic specialization for WSD where contextualized embeddings are adapted to the WSD task using solely lexical knowledge. The key idea is, for a given sense, to bring semantically related senses and contexts closer and send different/unrelated senses farther away. We realize this idea as the joint optimization of the Attract-Repel objective for sense pairs and the self-training objective for context-sense pairs while controlling deviations from the original embeddings. The proposed method outperformed previous studies that adapt contextualized embeddings. It achieved state-of-the-art performance on knowledge-based WSD when combined with the reranking heuristic that uses the sense inventory. We found that the similarity characteristics of specialized embeddings conform to the key idea. We also found that the (dis)similarity of embeddings between the related/different/unrelated senses correlates well with the performance of WSD.

**MetaQA: Combining Expert Agents for Multi-Skill Question Answering**
*Haritz Puerto, Gözde Şahin and Iryna Gurevych* 09:00-10:30 (Exhibit Hall)
The recent explosion of question-answering (QA) datasets and models has increased the interest in the generalization of models across multiple domains and formats by either training on multiple datasets or combining multiple models. Despite the promising results of multi-dataset models, some domains or QA formats may require specific architectures, and thus the adaptability of these models might be limited. In addition, current approaches for combining models disregard cues such as question-answer compatibility. In this work, we propose to combine expert agents with a novel, flexible, and training-efficient architecture that considers questions, answer predictions, and answer-prediction confidence scores to select the best answer among a list of answer predictions. Through quantitative and qualitative experiments, we show that our model i) creates a collaboration between agents that outperforms previous multi-agent and multi-dataset approaches, ii) is highly data-efficient to train, and iii) can be adapted to any QA format. We release our code and a dataset of answer predictions from expert agents for 16 QA datasets to foster future research of multi-agent systems.

**In-Depth Look at Word Filling Societal Bias Measures**
*Matúš Pikuliak, Ivana Beňová and Viktor Bachratý* 09:00-10:30 (Exhibit Hall)
Many measures of societal bias in language models have been proposed in recent years. A popular approach is to use a set of word filling prompts to evaluate the behavior of the language models. In this work, we analyze the validity of two such measures – StereoSet and CrowS-Pairs. We show that these measures produce unexpected and illogical results when appropriate control group samples are constructed. Based on this, we believe that they are problematic and using them in the future should be reconsidered. We propose a way forward with an improved testing protocol. Finally, we also introduce a new gender bias dataset for Slovak.

**AbLit: A Resource for Analyzing and Generating Abridged Versions of English Literature**
*Melissa Roemmele, Kyle Shaffer, Katrina Olsen, Yiyi Wang and Steve Deneefe* 09:00-10:30 (Exhibit Hall)
Creating an abridged version of a text involves shortening it while maintaining its linguistic qualities. In this paper, we examine this task

from an NLP perspective for the first time. We present a new resource, AbLit, which is derived from abridged versions of English literature books. The dataset captures passage-level alignments between the original and abridged texts. We characterize the linguistic relations of these alignments, and create automated models to predict these relations as well as to generate abridgements for new texts. Our findings establish abridgement as a challenging task, motivating future resources and research. The dataset is available at github.com/roemmele/AbLit.

**Creation and evaluation of timelines for longitudinal user posts**
*Anthony Hills, Adam Tsakalidis, Federico Nanni, Ioannis Zachos and Maria Liakata* 09:00-10:30 (Exhibit Hall)
There is increasing interest to work with user generated content in social media, especially textual posts over time. Currently there is no consistent way of segmenting user posts into timelines in a meaningful way that improves the quality and cost of manual annotation. Here we propose a set of methods for segmenting longitudinal user posts into timelines likely to contain interesting moments of change in a user's behaviour, based on their online posting activity. We also propose a novel framework for evaluating timelines and show its applicability in the context of two different social media datasets. Finally, we present a discussion of the linguistic content of highly ranked timelines.

**GrIPS: Gradient-free, Edit-based Instruction Search for Prompting Large Language Models**
*Archiki Prasad, Peter Hase, Xiang Zhou and Mohit Bansal* 09:00-10:30 (Exhibit Hall)
Providing natural language instructions in prompts is a useful new paradigm for improving task performance of large language models in a zero-shot setting. Recent work has aimed to improve such prompts via manual rewriting or gradient-based tuning. However, manual rewriting is time-consuming and requires subjective interpretation, while gradient-based tuning can be extremely computationally demanding for large models and may not be feasible for API-based models. In this work, we introduce Gradient-free Instructional Prompt Search (GrIPS), a gradient-free, edit-based search approach for improving task instructions for large language models. GrIPS takes in instructions designed for humans and automatically returns an improved, edited prompt, while allowing for API-based tuning. With InstructGPT models, GrIPS improves the average task performance by up to 4.30 percentage points on eight classification tasks from the Natural Instructions dataset (with similar improvements for OPT, BLOOM, and FLAN-T5). We see improvements for both instruction-only prompts and instruction + k-shot examples prompts. Notably, GrIPS outperforms manual rewriting and purely example-based prompts while controlling for the available compute and data budget. Further, performance of GrIPS is comparable to select gradient-based tuning approaches. Qualitatively, we show our edits can simplify instructions and at times make them incoherent but nonetheless improve accuracy.

**Incorporating Question Answering-Based Signals into Abstractive Summarization via Salient Span Selection**
*Daniel Deutsch and Dan Roth* 09:00-10:30 (Exhibit Hall)
In this work, we propose a method for incorporating question-answering (QA) signals into a summarization model. Our method identifies salient noun phrases (NPs) in the input document by automatically generating wh-questions that are answered by the NPs and automatically determining whether those questions are answered in the gold summaries. This QA-based signal is incorporated into a two-stage summarization model which first marks salient NPs in the input document using a classification model, then conditionally generates a summary. Our experiments demonstrate that the models trained using QA-based supervision generate higher-quality summaries than baseline methods of identifying salient spans on benchmark summarization datasets. Further, we show that the content of the generated summaries can be controlled based on which NPs are marked in the input document. Finally, we propose a method of augmenting the training data so the gold summaries are more consistent with the marked input spans used during training and show how this results in models which learn to better exclude unmarked document content.

**LoRaLay: A Multilingual and Multimodal Dataset for Long Range and Layout-Aware Summarization**
*Laura Nguyen, Thomas Scialom, Benjamin Piwowarski and Jacopo Staiano* 09:00-10:30 (Exhibit Hall)
Text Summarization is a popular task and an active area of research for the Natural Language Processing community. By definition, it requires to account for long input texts, a characteristic which poses computational challenges for neural models. Moreover, real-world documents come in a variety of complex, visually-rich, layouts. This information is of great relevance, whether to highlight salient content or to encode long-range interactions between textual passages. Yet, all publicly available summarization datasets only provide plain text content. To facilitate research on how to exploit visual/layout information to better capture long-range dependencies in summarization models, we present LoRaLay, a collection of datasets for long-range summarization with accompanying visual/layout information. We extend existing and popular English datasets (arXiv and PubMed) with layout information and propose four novel datasets – consistently built from scholar resources – covering French, Spanish, Portuguese, and Korean languages. Further, we propose new baselines merging layout-aware and long-range models – two orthogonal approaches – and obtain state-of-the-art results, showing the importance of combining both lines of research.

**Combining Parameter-efficient Modules for Task-level Generalisation**
*Edoardo Maria Ponti, Alessandro Sordoni, Yoshua Bengio and Siva Reddy* 09:00-10:30 (Exhibit Hall)
A modular design encourages neural models to disentangle and recombine different facets of knowledge to generalise more systematically to new tasks. In this work, we assume that each task is associated with a subset of latent skills from an (arbitrary size) inventory. In turn, each skill corresponds to a parameter-efficient (sparse / low-rank) model adapter. By jointly learning adapters and a routing function that allocates skills to each task, the full network is instantiated as the average of the parameters of active skills. We propose several inductive biases that encourage re-usage and composition of the skills, including variable-size skill allocation and a dual-speed learning rate. We evaluate our latent-skill model in two main settings: 1) multitask reinforcement learning for instruction following on 8 levels of the BabyAI platform; and 2) few-shot fine-tuning of language models on 160 NLP tasks of the CrossFit benchmark. We find that the modular design of our network enhances sample efficiency in reinforcement learning and few-shot generalisation in supervised learning, compared to a series of baselines. These include models where parameters are fully shared, task-specific, conditionally generated (HyperFormer), or sparse mixture-of-experts (TaskMoE).

**Social Influence Dialogue Systems: A Survey of Datasets and Models For Social Influence Tasks**
*Kushal Chawla, Weiyan Shi, Jingwen Zhang, Gale Lucas, Zhou Yu and Jonathan Gratch* 09:00-10:30 (Exhibit Hall)
Dialogue systems capable of social influence such as persuasion, negotiation, and therapy, are essential for extending the use of technology to numerous realistic scenarios. However, existing research primarily focuses on either task-oriented or open-domain scenarios, a categorization that has been inadequate for capturing influence skills systematically. There exists no formal definition or category for dialogue systems with these skills and data-driven efforts in this direction are highly limited. In this work, we formally define and introduce the category of social influence dialogue systems that influence users' cognitive and emotional responses, leading to changes in thoughts, opinions, and behaviors through natural conversations. We present a survey of various tasks, datasets, and methods, compiling the progress across seven diverse domains. We discuss the commonalities and differences between the examined systems, identify limitations, and recommend future directions. This study serves as a comprehensive reference for social influence dialogue systems to inspire more dedicated research and discussion in this emerging area.

**A Two-Sided Discussion of Preregistration of NLP Research**
*Anders Søgaard, Daniel Hershcovich and Miryam De Lhoneux* 09:00-10:30 (Exhibit Hall)
Van Miltenburg et al. (2021) suggest NLP research should adopt preregistration to prevent fishing expeditions and to promote publication of

negative results. At face value, this is a very reasonable suggestion, seemingly solving many methodological problems with NLP research. We discuss pros and cons - some old, some new: a) Preregistration is challenged by the practice of retrieving hypotheses after the results are known; b) preregistration may bias NLP toward confirmatory research; c) preregistration must allow for reclassification of research as exploratory; d) preregistration may increase publication bias; e) preregistration may increase flag-planting; f) preregistration may increase p-hacking; and finally, g) preregistration may make us less risk tolerant. We cast our discussion as a dialogue, presenting both sides of the debate.

**NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages**
*Genta Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo and Pascale Fung* 09:00-10:30 (Exhibit Hall)
Natural language processing (NLP) has a significant impact on society via technologies such as machine translation and search engines. Despite its success, NLP technology is only widely available for high-resource languages such as English and Chinese, while it remains inaccessible to many languages due to the unavailability of data resources and benchmarks. In this work, we focus on developing resources for languages in Indonesia. Despite being the second most linguistically diverse country, most languages in Indonesia are categorized as endangered and some are even extinct. We develop the first-ever parallel resource for 10 low-resource languages in Indonesia. Our resource includes sentiment and machine translation datasets, and bilingual lexicons. We provide extensive analyses and describe challenges for creating such resources. We hope this work can spark NLP research on Indonesian and other underrepresented languages.

**Zero and Few-Shot Localization of Task-Oriented Dialogue Agents with a Distilled Representation**
*Mehrad Moradshahi, Sina Semnani and Monica Lam* 09:00-10:30 (Exhibit Hall)
Task-oriented Dialogue (ToD) agents are mostly limited to a few widely-spoken languages, mainly due to the high cost of acquiring training data for each language. Existing low-cost approaches that rely on cross-lingual embeddings or naive machine translation sacrifice a lot of accuracy for data efficiency, and largely fail in creating a usable dialogue agent. We propose automatic methods that use ToD training data in a source language to build a high-quality functioning dialogue agent in another target language that has no training data (i.e. zero-shot) or a small training set (i.e. few-shot). Unlike most prior work in cross-lingual ToD that only focuses on Dialogue State Tracking (DST), we build an end-to-end agent.
We show that our approach closes the accuracy gap between few-shot and existing full-shot methods for ToD agents. We achieve this by (1) improving the dialogue data representation, (2) improving entity-aware machine translation, and (3) automatic filtering of noisy translations.
We evaluate our approach on the recent bilingual dialogue dataset BiToD. In Chinese to English transfer, in the zero-shot setting, our method achieves 46.7% and 22.0% in Task Success Rate (TSR) and Dialogue Success Rate (DSR) respectively. In the few-shot setting where 10% of the data in the target language is used, we improve the state-of-the-art by 15.2% and 14.0%, coming within 5% of full-shot training.

**Contextual Semantic Parsing for Multilingual Task-Oriented Dialogues**
*Mehrad Moradshahi, Victoria Tsai, Giovanni Campagna and Monica Lam* 09:00-10:30 (Exhibit Hall)
Robust state tracking for task-oriented dialogue systems currently remains restricted to a few popular languages. This paper shows that given a large-scale dialogue data set in one language, we can automatically produce an effective semantic parser for other languages using machine translation. We propose automatic translation of dialogue datasets with alignment to ensure faithful translation of slot values and eliminate costly human supervision used in previous benchmarks. We also propose a new contextual semantic parsing model, which encodes the formal slots and values, and only the last agent and user utterances. We show that the succinct representation reduces the compounding effect of translation errors, without harming the accuracy in practice.
We evaluate our approach on several dialogue state tracking benchmarks. On RiSAWOZ, CrossWOZ, CrossWOZ-EN, and MultiWOZ-ZH datasets we improve the state of the art by 11%, 17%, 20%, and 0.3% in joint goal accuracy. We present a comprehensive error analysis for all three datasets showing erroneous annotations can lead to misguided judgments on the quality of the model.
Finally, we present RiSAWOZ English and German datasets, created using our translation methodology. On these datasets, accuracy is within 11% of the original showing that high-accuracy multilingual dialogue datasets are possible without relying on expensive human annotations. We release our datasets and software open source.

**Evaluating and Improving the Coreference Capabilities of Machine Translation Models**
*Asaf Yehudai, Arie Cattan, Omri Abend and Gabriel Stanovsky* 09:00-10:30 (Exhibit Hall)
Machine translation (MT) requires a wide range of linguistic capabilities, which current end-to-end models are expected to learn implicitly by observing aligned sentences in bilingual corpora. In this work, we ask: \emph{How well MT models learn coreference resolution via implicit signal?} To answer this question, we develop an evaluation methodology that derives coreference clusters from MT output and evaluates them without requiring annotations in the target language. Following, we evaluate several prominent open-source and commercial MT systems, translating from English to six target languages, and compare them to state-of-the-art coreference resolvers on three challenging benchmarks. Our results show that the monolingual resolvers greatly outperform MT models. Motivated by this result, we experiment with different methods for incorporating the output of coreference resolution models in MT, showing improvement over strong baselines.

**Document-Level Planning for Text Simplification**
*Liam Cripwell, Joël Legrand and Claire Gardent* 09:00-10:30 (Exhibit Hall)
Most existing work on text simplification is limited to sentence-level inputs, with attempts to iteratively apply these approaches to document-level simplification failing to coherently preserve the discourse structure of the document. We hypothesise that by providing a high-level view of the target document, a simplification plan might help to guide generation. Building upon previous work on controlled, sentence-level simplification, we view a plan as a sequence of labels, each describing one of four sentence-level simplification operations (copy, rephrase, split, or delete). We propose a planning model that labels each sentence in the input document while considering both its context (a window of surrounding sentences) and its internal structure (a token-level representation). Experiments on two simplification benchmarks (Newsela-auto and Wiki-auto) show that our model outperforms strong baselines both on the planning task and when used to guide document-level simplification models.

**WinoDict: Probing language models for in-context word acquisition**
*Julian Eisenschlos, Jeremy Cole, Fangyu Liu and William Cohen* 09:00-10:30 (Exhibit Hall)
We introduce a new in-context learning paradigm to measure Large Language Models' (LLMs) ability to learn novel words during inference. In particular, we rewrite Winograd-style co-reference resolution problems by replacing the key concept word with a synthetic but plausible word that the model must understand to complete the task. Solving this task requires the model to make use of the dictionary definition of the new word given in the prompt. This benchmark addresses word acquisition, one important aspect of the diachronic degradation known to afflict LLMs. As LLMs are frozen in time at the moment they are trained, they are normally unable to reflect the way language changes over time. We show that the accuracy of LLMs compared to the original Winograd tasks decreases radically in our benchmark, thus identifying a limitation of current models and providing a benchmark to measure future improvements in LLMs ability to do in-context learning.

**Looking for a Needle in a Haystack: A Comprehensive Study of Hallucinations in Neural Machine Translation**
*Nuno M. Guerreiro, Elena Voita and André Martins* 09:00-10:30 (Exhibit Hall)

Although the problem of hallucinations in neural machine translation (NMT) has received some attention, research on this highly pathological phenomenon lacks solid ground. Previous work has been limited in several ways: it often resorts to artificial settings where the problem is amplified, it disregards some (common) types of hallucinations, and it does not validate adequacy of detection heuristics. In this paper, we set foundations for the study of NMT hallucinations. First, we work in a natural setting, i.e., in-domain data without artificial noise neither in training nor in inference. Next, we annotate a dataset of over 3.4k sentences indicating different kinds of critical errors and hallucinations. Then, we turn to detection methods and both revisit methods used previously and propose using glass-box uncertainty-based detectors. Overall, we show that for preventive settings, (i) previously used methods are largely inadequate, (ii) sequence log-probability works best and performs on par with reference-based methods. Finally, we propose DeHallucinator, a simple method for alleviating hallucinations at test time that significantly reduces the hallucinatory rate.

**Visual Writing Prompts: Character-Grounded Story Generation with Curated Image Sequences**
*Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg and Bernt Schiele* 09:00-10:30 (Exhibit Hall)
Current work on image-based story generation suffers from the fact that the existing image sequence collections do not have coherent plots behind them. We improve visual story generation by producing a new image-grounded dataset, Visual Writing Prompts (VWP). VWP contains almost 2K selected sequences of movie shots, each including 5-10 images. The image sequences are aligned with a total of 12K stories which were collected via crowdsourcing given the image sequences and a set of grounded characters from the corresponding image sequence. Our new image sequence collection and filtering process has allowed us to obtain stories that are more coherent and have more narrativity compared to previous work. We also propose a character-based story generation model driven by coherence as a strong baseline. Evaluations show that our generated stories are more coherent, visually grounded, and have more narrativity than stories generated with the current state-of-the-art model.

**Nationality Bias in Text Generation**
*Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-hao Huang and Shomir Wilson* 09:00-10:30 (Exhibit Hall)
Little attention is placed on analyzing nationality bias in language models, especially when nationality is highly used as a factor in increasing the performance of social NLP models. This paper examines how a text generation model, GPT-2, accentuates pre-existing societal biases about country-based demonyms. We generate stories using GPT-2 for various nationalities and use sensitivity analysis to explore how the number of internet users and the country's economic status impacts the sentiment of the stories. To reduce the propagation of biases through large language models (LLM), we explore the debiasing method of adversarial triggering. Our results show that GPT-2 demonstrates significant bias against countries with lower internet users, and adversarial triggering effectively reduces the same.

**Multilingual Normalization of Temporal Expressions with Masked Language Models**
*Lukas Lange, Jannik Strötgen, Heike Adel and Dietrich Klakow* 09:00-10:30 (Exhibit Hall)
The detection and normalization of temporal expressions is an important task and preprocessing step for many applications. However, prior work on normalization is rule-based, which severely limits the applicability in real-world multilingual settings, due to the costly creation of new rules. We propose a novel neural method for normalizing temporal expressions based on masked language modeling. Our multilingual method outperforms prior rule-based systems in many languages, and in particular, for low-resource languages with performance improvements of up to 33 F1 on average compared to the state of the art.

**Conversational Tree Search: A New Hybrid Dialog Task**
*Dirk Väth, Lindsey Vanderlyn and Ngoc Thang Vu* 09:00-10:30 (Exhibit Hall)
Conversational interfaces provide a flexible and easy way for users to seek information that may otherwise be difficult or inconvenient to obtain. However, existing interfaces generally fall into one of two categories: FAQs, where users must have a concrete question in order to retrieve a general answer, or dialogs, where users must follow a pre-defined path but may receive a personalized answer. In this paper, we introduce Conversational Tree Search (CTS) as a new task that bridges the gap between FAQ-style information retrieval and task-oriented dialog, allowing domain-experts to define dialog trees which can then be converted to an efficient dialog policy that learns only to ask the questions necessary to navigate a user to their goal. We collect a dataset for the travel reimbursement domain and demonstrate a baseline as well as a novel deep Reinforcement Learning architecture for this task. Our results show that the new architecture combines the positive aspects of both the FAQ and dialog system used in the baseline and achieves higher goal completion while skipping unnecessary questions.

**Robustification of Multilingual Language Models to Real-world Noise in Crosslingual Zero-shot Settings with Robust Contrastive Pretraining**
*Asa Cooper Stickland, Sailik Sengupta, Jason Krone, Saab Mansour and He He* 09:00-10:30 (Exhibit Hall)
Advances in neural modeling have achieved state-of-the-art (SOTA) results on public natural language processing (NLP) benchmarks, at times surpassing human performance. However, there is a gap between public benchmarks and real-world applications where noise, such as typographical or grammatical mistakes, is abundant and can result in degraded performance. Unfortunately, works which evaluate the robustness of neural models on noisy data and propose improvements, are limited to the English language. Upon analyzing noise in different languages, we observe that noise types vary greatly across languages. Thus, existing investigations do not generalize trivially to multilingual settings. To benchmark the performance of pretrained multilingual language models, we construct noisy datasets covering five languages and four NLP tasks and observe a clear gap in the performance between clean and noisy data in the zero-shot cross-lingual setting. After investigating several ways to boost the robustness of multilingual models in this setting, we propose Robust Contrastive Pretraining (RCP). RCP combines data augmentation with a contrastive loss term at the pretraining stage and achieves large improvements on noisy (and original test data) across two sentence-level (+3.2%) and two sequence-labeling (+10 F1-score) multilingual classification tasks.

**Made of Steel? Learning Plausible Materials for Components in the Vehicle Repair Domain**
*Annerose Eichel, Helena Schlipf and Sabine Schulte Im Walde* 09:00-10:30 (Exhibit Hall)
We propose a novel approach to learn domain-specific plausible materials for components in the vehicle repair domain by probing Pretrained Language Models (PLMs) in a cloze task style setting to overcome the lack of annotated datasets. We devise a new method to aggregate salient predictions from a set of cloze query templates and show that domain-adaptation using either a small, high-quality or a customized Wikipedia corpus boosts performance. When exploring resource-lean alternatives, we find a distilled PLM clearly outperforming a classic pattern-based algorithm. Further, given that 98% of our domain-specific components are multiword expressions, we successfully exploit the compositionality assumption as a way to address data sparsity.

**Semantic Frame Induction with Deep Metric Learning**
*Kosuke Yamada, Ryohei Sasano and Koichi Takeda* 09:00-10:30 (Exhibit Hall)
Recent studies have demonstrated the usefulness of contextualized word embeddings in unsupervised semantic frame induction. However, they have also revealed that generic contextualized embeddings are not always consistent with human intuitions about semantic frames, which causes unsatisfactory performance for frame induction based on contextualized embeddings. In this paper, we address supervised semantic frame induction, which assumes the existence of frame-annotated data for a subset of predicates in a corpus and aims to build a frame induction model that leverages the annotated data. We propose a model that uses deep metric learning to fine-tune a contextualized embed-

ding model, and we apply the fine-tuned contextualized embeddings to perform semantic frame induction. Our experiments on FrameNet show that fine-tuning with deep metric learning considerably improves the clustering evaluation scores, namely, the B-cubed F-score and Purity F-score, by about 8 points or more. We also demonstrate that our approach is effective even when the number of training instances is small.

**LEALLA: Learning Lightweight Language-agnostic Sentence Embeddings with Knowledge Distillation**
*Zhuoyuan Mao and Tetsuji Nakagawa* 09:00-10:30 (Exhibit Hall)
Large-scale language-agnostic sentence embedding models such as LaBSE (Feng et al., 2022) obtain state-of-the-art performance for parallel sentence alignment. However, these large-scale models can suffer from inference speed and computation overhead. This study systematically explores learning language-agnostic sentence embeddings with lightweight models. We demonstrate that a thin-deep encoder can construct robust low-dimensional sentence embeddings for 109 languages. With our proposed distillation methods, we achieve further improvements by incorporating knowledge from a teacher model. Empirical results on Tatoeba, United Nations, and BUCC show the effectiveness of our lightweight models. We release our lightweight language-agnostic sentence embedding models LEALLA on TensorFlow Hub.

**Span-based Named Entity Recognition by Generating and Compressing Information**
*Nhung Nguyen, Makoto Miwa and Sophia Ananiadou* 09:00-10:30 (Exhibit Hall)
The information bottleneck (IB) principle has been proven effective in various NLP applications. The existing work, however, only used either generative or information compression models to improve the performance of the target task. In this paper, we propose to combine the two types of IB models into one system to enhance Named Entity Recognition (NER). For one type of IB model, we incorporate two unsupervised generative components, span reconstruction and synonym generation, into a span-based NER system. The span reconstruction ensures that the contextualised span representation keeps the span information, while the synonym generation makes synonyms have similar representations even in different contexts. For the other type of IB model, we add a supervised IB layer that performs information compression into the system to preserve useful features for NER in the resulting span representations. Experiments on five different corpora indicate that jointly training both generative and information compression models can enhance the performance of the baseline span-based NER system. Our source code is publicly available at https://github.com/nguyennth/joint-ib-models.

**Pento-DIARef: A Diagnostic Dataset for Learning the Incremental Algorithm for Referring Expression Generation from Examples**
*Philipp Sadler and David Schlangen* 09:00-10:30 (Exhibit Hall)
NLP tasks are typically defined extensionally through datasets containing example instantiations. We present Pento-DIARef, a diagnostic dataset in a visual domain of puzzle pieces where referring expressions are generated by a well-known symbolic algorithm (the "Incremental Algorithm"), which itself is motivated by appeal to a hypothesised capability (eliminating distractors through application of Gricean maxims). Our question then is whether the extensional description (the dataset) is sufficient for a neural model to pick up the underlying regularity and exhibit this capability given the simple task definition of producing expressions from visual inputs. We find that a model supported by a vision detection step and a targeted data generation scheme achieves an almost perfect BLEU@1 score and sentence accuracy, whereas simpler baselines do not.

**Pento-DIARef: A Diagnostic Dataset for Learning the Incremental Algorithm for Referring Expression Generation from Examples**
*Philipp Sadler and David Schlangen* 09:00-10:30 (Exhibit Hall)
NLP tasks are typically defined extensionally through datasets containing example instantiations. We present Pento-DIARef, a diagnostic dataset in a visual domain of puzzle pieces where referring expressions are generated by a well-known symbolic algorithm (the "Incremental Algorithm"), which itself is motivated by appeal to a hypothesised capability (eliminating distractors through application of Gricean maxims). Our question then is whether the extensional description (the dataset) is sufficient for a neural model to pick up the underlying regularity and exhibit this capability given the simple task definition of producing expressions from visual inputs. We find that a model supported by a vision detection step and a targeted data generation scheme achieves an almost perfect BLEU@1 score and sentence accuracy, whereas simpler baselines do not.

**Shortcomings of Question Answering Based Factuality Frameworks for Error Localization**
*Ryo Kamoi, Tanya Goyal and Greg Durrett* 09:00-10:30 (Exhibit Hall)
Despite recent progress in abstractive summarization, models often generate summaries with factual errors. Numerous approaches to detect these errors have been proposed, the most popular of which are question answering (QA)-based factuality metrics. These have been shown to work well at predicting summary-level factuality and have potential to localize errors within summaries, but this latter capability has not been systematically evaluated in past research. In this paper, we conduct the first such analysis and find that, contrary to our expectations, QA-based frameworks fail to correctly identify error spans in generated summaries and are outperformed by trivial exact match baselines. Our analysis reveals a major reason for such poor localization: questions generated by the QG module often inherit errors from non-factual summaries which are then propagated further into downstream modules. Moreover, even human-in-the-loop question generation cannot easily offset these problems. Our experiments conclusively show that there exist fundamental issues with localization using the QA framework which cannot be fixed solely by stronger QA and QG models.

**Towards a Unified Multi-Domain Multilingual Named Entity Recognition Model**
*Mayank Kulkarni, Daniel Preotiuc-pietro, Karthik Radhakrishnan, Genta Winata, Shijie Wu, Lingjue Xie and Shaohua Yang* 09:00-10:30 (Exhibit Hall)
Named Entity Recognition is a key Natural Language Processing task whose performance is sensitive to choice of genre and language. A unified NER model across multiple genres and languages is more practical and efficient by leveraging commonalities across genres or languages. In this paper, we propose a novel setup for NER which includes multi-domain and multilingual training and evaluation across 13 domains and 4 languages. We explore a range of approaches to building a unified model using domain and language adaptation techniques. Our experiments highlight multiple nuances to consider while building a unified model, including that naive data pooling fails to obtain good performance, that domain-specific adaptations are more important than language-specific ones and that including domain-specific adaptations in a unified model nears the performance of training multiple dedicated monolingual models at a fraction of their parameter count.

**A Psycholinguistic Analysis of BERT's Representations of Compounds**
*Lars Buijtelaar and Sandro Pezzelle* 09:00-10:30 (Exhibit Hall)
This work studies the semantic representations learned by BERT for compounds, that is, expressions such as sunlight or bodyguard. We build on recent studies that explore semantic information in Transformers at the word level and test whether BERT aligns with human semantic intuitions when dealing with expressions (e.g., sunlight) whose overall meaning depends—to a various extent—on the semantics of the constituent words (sun, light). We leverage a dataset that includes human judgments on two psycholinguistic measures of compound semantic analysis: lexeme meaning dominance (LMD; quantifying the weight of each constituent toward the compound meaning) and semantic transparency (ST; evaluating the extent to which the compound meaning is recoverable from the constituents' semantics). We show that BERT-based measures moderately align with human intuitions, especially when using contextualized representations, and that LMD is overall

more predictable than ST. Contrary to the results reported for 'standard' words, higher, more contextualized layers are the best at representing compound meaning. These findings shed new light on the abilities of BERT in dealing with fine-grained semantic phenomena. Moreover, they can provide insights into how speakers represent compounds.

**Measuring Normative and Descriptive Biases in Language Models Using Census Data**
*Samia Touileb, Lilja Øvrelid and Erik Velldal* 09:00-10:30 (Exhibit Hall)
We investigate in this paper how distributions of occupations with respect to gender is reflected in pre-trained language models. Such distributions are not always aligned to normative ideals, nor do they necessarily reflect a descriptive assessment of reality. In this paper, we introduce an approach for measuring to what degree pre-trained language models are aligned to normative and descriptive occupational distributions. To this end, we use official demographic information about gender–occupation distributions provided by the national statistics agencies of France, Norway, United Kingdom, and the United States. We manually generate template-based sentences combining gendered pronouns and nouns with occupations, and subsequently probe a selection of ten language models covering the English, French, and Norwegian languages. The scoring system we introduce in this work is language independent, and can be used on any combination of template-based sentences, occupations, and languages. The approach could also be extended to other dimensions of national census data and other demographic variables.

**Shapley Head Pruning: Identifying and Removing Interference in Multilingual Transformers**
*William Held and Diyi Yang* 09:00-10:30 (Exhibit Hall)
Multilingual transformer-based models demonstrate remarkable zero and few-shot transfer across languages by learning and reusing language-agnostic features. However, as a fixed-size model acquires more languages, its performance across all languages degrades. Those who attribute this interference phenomenon to limited model capacity address the problem by adding additional parameters, despite evidence that transformer-based models are overparameterized. In this work, we show that it is possible to reduce interference by instead identifying and pruning language-specific attention heads. First, we use Shapley Values, a credit allocation metric from coalitional game theory, to identify attention heads that introduce interference. Then, we show that pruning such heads from a fixed model improves performance for a target language on both sentence classification and structural prediction. Finally, we provide insights on language-agnostic and language-specific attention heads using attention visualization.

**Why Don't You Do It Right? Analysing Annotators' Disagreement in Subjective Tasks**
*Marta Sandri, Elisa Leonardelli, Sara Tonelli and Elisabetta Jezek* 09:00-10:30 (Exhibit Hall)
Annotators' disagreement in linguistic data has been recently the focus of multiple initiatives aimed at raising awareness on issues related to 'majority voting' when aggregating diverging annotations. Disagreement can indeed reflect different aspects of linguistic annotation, from annotators' subjectivity to sloppiness or lack of enough context to interpret a text. In this work we first propose a taxonomy of possible reasons leading to annotators' disagreement in subjective tasks. Then, we manually label part of a Twitter dataset for offensive language detection in English following this taxonomy, identifying how the different categories are distributed. Finally we run a set of experiments aimed at assessing the impact of the different types of disagreement on classification performance. In particular, we investigate how accurately tweets belonging to different categories of disagreement can be classified as offensive or not, and how injecting data with different types of disagreement in the training set affects performance. We also perform offensive language detection as a multi-task framework, using disagreement classification as an auxiliary task.

**How people talk about each other: Modeling Generalized Intergroup Bias and Emotion**
*Venkata Subrahmanyan Govindarajan, Katherine Atwell, Barea Sinno, Malihe Alikhani, David Beaver and Junyi Jessy Li* 09:00-10:30 (Exhibit Hall)
Current studies of bias in NLP rely mainly on identifying (unwanted or negative) bias towards a specific demographic group. While this has led to progress recognizing and mitigating negative bias, and having a clear notion of the targeted group is necessary, it is not always practical. In this work we extrapolate to a broader notion of bias, rooted in social science and psychology literature. We move towards predicting interpersonal group relationship (IGR) - modeling the relationship between the speaker and the target in an utterance - using fine-grained interpersonal emotions as an anchor. We build and release a dataset of English tweets by US Congress members annotated for interpersonal emotion - the first of its kind, and 'found supervision' for IGR labels; our analyses show that subtle emotional signals are indicative of different biases. While humans can perform better than chance at identifying IGR given an utterance, we show that neural models perform much better; furthermore, a shared encoding between IGR and interpersonal perceived emotion enabled performance gains in both tasks.

**ComSearch: Equation Searching with Combinatorial Strategy for Solving Math Word Problems with Weak Supervision**
*Qianying Liu, Wenyu Guan, Jianhao Shen, Fei Cheng and Sadao Kurohashi* 09:00-10:30 (Exhibit Hall)
Previous studies have introduced a weakly-supervised paradigm for solving math word problems requiring only the answer value annotation. While these methods search for correct value equation candidates as pseudo labels, they search among a narrow sub-space of the enormous equation space. To address this problem, we propose a novel search algorithm with combinatorial strategy ComSearch, which can compress the search space by excluding mathematically equivalent equations. The compression allows the searching algorithm to enumerate all possible equations and obtain high-quality data. We investigate the noise in the pseudo labels that hold wrong mathematical logic, which we refer to as the false-matching problem, and propose a ranking model to denoise the pseudo labels. Our approach holds a flexible framework to utilize two existing supervised math word problem solvers to train pseudo labels, and both achieve state-of-the-art performance in the weak supervision task.

**Towards preserving word order importance through Forced Invalidation**
*Hadeel Al-negheimish, Pranava Madhyastha and Alessandra Russo* 09:00-10:30 (Exhibit Hall)
Large pre-trained language models such as BERT have been widely used as a framework for natural language understanding (NLU) tasks. However, recent findings have revealed that pre-trained language models are insensitive to word order. The performance on NLU tasks remains unchanged even after randomly permuting the word of a sentence, where crucial syntactic information is destroyed. To help preserve the importance of word order, we propose a simple approach called Forced Invalidation (FI): forcing the model to identify permuted sequences as invalid samples. We perform an extensive evaluation of our approach on various English NLU and QA based tasks over BERT-based and attention-based models over word embeddings. Our experiments demonstrate that FI significantly improves the sensitivity of the models to word order.

# Parallel Session 7 - 11:15-12:45

## Session 7 Orals – Language Grounding and Multi-Modality – Room C

11:15-12:45 (Elafiti 4)

**COVID-VTS: Fact Extraction and Verification on Short Video Platforms**
*Fuxiao Liu, Yaser Yacoob and Abhinav Shrivastava* 11:15-11:30 (Elafiti 4)
We introduce a new benchmark, COVID-VTS, for fact-checking multi-modal information involving short-duration videos with COVID19-focused information from both the real world and machine generation. We propose, TwtrDetective, an effective model incorporating cross-media consistency checking to detect token-level malicious tampering in different modalities, and generate explanations. Due to the scarcity of training data, we also develop an efficient and scalable approach to automatically generate misleading video posts by event manipulation or adversarial matching. We investigate several state-of-the-art models and demonstrate the superiority of TwtrDetective.

**Learning the Legibility of Visual Text Perturbations**
*Dev Seth, Rickard Stureborg, Danish Pruthi and Bhuwan Dhingra* 11:30-11:45 (Elafiti 4)
Many adversarial attacks in NLP perturb text in puts to produce visually similar strings which are legible to humans but degrade model performance. Although preserving legibility is a necessary condition for text perturbation, little work has been done to systematically characterize it; instead, legibility is typically loosely enforced via intuitions around the nature and extent of perturbations. Particularly, it is unclear to what extent can inputs be perturbed while preserving legibility, or how to quantify the legibility of a perturbed string. In this work, we address this gap by learning models that predict the legibility of a perturbed string, and rank candidate perturbations based on their legibility. To do so, we collect and release LEGIT, a human-annotated dataset comprising the legibility of visually perturbed text. Using this dataset, we build both text- and vision-based models which achieve up to 0.91 F score in predicting whether an input is legible, and an accuracy of 0.86 in predicting which of two given perturbations is more legible. Additionally, we discover that legible perturbations from the LEGIT dataset are more effective at lowering the performance of NLP models than best-known attack strategies, suggesting that current models may be vulnerable to a broad range of perturbations beyond what is captured by existing visual attacks.

**MAPL: Parameter-Efficient Adaptation of Unimodal Pre-Trained Models for Vision-Language Few-Shot Prompting**
*Oscar Mañas, Pau Rodriguez Lopez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal and Aishwarya Agrawal* 11:45-12:00 (Elafiti 4)
Large pre-trained models have proved to be remarkable zero- and (prompt-based) few-shot learners in unimodal vision and language tasks. We propose MAPL, a simple and parameter-efficient method that reuses frozen pre-trained unimodal models and leverages their strong generalization capabilities in multimodal vision-language (VL) settings. MAPL learns a lightweight mapping between the representation spaces of unimodal models using aligned image-text data, and can generalize to unseen VL tasks from just a few in-context examples. The small number of trainable parameters makes MAPL effective at low-data and in-domain learning. Moreover, MAPL's modularity enables easy extension to other pre-trained models. Extensive experiments on several visual question answering and image captioning benchmarks show that MAPL achieves superior or competitive performance compared to similar methods while training orders of magnitude fewer parameters. MAPL can be trained in just a few hours using modest computational resources and public datasets. We release our code and pre-trained model weights at https://github.com/oscmansan/mapl.

**Plausible May Not Be Faithful: Probing Object Hallucination in Vision-Language Pre-training**
*Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su and Pascale Fung* 12:00-12:15 (Elafiti 4)
Large-scale vision-language pre-trained (VLP) models are prone to hallucinate non-existent visual objects when generating text based on visual information. In this paper, we systematically study the object hallucination problem from three aspects. First, we examine recent state-of-the-art VLP models, showing that they still hallucinate frequently and models achieving better scores on standard metrics (e.g., CIDEr) could be more unfaithful. Second, we investigate how different types of image encoding in VLP influence hallucination, including region-based, grid-based, and patch-based. Surprisingly, we find that patch-based features perform the best and smaller patch resolution yields a non-trivial reduction in object hallucination. Third, we decouple various VLP objectives and demonstrate that token-level image-text alignment and controlled generation are crucial to reducing hallucination. Based on that, we propose a simple yet effective VLP loss named ObjMLM to further mitigate object hallucination. Results show that it reduces object hallucination by up to 17.4% when tested on two benchmarks (COCO Caption for in-domain and NoCaps for out-of-domain evaluation).

**Know your audience: specializing grounded language models with listener subtraction**
*Aaditya Singh, David Ding, Andrew Saxe, Felix Hill and Andrew Lampinen* 12:15-12:30 (Elafiti 4)
Effective communication requires adapting to the idiosyncrasies of each communicative context—such as the common ground shared with each partner. Humans demonstrate this ability to specialize to their audience in many contexts, such as the popular game Dixit. We take inspiration from Dixit to formulate a multi-agent image reference game where a (trained) speaker model is rewarded for describing a target image such that one (pretrained) listener model can correctly identify it among distractors, but another listener cannot. To adapt, the speaker must exploit differences in the knowledge it shares with the different listeners. We show that finetuning an attention-based adapter between a CLIP vision encoder and a large language model in this contrastive, multi-agent setting gives rise to context-dependent natural language specialization from rewards only, without direct supervision. Through controlled experiments, we show that training a speaker with two listeners that perceive differently, using our method, allows the speaker to adapt to the idiosyncracies of the listeners. Furthermore, we show zero-shot transfer of the specialization to real-world data. Our experiments demonstrate a method for specializing grounded language models without direct supervision and highlight the interesting research challenges posed by complex multi-agent communication.

**Improving Cross-modal Alignment for Text-Guided Image Inpainting**
*Yucheng Zhou and Guodong Long* 12:30-12:45 (Elafiti 4)
Text-guided image inpainting (TGII) aims to restore missing regions based on a given text in a damaged image. Existing methods are based on a strong vision encoder and a cross-modal fusion model to integrate cross-modal features. However, these methods allocate most of the computation to visual encoding, while light computation on modeling modality interactions. Moreover, they take cross-modal fusion for depth features, which ignores a fine-grained alignment between text and image. Recently, vision-language pre-trained models (VLPM), encapsulating rich cross-modal alignment knowledge, have advanced in most multimodal tasks. In this work, we propose a novel model for TGII by improving cross-modal alignment (CMA). CMA model consists of a VLPM as a vision-language encoder, an image generator and global-local discriminators. To explore cross-modal alignment knowledge for image restoration, we introduce cross-modal alignment distillation and in-sample distribution distillation. In addition, we employ adversarial training to enhance the model to fill the missing region in complicated structures effectively. Experiments are conducted on two popular vision-language datasets. Results show that our model achieves state-of-the-art performance compared with other strong competitors.

## Session 7 Orals – Language Resources and Evaluation 2 – Room A

11:15-12:45 (Elafiti 2)

**A Human Subject Study of Named Entity Recognition in Conversational Music Recommendation Queries**
*Elena Epure and Romain Hennequin* 11:15-11:30 (Elafiti 2)
We conducted a human subject study of named entity recognition on a noisy corpus of conversational music recommendation queries, with many irregular and novel named entities. We evaluated the human NER linguistic behaviour in these challenging conditions and compared it with the most common NER systems nowadays, fine-tuned transformers. Our goal was to learn about the task to guide the design of better evaluation methods and NER algorithms. The results showed that NER in our context was quite hard for both human and algorithms under a strict evaluation schema; humans had higher precision, while the model higher recall because of entity exposure especially during pre-training; and entity types had different error patterns (e.g. frequent typing errors for artists). The released corpus goes beyond predefined frames of interaction and can support future work in conversational music recommendation.

**DiscoScore: Evaluating Text Generation with BERT and Discourse Coherence**
*Wei Zhao, Michael Strube and Steffen Eger* 11:30-11:45 (Elafiti 2)
Recently, there has been a growing interest in designing text generation systems from a discourse coherence perspective, e.g., modeling the interdependence between sentences. Still, recent BERT-based evaluation metrics are weak in recognizing coherence, and thus are not reliable in a way to spot the discourse-level improvements of those text generation systems. In this work, we introduce DiscoScore, a parametrized discourse metric, which uses BERT to model discourse coherence from different perspectives, driven by Centering theory. Our experiments encompass 16 non-discourse and discourse metrics, including DiscoScore and popular coherence models, evaluated on summarization and document-level machine translation (MT). We find that (i) the majority of BERT-based metrics correlate much worse with human rated coherence than early discourse metrics, invented a decade ago; (ii) the recent state-of-the-art BARTScore is weak when operated at system level—which is particularly problematic as systems are typically compared in this manner. DiscoScore, in contrast, achieves strong system-level correlation with human ratings, not only in coherence but also in factual consistency and other aspects, and surpasses BARTScore by over 10 correlation points on average. Further, aiming to understand DiscoScore, we provide justifications to the importance of discourse coherence for evaluation metrics, and explain the superiority of one variant over another. Our code is available at \url{https://github.com/AIPHES/DiscoScore}.

**Large Scale Multi-Lingual Multi-Modal Summarization Dataset**
*Yash Verma, Anubhav Jangra, Raghvendra Verma and Sriparna Saha* 11:45-12:00 (Elafiti 2)
Significant developments in techniques such as encoder-decoder models have enabled us to represent information comprising multiple modalities. This information can further enhance many downstream tasks in the field of information retrieval and natural language processing; however, improvements in multi-modal techniques and their performance evaluation require large-scale multi-modal data which offers sufficient diversity. Multi-lingual modeling for a variety of tasks like multi-modal summarization, text generation, and translation leverages information derived from high-quality multi-lingual annotated data. In this work, we present the current largest multi-lingual multi-modal summarization dataset (M3LS), and it consists of over a million instances of document-image pairs along with a professionally annotated multi-modal summary for each pair. It is derived from news articles published by British Broadcasting Corporation(BBC) over a decade and spans 20 languages, targeting diversity across five language roots, it is also the largest summarization dataset for 13 languages and consists of cross-lingual summarization data for 2 languages. We formally define the multi-lingual multi-modal summarization task utilizing our dataset and report baseline scores from various state-of-the-art summarization techniques in a multi-lingual setting. We also compare it with many similar datasets to analyze the uniqueness and difficulty of M3LS. The dataset and code used in this work are made available at "https://github.com/anubhav-jangra/M3LS".

**LoRaLay: A Multilingual and Multimodal Dataset for Long Range and Layout-Aware Summarization**
*Laura Nguyen, Thomas Scialom, Benjamin Piwowarski and Jacopo Staiano* 12:00-12:15 (Elafiti 2)
Text Summarization is a popular task and an active area of research for the Natural Language Processing community. By definition, it requires to account for long input texts, a characteristic which poses computational challenges for neural models. Moreover, real-world documents come in a variety of complex, visually-rich, layouts. This information is of great relevance, whether to highlight salient content or to encode long-range interactions between textual passages. Yet, all publicly available summarization datasets only provide plain text content. To facilitate research on how to exploit visual/layout information to better capture long-range dependencies in summarization models, we present LoRaLay, a collection of datasets for long-range summarization with accompanying visual/layout information. We extend existing and popular English datasets (arXiv and PubMed) with layout information and propose four novel datasets – consistently built from scholar resources – covering French, Spanish, Portuguese, and Korean languages. Further, we propose new baselines merging layout-aware and long-range models – two orthogonal approaches – and obtain state-of-the-art results, showing the importance of combining both lines of research.

**NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages**
*Genta Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo and Pascale Fung* 12:15-12:30 (Elafiti 2)
Natural language processing (NLP) has a significant impact on society via technologies such as machine translation and search engines. Despite its success, NLP technology is only widely available for high-resource languages such as English and Chinese, while it remains inaccessible to many languages due to the unavailability of data resources and benchmarks. In this work, we focus on developing resources for languages in Indonesia. Despite being the second most linguistically diverse country, most languages in Indonesia are categorized as endangered and some are even extinct. We develop the first-ever parallel resource for 10 low-resource languages in Indonesia. Our resource includes sentiment and machine translation datasets, and bilingual lexicons. We provide extensive analyses and describe challenges for creating such resources. We hope this work can spark NLP research on Indonesian and other underrepresented languages.

**UScore: An Effective Approach to Fully Unsupervised Evaluation Metrics for Machine Translation**
*Jonas Belouadi and Steffen Eger* 12:30-12:45 (Elafiti 2)
The vast majority of evaluation metrics for machine translation are supervised, i.e., (i) are trained on human scores, (ii) assume the existence of reference translations, or (iii) leverage parallel data. This hinders their applicability to cases where such supervision signals are not available. In this work, we develop fully unsupervised evaluation metrics. To do so, we leverage similarities and synergies between evaluation metric induction, parallel corpus mining, and MT systems. In particular, we use an unsupervised evaluation metric to mine pseudo-parallel data, which we use to remap deficient underlying vector spaces (in an iterative manner) and to induce an unsupervised MT system, which then provides pseudo-references as an additional component in the metric. Finally, we also induce unsupervised multilingual sentence embeddings from pseudo-parallel data. We show that our fully unsupervised metrics are effective, i.e., they beat supervised competitors on 4 out of our 5 evaluation datasets. We make our code publicly available.

## Session 7 Orals – Machine Learning for NLP – Room B

11:15-12:45 (Elafiti 3)

**Combining Parameter-efficient Modules for Task-level Generalisation**
*Edoardo Maria Ponti, Alessandro Sordoni, Yoshua Bengio and Siva Reddy* 11:15-11:30 (Elafiti 3)
A modular design encourages neural models to disentangle and recombine different facets of knowledge to generalise more systematically to new tasks. In this work, we assume that each task is associated with a subset of latent skills from an (arbitrary size) inventory. In turn, each skill corresponds to a parameter-efficient (sparse / low-rank) model adapter. By jointly learning adapters and a routing function that allocates skills to each task, the full network is instantiated as the average of the parameters of active skills. We propose several inductive biases that encourage re-usage and composition of the skills, including variable-size skill allocation and a dual-speed learning rate. We evaluate our latent-skill model in two main settings: 1) multitask reinforcement learning for instruction following on 8 levels of the BabyAI platform; and 2) few-shot fine-tuning of language models on 160 NLP tasks of the CrossFit benchmark. We find that the modular design of our network enhances sample efficiency in reinforcement learning and few-shot generalisation in supervised learning, compared to a series of baselines. These include models where parameters are fully shared, task-specific, conditionally generated (HyperFormer), or sparse mixture-of-experts (TaskMoE).

**DyLoRA: Parameter-Efficient Tuning of Pre-trained Models using Dynamic Search-Free Low-Rank Adaptation**
*Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev and Ali Ghodsi* 11:30-11:45 (Elafiti 3)
With the ever-growing size of pretrained models (PMs), fine-tuning them has become more expensive and resource-hungry. As a remedy, low-rank adapters (LoRA) keep the main pretrained weights of the model frozen and just introduce some learnable truncated SVD modules (so-called LoRA blocks) to the model. While LoRA blocks are parameter-efficient, they suffer from two major problems: first, the size of these blocks is fixed and cannot be modified after training (for example, if we need to change the rank of LoRA blocks, then we need to re-train them from scratch); second, optimizing their rank requires an exhaustive search and effort. In this work, we introduce a dynamic low-rank adaptation (DyLoRA) technique to address these two problems together. Our DyLoRA method trains LoRA blocks for a range of ranks instead of a single rank by sorting the representation learned by the adapter module at different ranks during training. We evaluate our solution on different natural language understanding (GLUE benchmark) and language generation tasks (E2E, DART and WebNLG) using different pretrained models such as RoBERTa and GPT with different sizes. Our results show that we can train dynamic search-free models with DyLoRA at least 4 to 7 times (depending to the task) faster than LoRA without significantly compromising performance. Moreover, our models can perform consistently well on a much larger range of ranks compared to LoRA.

**On the inconsistency of separable losses for structured prediction**
*Caio Corro* 11:45-11:55 (Elafiti 3)
In this paper, we prove that separable negative log-likelihood losses for structured prediction are not necessarily Bayes consistent, that is minimizing these losses may not result in a model that predicts the most probable structure in the data distribution for a given input. This fact opens the question of whether these losses are well-adapted for structured prediction and, if so, why.

**Probabilistic Robustness for Data Filtering**
*Yu Yu, Abdul Khan, Shahram Khadivi and Jia Xu* 11:55-12:10 (Elafiti 3)
We introduce our probabilistic robustness rewarded data optimization (PRoDO) approach as a framework to enhance the model's generalization power by selecting training data that optimizes our probabilistic robustness metrics. We use proximal policy optimization (PPO) reinforcement learning to approximately solve the computationally intractable training subset selection problem. The PPO's reward is defined as our ($\alpha,\epsilon, \gamma$)-Robustness that measures performance consistency over multiple domains by simulating unknown test sets in real-world scenarios using a leaving-one-out strategy. We demonstrate that our PRoDO effectively filters data that lead to significantly higher prediction accuracy and robustness on unknown-domain test sets. Our experiments achieve up to +17.2\% increase of accuracy (+25.5\% relatively) in sentiment analysis, and -28.05 decrease of perplexity (-32.1\% relatively) in language modeling. In addition, our probabilistic ($\alpha,\epsilon, \gamma$)-Robustness definition serves as an evaluation metric with higher levels of agreement with human annotations than typical performance-based metrics.

**RevUp: Revise and Update Information Bottleneck for Event Representation**
*Mehdi Rezaee and Francis Ferraro* 12:10-12:25 (Elafiti 3)
The existence of external ("side") semantic knowledge has been shown to result in more expressive computational event models. To enable the use of side information that may be noisy or missing, we propose a semi-supervised information bottleneck-based discrete latent variable model. We reparameterize the model's discrete variables with auxiliary continuous latent variables and a light-weight hierarchical structure. Our model is learned to minimize the mutual information between the observed data and optional side knowledge that is not already captured by the new, auxiliary variables. We theoretically show that our approach generalizes past approaches, and perform an empirical case study of our approach on event modeling. We corroborate our theoretical results with strong empirical experiments, showing that the proposed method outperforms previous proposed approaches on multiple datasets.

**Self-imitation Learning for Action Generation in Text-based Games**
*Zijing Shi, Yunqiu Xu, Meng Fang and Ling Chen* 12:25-12:40 (Elafiti 3)
In this work, we study reinforcement learning (RL) in solving text-based games. We address the challenge of combinatorial action space, by proposing a confidence-based self-imitation model to generate action candidates for the RL agent. Firstly, we leverage the self-imitation learning to rank and exploit past valuable trajectories to adapt a pre-trained language model (LM) towards a target game. Then, we devise a confidence-based strategy to measure the LM's confidence with respect to a state, thus adaptively pruning the generated actions to yield a more compact set of action candidates. In multiple challenging games, our model demonstrates promising performance in comparison to the baselines.

## Session 7 Posters

11:15-12:45 (Exhibit Hall)

**Adding Instructions during Pretraining: Effective way of Controlling Toxicity in Language Models**
*Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi and Bryan Catanzaro* 11:15-12:45 (Exhibit Hall)
Pretrained large language models have become indispensable for solving various natural language processing (NLP) tasks. However, safely

deploying them in real world applications is challenging because they generate toxic content. To address this challenge, we propose two novel pretraining data augmentation strategies that significantly reduce model toxicity without compromising its utility. Our two strategies are: (1) MEDA: adds raw toxicity score as meta-data to the pretraining samples, and (2) INST: adds instructions to those samples indicating their toxicity. Our results indicate that our best performing strategy (INST) substantially reduces the toxicity probability up to 61% while preserving the accuracy on five benchmark NLP tasks as well as improving AUC scores on four bias detection tasks by 1.3%. We also demonstrate the generalizability of our techniques by scaling the number of training samples and the number of model parameters.

**LingMess: Linguistically Informed Multi Expert Scorers for Coreference Resolution**
*Shon Otmazgin, Arie Cattan and Yoav Goldberg* 11:15-12:45 (Exhibit Hall)
Current state-of-the-art coreference systems are based on a single pairwise scoring component, which assigns to each pair of mention spans a score reflecting their tendency to corefer to each other. We observe that different kinds of mention pairs require different information sources to assess their score. We present LingMess, a linguistically motivated categorization of mention-pairs into 6 types of coreference decisions and learn a dedicated trainable scoring function for each category. This significantly improves the accuracy of the pairwise scorer as well as of the overall coreference performance on the English Ontonotes coreference corpus and 5 additional datasets.

**Finding the Law: Enhancing Statutory Article Retrieval via Graph Neural Networks**
*Antoine Louis, Gijs Van Dijck and Gerasimos Spanakis* 11:15-12:45 (Exhibit Hall)
Statutory article retrieval (SAR), the task of retrieving statute law articles relevant to a legal question, is a promising application of legal text processing. In particular, high-quality SAR systems can improve the work efficiency of legal professionals and provide basic legal assistance to citizens in need at no cost. Unlike traditional ad-hoc information retrieval, where each document is considered a complete source of information, SAR deals with texts whose full sense depends on complementary information from the topological organization of statute law. While existing works ignore these domain-specific dependencies, we propose a novel graph-augmented dense statute retriever (G-DSR) model that incorporates the structure of legislation via a graph neural network to improve dense retrieval performance. Experimental results show that our approach outperforms strong retrieval baselines on a real-world expert-annotated SAR dataset.

**Question Generation Using Sequence-to-Sequence Model with Semantic Role Labels**
*Alireza Naeiji, Aijun An, Heidar Davoudi, Marjan Delpisheh and Muath Alzghool* 11:15-12:45 (Exhibit Hall)
Automatic generation of questions from text has gained increasing attention due to its useful applications. We propose a novel question generation method that combines the benefits of rule-based and neural sequence-to-sequence (Seq2Seq) models. The proposed method can automatically generate multiple questions from an input sentence covering different views of the sentence as in rule-based methods, while more complicated "rules" can be learned via the Seq2Seq model. The method utilizes semantic role labeling to convert training examples into their semantic representations, and then trains a Seq2Seq model over the semantic representations. Our extensive experiments on three real-world data sets show that the proposed method significantly improves the state-of-the-art neural question generation approaches.

**Friend-training: Learning from Models of Different but Related Tasks**
*Mian Zhang, Lifeng Jin, Linfeng Song, Haitao Mi, Xiabing Zhou and Dong Yu* 11:15-12:45 (Exhibit Hall)
Current self-training methods such as standard self-training, co-training, tri-training, and others often focus on improving model performance on a single task, utilizing differences in input features, model architectures, and training processes. However, many tasks in natural language processing are about different but related aspects of language, and models trained for one task can be great teachers for other related tasks. In this work, we propose friend-training, a cross-task self-training framework, where models trained to do different tasks are used in an iterative training, pseudo-labeling, and retraining process to help each other for better selection of pseudo-labels. With two dialogue understanding tasks, conversational semantic role labeling and dialogue rewriting, chosen for a case study, we show that the models trained with the friend-training framework achieve the best performance compared to strong baselines.

**Comparing Intrinsic Gender Bias Evaluation Measures without using Human Annotated Examples**
*Masahiro Kaneko, Danushka Bollegala and Naoaki Okazaki* 11:15-12:45 (Exhibit Hall)
Numerous types of social biases have been identified in pre-trained language models (PLMs), and various intrinsic bias evaluation measures have been proposed for quantifying those social biases. Prior works have relied on human annotated examples to compare existing intrinsic bias evaluation measures. However, this approach is not easily adaptable to different languages nor amenable to large scale evaluations due to the costs and difficulties when recruiting human annotators. To overcome this limitation, we propose a method to compare intrinsic gender bias evaluation measures without relying on human-annotated examples. Specifically, we create multiple bias-controlled versions of PLMs using varying amounts of male vs. female gendered sentences, mined automatically from an unannotated corpus using gender-related word lists. Next, each bias-controlled PLM is evaluated using an intrinsic bias evaluation measure, and the rank correlation between the computed bias scores and the gender proportions used to fine-tune the PLMs is computed. Experiments on multiple corpora and PLMs repeatedly show that the correlations reported by our proposed method that does not require human annotated examples are comparable to those computed using human annotated examples in prior work.

**Should You Mask 15% in Masked Language Modeling?**
*Alexander Wettig, Tianyu Gao, Zexuan Zhong and Danqi Chen* 11:15-12:45 (Exhibit Hall)
Masked language models (MLMs) conventionally mask 15% of tokens due to the belief that more masking would leave insufficient context to learn good representations; this masking rate has been widely used, regardless of model sizes or masking strategies. In this work, we revisit this important choice of MLM pre-training. We first establish that 15% is not universally optimal, and larger models should adopt a higher masking rate. Specifically, we find that masking 40% outperforms 15% for BERT-large size models on GLUE and SQuAD. Interestingly, an extremely high masking rate of 80% can still preserve 95% fine-tuning performance and most of the accuracy in linguistic probing, challenging the conventional wisdom about the role of the masking rate. We then examine the interplay between masking rates and masking strategies and find that uniform masking requires a higher masking rate compared to sophisticated masking strategies such as span or PMI masking. Finally, we argue that increasing the masking rate has two distinct effects: it leads to more corruption, which makes the prediction task more difficult; it also enables more predictions, which benefits optimization. Using this framework, we revisit BERT's 80-10-10 corruption strategy. Together, our results contribute to a better understanding of MLM pre-training.

**DeepMaven: Deep Question Answering on Long-Distance Movie/TV Show Videos with Multimedia Knowledge Extraction and Synthesis**
*Yi Fung, Han Wang, Tong Wang, Ali Kebarighotbi, Mohit Bansal, Heng Ji and Prem Natarajan* 11:15-12:45 (Exhibit Hall)
Long video content understanding poses a challenging set of research questions as it involves long-distance, cross-media reasoning and knowledge awareness. In this paper, we present a new benchmark for this problem domain, targeting the task of deep movie/TV question answering (QA) beyond previous work's focus on simple plot summary and short video moment settings. We define several baselines based on direct retrieval of relevant context for long-distance movie QA. Observing that real-world QAs may require higher-order multi-hop inferences, we further propose a novel framework, called the DeepMaven, which extracts events, entities, and relations from the rich multimedia content in long videos to pre-construct movie knowledge graphs (movieKGs), and at the time of QA inference, complements general

semantics with structured knowledge for more effective information retrieval and knowledge reasoning. We also introduce our recently collected DeepMovieQA dataset, including 1,000 long-form QA pairs from 41 hours of videos, to serve as a new and useful resource for future work. Empirical results show the DeepMaven performs competitively for both the new DeepMovieQA and the pre-existing MovieQA dataset.

**DiffQG: Generating Questions to Summarize Factual Changes**
*Jeremy Cole, Palak Jain, Julian Eisenschlos, Michael Zhang, Eunsol Choi and Bhuwan Dhingra* 11:15-12:45 (Exhibit Hall)
Identifying the difference between two versions of the same article is useful to update knowledge bases and to understand how articles evolve. Paired texts occur naturally in diverse situations: reporters write similar news stories and maintainers of authoritative websites must keep their information up to date. We propose representing factual changes between paired documents as question-answer pairs, where the answer to the same question differs between two versions. We find that question-answer pairs can flexibly and concisely capture the updated contents. Provided with paired documents, annotators identify questions that are answered by one passage but answered differently or cannot be answered by the other. We release DiffQG which consists of 759 QA pairs and 1153 examples of paired passages with no factual change. These questions are intended to be both unambiguous and information-seeking and involve complex edits, pushing beyond the capabilities of current question generation and factual change detection systems. Our dataset summarizes the changes between two versions of the document as questions and answers, studying automatic update summarization in a novel way.

**Paraphrase Acquisition from Image Captions**
*Marcel Gohsen, Matthias Hagen, Martin Potthast and Benno Stein* 11:15-12:45 (Exhibit Hall)
We propose to use image captions from the Web as a previously underutilized resource for paraphrases (i.e., texts with the same "message") and to create and analyze a corresponding dataset. When an image is reused on the Web, an original caption is often assigned. We hypothesize that different captions for the same image naturally form a set of mutual paraphrases. To demonstrate the suitability of this idea, we analyze captions in the English Wikipedia, where editors frequently relabel the same image for different articles. The paper introduces the underlying mining technology, the resulting Wikipedia-IPC dataset, and compares known paraphrase corpora with respect to their syntactic and semantic paraphrase similarity to our new resource. In this context, we introduce characteristic maps along the two similarity dimensions to identify the style of paraphrases coming from different sources. An annotation study demonstrates the high reliability of the algorithmically determined characteristic maps.

**What Did You Learn To Hate? A Topic-Oriented Analysis of Generalization in Hate Speech Detection**
*Tom Bourgeade, Patricia Chiril, Farah Benamara and Véronique Moriceau* 11:15-12:45 (Exhibit Hall)
Hate speech has unfortunately become a significant phenomenon on social media platforms, and it can cover various topics (misogyny, sexism, racism, xenophobia, etc.) and targets (e.g., black people, women). Various hate speech detection datasets have been proposed, some annotated for specific topics, and others for hateful speech in general. In either case, they often employ different annotation guidelines, which can lead to inconsistencies, even in datasets focusing on the same topics. This can cause issues in models trying to generalize across more data and more topics in order to improve detection accuracy. In this paper, we propose, for the first time, a topic-oriented approach to study generalization across popular hate speech datasets. We first perform a comparative analysis of the performances of Transformer-based models in capturing topic-generic and topic-specific knowledge when trained on different datasets. We then propose a novel, simple yet effective approach to study more precisely which topics are best captured in implicit manifestations of hate, showing that selecting combinations of datasets with better out-of-domain topical coverage improves the reliability of automatic hate speech detection.

**CHARD: Clinical Health-Aware Reasoning Across Dimensions for Text Generation Models**
*Steven Y. Feng, Vivek Khetan, Bogdan Sacaleanu, Anatole Gershman and Eduard Hovy* 11:15-12:45 (Exhibit Hall)
We motivate and introduce CHARD: Clinical Health-Aware Reasoning across Dimensions, to investigate the capability of text generation models to act as implicit clinical knowledge bases and generate free-flow textual explanations about various health-related conditions across several dimensions. We collect and present an associated dataset, CHARDat, consisting of explanations about 52 health conditions across three clinical dimensions. We conduct extensive experiments using BART and T5 along with data augmentation, and perform automatic, human, and qualitative analyses. We show that while our models can perform decently, CHARD is very challenging with strong potential for further exploration.

**End-to-end Case-Based Reasoning for Commonsense Knowledge Base Completion**
*Zonglin Yang, Xinya Du, Erik Cambria and Claire Cardie* 11:15-12:45 (Exhibit Hall)
Pretrained language models have been shown to store knowledge in their parameters and have achieved reasonable performance in commonsense knowledge base completion (CKBC) tasks. However, CKBC is knowledge-intensive and it is reported that pretrained language models' performance in knowledge-intensive tasks are limited because of their incapability of accessing and manipulating knowledge. As a result, we hypothesize that providing retrieved passages that contain relevant knowledge as additional input to the CKBC task will improve performance. In particular, we draw insights from Case-Based Reasoning (CBR) – which aims to solve a new problem by reasoning with retrieved relevant cases, and investigate the direct application of it to CKBC. On two benchmark datasets, we demonstrate through automatic and human evaluations that our End-to-end Case-Based Reasoning Framework (ECBRF) generates more valid, informative, and novel knowledge than the state-of-the-art COMET model for CKBC in both the fully supervised and few-shot settings. We provide insights on why previous retrieval-based methods only achieve merely the same performance with COMET. From the perspective of CBR, our framework addresses a fundamental question on whether CBR methodology can be utilized to improve deep learning models.

**Prompt Tuning with Contradictory Intentions for Sarcasm Recognition**
*Yiyi Liu, Ruqing Zhang, Yixing Fan, Jiafeng Guo and Xueqi Cheng* 11:15-12:45 (Exhibit Hall)
Recently, prompt tuning has achieved promising results in a variety of natural language processing (NLP) tasks. The typical approach is to insert text pieces (i.e. templates) into the input and transform downstream tasks into the same form as pre-training. In essence, a high-quality template is the foundation of prompt tuning to support the performance of the converted cloze-style task. However, for sarcasm recognition, it is time-consuming and requires increasingly sophisticated domain knowledge to determine the appropriate templates and label words due to its highly figurative nature. In this work, we propose SarcPrompt, to incorporate the prior knowledge about contradictory intentions into prompt tuning for sarcasm recognition. SarcPrompt is inspired by that the speaker usually says the opposite of what they actually mean in the sarcastic text. Based on this idea, we explicitly mimic the actual intention by prompt construction and indicate whether the actual intention is contradictory to the literal content by verbalizer engineering. Experiments on three public datasets with standard and low-resource settings demonstrate the effectiveness of our SarcPrompt for sarcasm recognition.

**Retrieval-augmented Image Captioning**
*Rita Ramos, Desmond Elliott and Bruno Martins* 11:15-12:45 (Exhibit Hall)
Inspired by retrieval-augmented language generation and pretrained Vision and Language (VL) encoders, we present a new approach to image captioning that generates sentences given the input image and a set of captions retrieved from a datastore, as opposed to the image alone. The encoder in our model jointly processes the image and retrieved captions using a pretrained VL BERT, while the decoder attends to the multimodal encoder representations, benefiting from the extra textual evidence from the retrieved captions. Experimental results on the COCO

dataset show that image captioning can be effectively formulated from this new perspective. Our model, named EXTRA, benefits from using captions retrieved from the training dataset, and it can also benefit from using an external dataset without the need for retraining. Ablation studies show that retrieving a sufficient number of captions (e.g., k=5) can improve captioning quality. Our work contributes towards using pretrained VL encoders for generative tasks, instead of standard classification tasks.

**Self-training Reduces Flicker in Retranslation-based Simultaneous Translation**
*Sukanta Sen, Rico Sennrich, Biao Zhang and Barry Haddow* 11:15-12:45 (Exhibit Hall)
In simultaneous translation, the retranslation approach has the advantage of requiring no modifications to the inference engine. However, in order to reduce the undesirable flicker in the output, previous work has resorted to increasing the latency through masking, and introducing specialised inference, thus losing the simplicity of the approach. In this work, we show that self-training improves the flicker-latency tradeoff, while maintaining similar translation quality to the original. Our analysis indicates that self-training reduces flicker by controlling monotonicity. Furthermore, self-training can be combined with biased beam search to further improve the flicker-latency tradeoff.

**Meeting the Needs of Low-Resource Languages: The Value of Automatic Alignments via Pretrained Models**
*Abteen Ebrahimi, Arya D. Mccarthy, Arturo Oncevay, John Ortega, Luis Chiruzzo, Gustavo Giménez-lugo, Rolando Coto-solano and Katharina Kann* 11:15-12:45 (Exhibit Hall)
Large multilingual models have inspired a new class of word alignment methods, which work well for the model's pretraining languages. However, the languages most in need of automatic alignment are low-resource and, thus, not typically included in the pretraining data. In this work, we ask: How do modern aligners perform on unseen languages, and are they better than traditional methods? We contribute gold-standard alignments for Bribri–Spanish, Guarani–Spanish, Quechua–Spanish, and Shipibo-Konibo–Spanish. With these, we evaluate state-of-the-art aligners with and without model adaptation to the target language. Finally, we also evaluate the resulting alignments extrinsically through two downstream tasks: named entity recognition and part-of-speech tagging. We find that although transformer-based methods generally outperform traditional models, the two classes of approach remain competitive with each other.

**Assistive Recipe Editing through Critiquing**
*Diego Antognini, Shuyang Li, Boi Faltings and Julian Mcauley* 11:15-12:45 (Exhibit Hall)
There has recently been growing interest in the automatic generation of cooking recipes that satisfy some form of dietary restrictions, thanks in part to the availability of online recipe data. Prior studies have used pre-trained language models, or relied on small paired recipe data (e.g., a recipe paired with a similar one that satisfies a dietary constraint). However, pre-trained language models generate inconsistent or incoherent recipes, and paired datasets are not available at scale. We address these deficiencies with RecipeCrit, a hierarchical denoising auto-encoder that edits recipes given ingredient-level critiques. The model is trained for recipe completion to learn semantic relationships within recipes. Our work's main innovation is our unsupervised critiquing module that allows users to edit recipes by interacting with the predicted ingredients; the system iteratively rewrites recipes to satisfy users' feedback. Experiments on the Recipe1M recipe dataset show that our model can more effectively edit recipes compared to strong language-modeling baselines, creating recipes that satisfy user constraints and are more correct, serendipitous, coherent, and relevant as measured by human judges.

**Realistic Conversational Question Answering with Answer Selection based on Calibrated Confidence and Uncertainty Measurement**
*Soyeong Jeong, Jinheon Baek, Sung Ju Hwang and Jong Park* 11:15-12:45 (Exhibit Hall)
Conversational Question Answering (ConvQA) models aim at answering a question with its relevant paragraph and previous question-answer pairs that occurred during conversation multiple times. To apply such models to a real-world scenario, some existing work uses predicted answers, instead of unavailable ground-truth answers, as the conversation history for inference. However, since these models usually predict wrong answers, using all the predictions without filtering significantly hampers the model performance. To address this problem, we propose to filter out inaccurate answers in the conversation history based on their estimated confidences and uncertainties from the ConvQA model, without making any architectural changes. Moreover, to make the confidence and uncertainty values more reliable, we propose to further calibrate them, thereby smoothing the model predictions. We validate our models, Answer Selection-based realistic Conversation Question Answering, on two standard ConvQA datasets, and the results show that our models significantly outperform relevant baselines. Code is available at: https://github.com/starsuzi/AS-ConvQA.

**PromptDA: Label-guided Data Augmentation for Prompt-based Few Shot Learners**
*Canyu Chen and Kai Shu* 11:15-12:45 (Exhibit Hall)
Recent advances in large pre-trained language models (PLMs) lead to impressive gains on natural language understanding (NLU) tasks with task-specific fine-tuning. However, directly fine-tuning PLMs heavily relies on sufficient labeled training instances, which are usually hard to obtain. Prompt-based tuning on PLMs has shown to be powerful for various downstream few-shot tasks. Existing works studying prompt-based tuning for few-shot NLU tasks mainly focus on deriving proper label words with a verbalizer or generating prompt templates to elicit semantics from PLMs. In addition, conventional data augmentation strategies such as synonym substitution are also widely adopted in low-resource scenarios. However, the improvements they bring to prompt-based few-shot learning have been demonstrated to be marginal. Thus, an important research question arises as follows: how to design effective data augmentation methods for prompt-based few-shot tuning? To this end, considering the label semantics are essential in prompt-based tuning, we propose a novel label-guided data augmentation framework PromptDA, which exploits the enriched label semantic information for data augmentation. Extensive experiment results on few-shot text classification tasks show that our proposed framework achieves superior performances by effectively leveraging label semantics and data augmentation for natural language understanding.

**What Makes Sentences Semantically Related? A Textual Relatedness Dataset and Empirical Study**
*Mohamed Abdalla, Krishnapriya Vishnubhotla and Saif Mohammad* 11:15-12:45 (Exhibit Hall)
The degree of semantic relatedness of two units of language has long been considered fundamental to understanding meaning. Additionally, automatically determining relatedness has many applications such as question answering and summarization. However, prior NLP work has largely focused on semantic similarity, a subset of relatedness, because of a lack of relatedness datasets. In this paper, we introduce a dataset for Semantic Textual Relatedness, STR-2022, that has 5,500 English sentence pairs manually annotated using a comparative annotation framework, resulting in fine-grained scores. We show that human intuition regarding relatedness of sentence pairs is highly reliable, with a repeat annotation correlation of 0.84. We use the dataset to explore questions on what makes sentences semantically related. We also show the utility of STR-2022 for evaluating automatic methods of sentence representation and for various downstream NLP tasks.

Our dataset, data statement, and annotation questionnaire can be found at: https://doi.org/10.5281/zenodo.7599667.

**The Functional Relevance of Probed Information: A Case Study**
*Michael Hanna, Roberto Zamparelli and David Mareček* 11:15-12:45 (Exhibit Hall)
Recent studies have shown that transformer models like BERT rely on number information encoded in their representations of sentences' subjects and head verbs when performing subject-verb agreement. However, probing experiments suggest that subject number is also encoded in the representations of all words in such sentences. In this paper, we use causal interventions to show that BERT only uses the subject plurality information encoded in its representations of the subject and words that agree with it in number. We also demonstrate that current

probing metrics are unable to determine which words' representations contain functionally relevant information. This both provides a revised view of subject-verb agreement in language models, and suggests potential pitfalls for current probe usage and evaluation.

**Probing Power by Prompting: Harnessing Pre-trained Language Models for Power Connotation Framing**
*Shima Khanehzar, Trevor Cohn, Gosia Mikolajczak and Lea Frermann* 11:15-12:45 (Exhibit Hall)
When describing actions, subtle changes in word choice can evoke very different associations with the involved entities. For instance, a company '{\it employing} workers' evokes a more positive connotation than the one '{\it exploiting}' them. This concept is called {\it connotation}. This paper investigates whether pre-trained language models (PLMs) encode such subtle connotative information about {\it power differentials} between involved entities. We design a probing framework for power connotation, building on~\citet{sap-etal-2017-connotation}'s operationalization of {\it connotation frames}. We show that zero-shot prompting of PLMs leads to above chance prediction of power connotation, however fine-tuning PLMs using our framework drastically improves their accuracy. Using our fine-tuned models, we present a case study of {\it power dynamics} in US news reporting on immigration, showing the potential of our framework as a tool for understanding subtle bias in the media.

**Find Parent then Label Children: A Two-stage Taxonomy Completion Method with Pre-trained Language Model**
*Fei Xia, Yixuan Weng, Shizhu He, Kang Liu and Jun Zhao* 11:15-12:45 (Exhibit Hall)
Taxonomies, which organize domain concepts into hierarchical structures, are crucial for building knowledge systems and downstream applications. As domain knowledge evolves, taxonomies need to be continuously updated to include new concepts. Previous approaches have mainly focused on adding concepts to the leaf nodes of the existing hierarchical tree, which does not fully utilize the taxonomy's knowledge and is unable to update the original taxonomy structure (usually involving non-leaf nodes). In this paper, we propose a two-stage method called ATTEMPT for taxonomy completion. Our method inserts new concepts into the correct position by finding a parent node and labeling child nodes. Specifically, by combining local nodes with prompts to generate natural sentences, we take advantage of pre-trained language models for hypernym/hyponymy recognition. Experimental results on two public datasets (including six domains) show that ATTEMPT performs best on both taxonomy completion and extension tasks, surpassing existing methods.

**Cross-Lingual Dialogue Dataset Creation via Outline-Based Generation**
*Majewska Olga, Edoardo M. Ponti, Ivan Vulić and Anna Korhonen* 11:15-12:45 (Exhibit Hall)
Multilingual task-oriented dialogue (ToD) facilitates access to services and information for many (communities of) speakers. Nevertheless, its potential is not fully realized, as current multilingual ToD datasets—both for modular and end-to-end modeling—suffer from severe limitations. 1) When created from scratch, they are usually small in scale and fail to cover many possible dialogue flows. 2) Translation-based ToD datasets might lack naturalness and cultural specificity in the target language. In this work, to tackle these limitations we propose a novel outline-based annotation process for multilingual ToD datasets, where domain-specific abstract schemata of dialogue are mapped into natural language outlines. These in turn guide the target language annotators in writing dialogues by providing instructions about each turn's intents and slots. Through this process we annotate a new large-scale dataset for evaluation of multilingual and cross-lingual ToD systems. Our Cross-lingual Outline-based Dialogue dataset (cod) enables natural language understanding, dialogue state tracking, and end-to-end dialogue evaluation in 4 diverse languages: Arabic, Indonesian, Russian, and Kiswahili. Qualitative and quantitative analyses of cod versus an equivalent translation-based dataset demonstrate improvements in data quality, unlocked by the outline-based approach. Finally, we benchmark a series of state-of-the-art systems for cross-lingual ToD, setting reference scores for future work and demonstrating that cod prevents over-inflated performance, typically met with prior translation-based ToD datasets.

**Event Temporal Relation Extraction with Bayesian Translational Model**
*Xingwei Tan, Gabriele Pergola and Yulan He* 11:15-12:45 (Exhibit Hall)
Existing models to extract temporal relations between events lack a principled method to incorporate external knowledge. In this study, we introduce Bayesian-Trans, a Bayesian learning-based method that models the temporal relation representations as latent variables and infers their values via Bayesian inference and translational functions. Compared to conventional neural approaches, instead of performing point estimation to find the best set parameters, the proposed model infers the parameters' posterior distribution directly, enhancing the model's capability to encode and express uncertainty about the predictions. Experimental results on the three widely used datasets show that Bayesian-Trans outperforms existing approaches for event temporal relation extraction. We additionally present detailed analyses on uncertainty quantification, comparison of priors, and ablation studies, illustrating the benefits of the proposed approach.

**Poor Man's Quality Estimation: Predicting Reference-Based MT Metrics Without the Reference**
*Vilém Zouhar, Shehzaad Dhuliawala, Wangchunshu Zhou, Nico Daheim, Tom Kocmi, Yuchen Eleanor Jiang and Mrinmaya Sachan* 11:15-12:45 (Exhibit Hall)
Machine translation quality estimation (QE) predicts human judgements of a translation hypothesis without seeing the reference. State-of-the-art QE systems based on pretrained language models have been achieving remarkable correlations with human judgements yet they are computationally heavy and require human annotations, which are slow and expensive to create. To address these limitations, we define the problem of metric estimation (ME) where one predicts the automated metric scores also without the reference. We show that even without access to the reference, our model can estimate automated metrics at the sentence-level. Because automated metrics correlate with human judgements, we can leverage the ME task for pre-training a QE model. For the QE task, we find that pre-training on TER is better than training for scratch.

**Unsupervised Anomaly Detection in Multi-Topic Short-Text Corpora**
*Mira Ait-saada and Mohamed Nadif* 11:15-12:45 (Exhibit Hall)
Unsupervised anomaly detection seeks to identify deviant data samples in a dataset without using labels and constitutes a challenging task, particularly when the majority class is heterogeneous. This paper addresses this topic for textual data and aims to determine whether a text sample is an outlier within a potentially multi-topic corpus. To this end, it is crucial to grasp the semantic aspects of words, particularly when dealing with short texts, since it is difficult to syntactically discriminate data samples based only on a few words. Thereby we make use of word embeddings to represent each sample by a dense vector, efficiently capturing the underlying semantics. Then, we rely on the Mixture Model approach to detect which samples deviate the most from the underlying distributions of the corpus. Experiments carried out on real datasets show the effectiveness of the proposed approach in comparison to state-of-the-art techniques both in terms of performance and time efficiency, especially when more than one topic is present in the corpus.

**Low-Resource Compositional Semantic Parsing with Concept Pretraining**
*Subendhu Rongali, Mukund Sridhar, Haidar Khan, Konstantine Arkoudas, Wael Hamza and Andrew Mccallum* 11:15-12:45 (Exhibit Hall)
Semantic parsing plays a key role in digital voice assistants such as Alexa, Siri, and Google Assistant by mapping natural language to structured meaning representations. When we want to improve the capabilities of a voice assistant by adding a new domain, the underlying semantic parsing model needs to be retrained using thousands of annotated examples from the new domain, which is time-consuming and expensive. In this work, we present an architecture to perform such domain adaptation automatically, with only a small amount of metadata about the new domain and without any new training data (zero-shot) or with very few examples (few-shot). We use a base seq2seq

(sequence-to-sequence) architecture and augment it with a concept encoder that encodes intent and slot tags from the new domain. We also introduce a novel decoder-focused approach to pretrain seq2seq models to be concept aware using Wikidata and use it to help our model learn important concepts and perform well in low-resource settings. We report few-shot and zero-shot results for compositional semantic parsing on the TOPv2 dataset and show that our model outperforms prior approaches in few-shot settings for the TOPv2 and SNIPS datasets.

**FrameBERT: Conceptual Metaphor Detection with Frame Embedding Learning**
*Yucheng Li, Shun Wang, Chenghua Lin, Frank Guerin and Loic Barrault* 11:15-12:45 (Exhibit Hall)
In this paper, we propose FrameBERT, a BERT-based model that can explicitly learn and incorporate FrameNet Embeddings for concept-level metaphor detection. FrameBERT not only achieves better or comparable performance to the state-of-the-art, but also is more explainable and interpretable compared to existing models, attributing to its ability of accounting for external knowledge of FrameNet.

**CTC Alignments Improve Autoregressive Translation**
*Brian Yan, Siddharth Dalmia, Yosuke Higuchi, Graham Neubig, Florian Metze, Alan W Black and Shinji Watanabe* 11:15-12:45 (Exhibit Hall)
Connectionist Temporal Classification (CTC) is a widely used approach for automatic speech recognition (ASR) that performs conditionally independent monotonic alignment. However for translation, CTC exhibits clear limitations due to the contextual and non-monotonic nature of the task and thus lags behind attentional decoder approaches in terms of translation quality. In this work, we argue that CTC does in fact make sense for translation if applied in a joint CTC/attention framework wherein CTC's core properties can counteract several key weaknesses of pure-attention models during training and decoding. To validate this conjecture, we modify the Hybrid CTC/Attention model originally proposed for ASR to support text-to-text translation (MT) and speech-to-text translation (ST). Our proposed joint CTC/attention models outperform pure-attention baselines across six benchmark translation tasks.

**Cluster-Guided Label Generation in Extreme Multi-Label Classification**
*Taehee Jung, Joo-kyung Kim, Sungjin Lee and Dongyeop Kang* 11:15-12:45 (Exhibit Hall)
For extreme multi-label classification (XMC), existing classification-based models poorly per- form for tail labels and often ignore the semantic relations among labels, like treating "Wikipedia" and "Wiki" as independent and separate labels. In this paper, we cast XMC as a generation task (XLGen), where we benefit from pre-trained text-to-text models. However, generating labels from the extremely large label space is challenging without any constraints or guidance. We, therefore, propose to guide label generation using label cluster information to hierarchically generate lower-level labels. We also find that frequency-based label ordering and using decoding ensemble methods are critical factors for the improvements in XLGen. XLGen with cluster guidance significantly outperforms the classification and generation baselines on tail labels, and also generally improves the overall performance in four popular XMC benchmarks. In human evaluation, we also find XLGen generates unseen but plausible labels. Our code is now available at https:// github.com/alexa/xlgen-eacl-2023.

**CylE: Cylinder Embeddings for Multi-hop Reasoning over Knowledge Graphs**
*Chau Nguyen, Tim French, Wei Liu and Michael Stewart* 11:15-12:45 (Exhibit Hall)
Recent geometric-based approaches have been shown to efficiently model complex logical queries (including the intersection operation) over Knowledge Graphs based on the natural representation of Venn diagram. Existing geometric-based models (using points, boxes embeddings), however, cannot handle the logical negation operation. Further, those using cones embeddings are limited to representing queries by two-dimensional shapes, which reduced their effectiveness in capturing entities query relations for correct answers. To overcome this challenge, we propose unbounded cylinder embeddings (namely CylE), which is a novel geometric-based model based on three-dimensional shapes. Our approach can handle a complete set of basic first-order logic operations (conjunctions, disjunctions and negations). CylE considers queries as Cartesian products of unbounded sector-cylinders and consider a set of nearest boxes corresponds to the set of answer entities. Precisely, the conjunctions can be represented via the intersections of unbounded sector-cylinders. Transforming queries to Disjunctive Normal Form can handle queries with disjunctions. The negations can be represented by considering the closure of complement for an arbitrary unbounded sector-cylinder. Empirical results show that the performance of multi-hop reasoning task using CylE significantly increases over state-of-the-art geometric-based query embedding models for queries without negation. For queries with negation operations, though the performance is on a par with the best performing geometric-based model, CylE significantly outperforms a recent distribution-based model.

**CLICK: Contrastive Learning for Injecting Contextual Knowledge to Conversational Recommender System**
*Hyeongjun Yang, Heesoo Won, Youbin Ahn and Kyong-ho Lee* 11:15-12:45 (Exhibit Hall)
Conversational recommender systems (CRSs) capture a user preference through a conversation. However, the existing CRSs lack capturing comprehensive user preferences. This is because the items mentioned in a conversation are mainly regarded as a user preference. Thus, they have limitations in identifying a user preference from a dialogue context expressed without preferred items. Inspired by the characteristic of an online recommendation community where participants identify a context of a recommendation request and then comment with appropriate items, we exploit the Reddit data. Specifically, we propose a Contrastive Learning approach for Injecting Contextual Knowledge (CLICK) from the Reddit data to the CRS task, which facilitates the capture of a context-level user preference from a dialogue context, regardless of the existence of preferred item-entities. Moreover, we devise a relevance-enhanced contrastive learning loss to consider the fine-grained reflection of multiple recommendable items. We further develop a response generation module to generate a persuasive rationale for a recommendation. Extensive experiments on the benchmark CRS dataset show the effectiveness of CLICK, achieving significant improvements over state-of-the-art methods.

**Guide the Learner: Controlling Product of Experts Debiasing Method Based on Token Attribution Similarities**
*Ali Modarressi, Hossein Amirkhani and Mohammad Taher Pilehvar* 11:15-12:45 (Exhibit Hall)
Several proposals have been put forward in recent years for improving out-of-distribution (OOD) performance through mitigating dataset biases. A popular workaround is to train a robust model by re-weighting training examples based on a secondary biased model. Here, the underlying assumption is that the biased model resorts to shortcut features. Hence, those training examples that are correctly predicted by the biased model are flagged as being biased and are down-weighted during the training of the main model. However, assessing the importance of an instance merely based on the predictions of the biased model may be too naive. It is possible that the prediction of the main model can be derived from another decision-making process that is distinct from the behavior of the biased model. To circumvent this, we introduce a fine-tuning strategy that incorporates the similarity between the main and biased model attribution scores in a Product of Experts (PoE) loss function to further improve OOD performance. With experiments conducted on natural language inference and fact verification benchmarks, we show that our method improves OOD results while maintaining in-distribution (ID) performance.

**Task and Sentiment Adaptation for Appraisal Tagging**
*Lin Tian, Xiuzhen Zhang, Myung Hee Kim and Jennifer Biggs* 11:15-12:45 (Exhibit Hall)
The Appraisal framework in linguistics defines the framework for fine-grained evaluations and opinions and has contributed to sentiment analysis and opinion mining. As developing appraisal-annotated resources requires tagging of several dimensions with distinct semantic taxonomies, it has been primarily conducted manually by human experts through expensive and time-consuming processes. In this paper, we study how to automatically identify and annotate text segments for appraisal. We formulate the problem as a sequence tagging problem and

propose novel task and sentiment adapters based on language models for appraisal tagging. Our model, named Adaptive Appraisal (A$ˆ2$), achieves superior performance than baseline adapter-based models and other neural classification models, especially for cross-domain and cross-language settings. Source code for A$ˆ2$ is available at: https://github.com/ltian678/AA-code.git

**An In-depth Analysis of Implicit and Subtle Hate Speech Messages**
*Nicolas Ocampo, Ekaterina Sviridova, Elena Cabrio and Serena Villata* 11:15-12:45 (Exhibit Hall)
The research carried out so far in detecting abusive content in social media has primarily focused on overt forms of hate speech. While explicit hate speech (HS) is more easily identifiable by recognizing hateful words, messages containing linguistically subtle and implicit forms of HS (as circumlocution, metaphors and sarcasm) constitute a real challenge for automatic systems. While the sneaky and tricky nature of subtle messages might be perceived as less hurtful with respect to the same content expressed clearly, such abuse is at least as harmful as overt abuse. In this paper, we first provide an in-depth and systematic analysis of 7 standard benchmarks for HS detection, relying on a fine-grained and linguistically-grounded definition of implicit and subtle messages. Then, we experiment with state-of-the-art neural network architectures on two supervised tasks, namely implicit HS and subtle HS message classification. We show that while such models perform satisfactory on explicit messages, they fail to detect implicit and subtle content, highlighting the fact that HS detection is not a solved problem and deserves further investigation.

**Step by Step Loss Goes Very Far: Multi-Step Quantization for Adversarial Text Attacks**
*Piotr Gaiński and Klaudia Bałazy* 11:15-12:45 (Exhibit Hall)
We propose a novel gradient-based attack against transformer-based language models that searches for an adversarial example in a continuous space of tokens probabilities. Our algorithm mitigates the gap between adversarial loss for continuous and discrete text representations by performing multi-step quantization in a quantization-compensation loop. Experiments show that our method significantly outperforms other approaches on various natural language processing (NLP) tasks.

**UDAPTER - Efficient Domain Adaptation Using Adapters**
*Bhavitvya Malik, Abhinav Ramesh Kashyap, Min-yen Kan and Soujanya Poria* 11:15-12:45 (Exhibit Hall)
We propose two methods to make unsupervised domain adaptation (UDA) more parameter efficient using adapters – small bottleneck layers interspersed with every layer of the large-scale pre-trained language model (PLM). The first method deconstructs UDA into a two-step process: first by adding a domain adapter to learn domain-invariant information and then by adding a task adapter that uses domain-invariant information to learn task representations in the source domain. The second method jointly learns a supervised classifier while reducing the divergence measure. Compared to strong baselines, our simple methods perform well in natural language inference (MNLI) and the cross-domain sentiment classification task. We even outperform unsupervised domain adaptation methods such as DANN and DSN in sentiment classification, and we are within 0.85% F1 for natural language inference task, by fine-tuning only a fraction of the full model parameters. We release our code at this URL.

**An Empirical Study of Clinical Note Generation from Doctor-Patient Encounters**
*Asma Ben Abacha, Wen-wai Yim, Yadan Fan and Thomas Lin* 11:15-12:45 (Exhibit Hall)
Medical doctors spend on average 52 to 102 minutes per day writing clinical notes from their patient encounters (Hripcsak et al., 2011). Reducing this workload calls for relevant and efficient summarization methods. In this paper, we introduce new resources and empirical investigations for the automatic summarization of doctor-patient conversations in a clinical setting. In particular, we introduce the MTS-Dialog dataset; a new collection of 1,700 doctor-patient dialogues and corresponding clinical notes. We use this new dataset to investigate the feasibility of this task and the relevance of existing language models, data augmentation, and guided summarization techniques. We compare standard evaluation metrics based on n-gram matching, contextual embeddings, and Fact Extraction to assess the accuracy and the factual consistency of the generated summaries. To ground these results, we perform an expert-based evaluation using relevant natural language generation criteria and task-specific criteria such as critical omissions, and study the correlation between the automatic metrics and expert judgments. To the best of our knowledge, this study is the first attempt to introduce an open dataset of doctor-patient conversations and clinical notes, with detailed automated and manual evaluations of clinical note generation.

**Can Synthetic Text Help Clinical Named Entity Recognition? A Study of Electronic Health Records in French**
*Nicolas Hiebel, Olivier Ferret, Karen Fort and Aurélie Névéol* 11:15-12:45 (Exhibit Hall)
In sensitive domains, the sharing of corpora is restricted due to confidentiality, copyrights or trade secrets. Automatic text generation can help alleviate these issues by producing synthetic texts that mimic the linguistic properties of real documents while preserving confidentiality. In this study, we assess the usability of synthetic corpus as a substitute training corpus for clinical information extraction. Our goal is to automatically produce a clinical case corpus annotated with clinical entities and to evaluate it for a named entity recognition (NER) task. We use two auto-regressive neural models partially or fully trained on generic French texts and fine-tuned on clinical cases to produce a corpus of synthetic clinical cases. We study variants of the generation process: (i) fine-tuning on annotated vs. plain text (in that case, annotations are obtained a posteriori) and (ii) selection of generated texts based on models parameters and filtering criteria. We then train NER models with the resulting synthetic text and evaluate them on a gold standard clinical corpus. Our experiments suggest that synthetic text is useful for clinical NER.

**Parameter-Efficient Korean Character-Level Language Modeling**
*Marco Cognetta, Sangwhan Moon, Lawrence Wolf-sonkin and Naoaki Okazaki* 11:15-12:45 (Exhibit Hall)
Character-level language modeling has been shown empirically to perform well on highly agglutinative or morphologically rich languages while using only a small fraction of the parameters required by (sub)word models. Korean fits nicely into this framework, except that, like other CJK languages, it has a very large character vocabulary of 11,172 unique syllables. However, unlike Japanese Kanji and Chinese Hanzi, each Korean syllable can be uniquely factored into a small set of subcharacters, called jamo. We explore a "three-hot" scheme, where we exploit the decomposability of Korean characters to model at the syllable level but using only jamo-level representations. We find that our three-hot embedding and decoding scheme alleviates the two major issues with prior syllable- and jamo-level models. Namely, it requires fewer than 1% of the embedding parameters of a syllable model, and it does not require tripling the sequence length, as with jamo models. In addition, it addresses a theoretical flaw in a prior three-hot modeling scheme. Our experiments show that, even when reducing the number of embedding parameters by >99.6% (from 11.4M to just 36k), our model suffers no loss in translation quality compared to the baseline syllable model.

**Do we need Label Regularization to Fine-tune Pre-trained Language Models?**
*Ivan Kobyzev, Aref Jafari, Mehdi Rezagholizadeh, Tianda Li, Alan Do-omri, Peng Lu, Pascal Poupart and Ali Ghodsi* 11:15-12:45 (Exhibit Hall)
Knowledge Distillation (KD) is a prominent neural model compression technique that heavily relies on teacher network predictions to guide the training of a student model. Considering the ever-growing size of pre-trained language models (PLMs), KD is often adopted in many NLP tasks involving PLMs. However, it is evident that in KD, deploying the teacher network during training adds to the memory and computational requirements of training. In the computer vision literature, the necessity of the teacher network is put under scrutiny by showing that KD is a

label regularization technique that can be replaced with lighter teacher-free variants such as the label-smoothing technique. However, to the best of our knowledge, this issue is not investigated in NLP. Therefore, this work concerns studying different label regularization techniques and whether we actually need them to improve the fine-tuning of smaller PLM networks on downstream tasks. In this regard, we did a comprehensive set of experiments on different PLMs such as BERT, RoBERTa, and GPT with more than 600 distinct trials and ran each configuration five times. This investigation led to a surprising observation that KD and other label regularization techniques do not play any meaningful role over regular fine-tuning when the student model is pre-trained. We further explore this phenomenon in different settings of NLP and computer vision tasks and demonstrate that pre-training itself acts as a kind of regularization, and additional label regularization is unnecessary.

**Evaluating the Robustness of Discrete Prompts**
*Yoichi Ishibashi, Danushka Bollegala, Katsuhito Sudoh and Satoshi Nakamura* 11:15-12:45 (Exhibit Hall)
Discrete prompts have been used for fine-tuning Pre-trained Language Models for diverse NLP tasks. In particular, automatic methods that generate discrete prompts from a small set of training instances have reported superior performance. However, a closer look at the learnt prompts reveals that they contain noisy and counter-intuitive lexical constructs that would not be encountered in manually-written prompts. This raises an important yet understudied question regarding the robustness of automatically learnt discrete prompts when used in downstream tasks. To address this question, we conduct a systematic study of the robustness of discrete prompts by applying carefully designed perturbations into an application using AutoPrompt and then measure their performance in two Natural Language Inference (NLI) datasets. Our experimental results show that although the discrete prompt-based method remains relatively robust against perturbations to NLI inputs, they are highly sensitive to other types of perturbations such as shuffling and deletion of prompt tokens. Moreover, they generalize poorly across different NLI datasets. We hope our findings will inspire future work on robust discrete prompt learning.

**Assessing Out-of-Domain Language Model Performance from Few Examples**
*Prasann Singhal, Jarad Forristal, Xi Ye and Greg Durrett* 11:15-12:45 (Exhibit Hall)
While pretrained language models have exhibited impressive generalization capabilities, they still behave unpredictably under certain domain shifts. In particular, a model may learn a reasoning process on in-domain training data that does not hold for out-of-domain test data. We address the task of predicting out-of-domain (OOD) performance in a few-shot fashion: given a few target-domain examples and a set of models with similar training performance, can we understand how these models will perform on OOD test data? We benchmark the performance on this task when looking at model accuracy on the few-shot examples, then investigate how to incorporate analysis of the models' behavior using feature attributions to better tackle this problem. Specifically, we explore a set of factors designed to reveal model agreement with certain pathological heuristics that may indicate worse generalization capabilities. On textual entailment, paraphrase recognition, and a synthetic classification task, we show that attribution-based factors can help rank relative model OOD performance. However, accuracy on a few-shot test set is a surprisingly strong baseline, particularly when the system designer does not have in-depth prior knowledge about the domain shift.

**Analyzing Challenges in Neural Machine Translation for Software Localization**
*Sai Koneru, Matthias Huck, Miriam Exel and Jan Niehues* 11:15-12:45 (Exhibit Hall)
Advancements in Neural Machine Translation (NMT) greatly benefit the software localization industry by decreasing the post-editing time of human annotators. Although the volume of the software being localized is growing significantly, techniques for improving NMT for user interface (UI) texts are lacking. These UI texts have different properties than other collections of texts, presenting unique challenges for NMT. For example, they are often very short, causing them to be ambiguous and needing additional context (button, title text, a table item, etc.) for disambiguation. However, no such UI data sets are readily available with contextual information for NMT models to exploit. This work aims to provide a first step in improving UI translations and highlight its challenges. To achieve this, we provide a novel multilingual UI corpus collection ($\sim1.3M$ for English $\leftrightarrow$ German) with a targeted test set and analyze the limitations of state-of-the-art methods on this challenging task. Specifically, we present a targeted test set for disambiguation from English to German to evaluate reliably and emphasize UI translation challenges. Furthermore, we evaluate several state-of-the-art NMT techniques from domain adaptation and document-level NMT on this challenging task. All the scripts to replicate the experiments and data sets are available here.

**Bootstrapping Multilingual Semantic Parsers using Large Language Models**
*Abhijeet Awasthi, Nitish Gupta, Bidisha Samanta, Shachi Dave, Sunita Sarawagi and Partha Talukdar* 11:15-12:45 (Exhibit Hall)
Despite cross-lingual generalization demonstrated by pre-trained multilingual models, the translate-train paradigm of transferring English datasets across multiple languages remains to be a key mechanism for training task-specific multilingual models. However, for many low-resource languages, the availability of a reliable translation service entails significant amounts of costly human-annotated translation pairs. Further, translation services may continue to be brittle due to domain mismatch between task-specific input text and general-purpose text used for training translation models. For multilingual semantic parsing, we demonstrate the effectiveness and flexibility offered by large language models (LLMs) for translating English datasets into several languages via few-shot prompting. Through extensive comparisons on two public datasets, MTOP and MASSIVE, spanning 50 languages and several domains, we show that our method of translating data using LLMs outperforms a strong translate-train baseline on 41 out of 50 languages. We study the key design choices that enable more effective multilingual data translation via prompted LLMs.

**Multimodal Graph Transformer for Multimodal Question Answering**
*Xuehai He and Xin Wang* 11:15-12:45 (Exhibit Hall)
Despite the success of Transformer models in vision and language tasks, they often learn knowledge from enormous data implicitly and cannot utilize structured input data directly. On the other hand, structured learning approaches such as graph neural networks (GNNs) that integrate prior information can barely compete with Transformer models. In this work, we aim to benefit from both worlds and propose a novel Multimodal Graph Transformer for question answering tasks that requires performing reasoning across multiple modalities. We introduce a graph-involved plug-and-play quasi-attention mechanism to incorporate multimodal graph information, acquired from text and visual data, to the vanilla self-attention as effective prior. In particular, we construct the text graph, dense region graph, and semantic graph to generate adjacency matrices, and then compose them with input vision and language features to perform downstream reasoning. Such a way of regularizing self-attention with graph information significantly improves the inferring ability and helps align features from different modalities. We validate the effectiveness of Multimodal Graph Transformer over its Transformer baselines on GQA, VQAv2, and MultiModalQA datasets.

**Semantic Parsing for Conversational Question Answering over Knowledge Graphs**
*Laura Perez-beltrachini, Parag Jain, Emilio Monti and Mirella Lapata* 11:15-12:45 (Exhibit Hall)
In this paper, we are interested in developing semantic parsers which understand natural language questions embedded in a conversation with a user and ground them to formal queries over definitions in a general purpose knowledge graph (KG) with very large vocabularies (covering thousands of concept names and relations, and millions of entities). To this end, we develop a dataset where user questions are annotated with Sparql parses and system answers correspond to execution results thereof. We present two different semantic parsing approaches and highlight the challenges of the task: dealing with large vocabularies, modelling conversation context, predicting queries with multiple entities, and generalising to new questions at test time. We hope our dataset will serve as useful testbed for the development of conversational semantic

parsers. Our dataset and models are released at https://github.com/EdinburghNLP/SPICE.

# Main Conference: Thursday, May 4, 2023

# Parallel Session 10 - 09:00-10:30

## Session 10 Orals – Language Resources and Evaluation 1 – Room C

09:00-10:30 (Elafiti 4)

**Investigating UD Treebanks via Dataset Difficulty Measures**
*Artur Kulmizev and Joakim Nivre* 09:00-09:15 (Elafiti 4)
Treebanks annotated with Universal Dependencies (UD) are currently available for over 100 languages and are widely utilized by the community. However, their inherent characteristics are hard to measure and are only partially reflected in parser evaluations via accuracy metrics like LAS. In this study, we analyze a large subset of the UD treebanks using three recently proposed accuracy-free dataset analysis methods: dataset cartography, $\mathcal{V}$-information, and minimum description length. Each method provides insights about UD treebanks that would remain undetected if only LAS was considered. Specifically, we identify a number of treebanks that, despite yielding high LAS, contain very little information that is usable by a parser to surpass what can be achieved by simple heuristics. Furthermore, we make note of several treebanks that score consistently low across numerous metrics, indicating a high degree of noise or annotation inconsistency present therein.

**Vote'n'Rank: Revision of Benchmarking with Social Choice Theory**
*Mark Rofin, Vladislav Mikhailov, Mikhail Florinsky, Andrey Kravchenko, Tatiana Shavrina, Elena Tutubalina, Daniel Karabekyan and Ekaterina Artemova* 09:15-09:30 (Elafiti 4)
The development of state-of-the-art systems in different applied areas of machine learning (ML) is driven by benchmarks, which have shaped the paradigm of evaluating generalisation capabilities from multiple perspectives. Although the paradigm is shifting towards more fine-grained evaluation across diverse tasks, the delicate question of how to aggregate the performances has received particular interest in the community. In general, benchmarks follow the unspoken utilitarian principles, where the systems are ranked based on their mean average score over task-specific metrics. Such aggregation procedure has been viewed as a sub-optimal evaluation protocol, which may have created the illusion of progress. This paper proposes Vote'n'Rank, a framework for ranking systems in multi-task benchmarks under the principles of the social choice theory. We demonstrate that our approach can be efficiently utilised to draw new insights on benchmarking in several ML sub-fields and identify the best-performing systems in research and development case studies. The Vote'n'Rank's procedures are more robust than the mean average while being able to handle missing performance scores and determine conditions under which the system becomes the winner.

**Incorporating Context into Subword Vocabularies**
*Shaked Yehezkel and Yuval Pinter* 09:30-09:45 (Elafiti 4)
Most current popular subword tokenizers are trained based on word frequency statistics over a corpus, without considering information about co-occurrence or context. Nevertheless, the resulting vocabularies are used in language models' highly contextualized settings. We present SaGe, a tokenizer that tailors subwords for their downstream use by baking in the contextualized signal at the vocabulary creation phase. We show that SaGe does a better job than current widespread tokenizers in keeping token contexts cohesive, while not incurring a large price in terms of encoding efficiency or domain robustness. SaGe improves performance on English GLUE classification tasks as well as on NER, and on Inference and NER in Turkish, demonstrating its robustness to language properties such as morphological exponence and agglutination.

**Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions**
*Mihir Parmar, Swaroop Mishra, Mor Geva and Chitta Baral* 09:45-10:00 (Elafiti 4)
In recent years, progress in NLU has been driven by benchmarks. These benchmarks are typically collected by crowdsourcing, where annotators write examples based on annotation instructions crafted by dataset creators. In this work, we hypothesize that annotators pick up on patterns in the crowdsourcing instructions, which bias them to write many similar examples that are then over-represented in the collected data. We study this form of bias, termed instruction bias, in 14 recent NLU benchmarks, showing that instruction examples often exhibit concrete patterns, which are propagated by crowdworkers to the collected data. This extends previous work (Geva et al., 2019) and raises a new concern of whether we are modeling the dataset creator's instructions, rather than the task. Through a series of experiments, we show that, indeed, instruction bias can lead to overestimation of model performance, and that models struggle to generalize beyond biases originating in the crowdsourcing instructions. We further analyze the influence of instruction bias in terms of pattern frequency and model size, and derive concrete recommendations for creating future NLU benchmarks.

**Annotation Error Detection: Analyzing the Past and Present for a More Coherent Future**
*Jan-Christoph Klie, Bonnie Webber and Iryna Gurevych* 10:00-10:15 (Elafiti 4)
Annotated data is an essential ingredient in natural language processing for training and evaluating machine learning models. It is therefore very desirable for the annotations to be of high quality. Recent work, however, has shown that several popular datasets contain a surprising amount of annotation errors or inconsistencies. To alleviate this issue, many methods for annotation error detection have been devised over the years. While researchers show that their approaches work well on their newly introduced datasets, they rarely compare their methods to previous work or on the same datasets. This raises strong concerns on methods' general performance and makes it difficult to asses their strengths and weaknesses. We therefore reimplement 18 methods for detecting potential annotation errors and evaluate them on 9 English datasets for text classification as well as token and span labeling. In addition, we define a uniform evaluation setup including a new formalization of the annotation error detection task, evaluation protocol and general best practices. To facilitate future research and reproducibility, we release our datasets and implementations in an easy-to-use and open source software package.

**ZELDA: A Comprehensive Benchmark for Supervised Entity Disambiguation**
*Marcel Milich and Alan Akbik* 10:15-10:30 (Elafiti 4)
Entity disambiguation (ED) is the task of disambiguating named entity mentions in text to unique entries in a knowledge base. Due to its industrial relevance, as well as current progress in leveraging pre-trained language models, a multitude of ED approaches have been proposed in recent years. However, we observe a severe lack of uniformity across experimental setups in current ED work,rendering a direct comparison of approaches based solely on reported numbers impossible: Current approaches widely differ in the data set used to train, the size of the covered entity vocabulary, and the usage of additional signals such as candidate lists. To address this issue, we present ZELDA , a novel entity disambiguation benchmark that includes a unified training data set, entity vocabulary, candidate lists, as well as challenging evaluation splits covering 8 different domains. We illustrate its design and construction, and present experiments in which we train and compare current state-of-the-art approaches on our benchmark. To encourage greater direct comparability in the entity disambiguation domain, we make our

benchmark publicly available to the research community.

## Session 10 Orals – Lexical Semantics, Discourse and Anaphora – Room B

09:00-10:30 (Elafiti 3)

**A Psycholinguistic Analysis of BERT's Representations of Compounds**
*Lars Buijtelaar and Sandro Pezzelle* 09:00-09:15 (Elafiti 3)
This work studies the semantic representations learned by BERT for compounds, that is, expressions such as sunlight or bodyguard. We build on recent studies that explore semantic information in Transformers at the word level and test whether BERT aligns with human semantic intuitions when dealing with expressions (e.g., sunlight) whose overall meaning depends—to a various extent—on the semantics of the constituent words (sun, light). We leverage a dataset that includes human judgments on two psycholinguistic measures of compound semantic analysis: lexeme meaning dominance (LMD; quantifying the weight of each constituent toward the compound meaning) and semantic transparency (ST; evaluating the extent to which the compound meaning is recoverable from the constituents' semantics). We show that BERT-based measures moderately align with human intuitions, especially when using contextualized representations, and that LMD is overall more predictable than ST. Contrary to the results reported for 'standard' words, higher, more contextualized layers are the best at representing compound meaning. These findings shed new light on the abilities of BERT in dealing with fine-grained semantic phenomena. Moreover, they can provide insights into how speakers represent compounds.

**A Systematic Search for Compound Semantics in Pretrained BERT Architectures**
*Filip Miletic and Sabine Schulte Im Walde* 09:15-09:30 (Elafiti 3)
To date, transformer-based models such as BERT have been less successful in predicting compositionality of noun compounds than static word embeddings. This is likely related to a suboptimal use of the encoded information, reflecting an incomplete grasp of how the models represent the meanings of complex linguistic structures. This paper investigates variants of semantic knowledge derived from pretrained BERT when predicting the degrees of compositionality for 280 English noun compounds associated with human compositionality ratings. Our performance strongly improves on earlier unsupervised implementations of pretrained BERT and highlights beneficial decisions in data preprocessing, embedding computation, and compositionality estimation. The distinct linguistic roles of heads and modifiers are reflected by differences in BERT-derived representations, with empirical properties such as frequency, productivity, and ambiguity affecting model performance. The most relevant representational information is concentrated in the initial layers of the model architecture.

**Bridging the Gap Between BabelNet and HowNet: Unsupervised Sense Alignment and Sememe Prediction**
*Xiang Zhang, Ning Shi, Bradley Hauer and Grzegorz Kondrak* 09:30-09:45 (Elafiti 3)
As the minimum semantic units of natural languages, sememes can provide precise representations of concepts. Despite the widespread utilization of lexical resources for semantic tasks, use of sememes is limited by a lack of available sememe knowledge bases. Recent efforts have been made to connect BabelNet with HowNet by automating sememe prediction. However, these methods depend on large manually annotated datasets. We propose to use sense alignment via a novel unsupervised and explainable method. Our method consists of four stages, each relaxing predefined constraints until a complete alignment of BabelNet synsets to HowNet senses is achieved. Experimental results demonstrate the superiority of our unsupervised method over previous supervised ones by an improvement of 12% overall F1 score, setting a new state of the art. Our work is grounded in an interpretable propagation of sememe information between lexical resources, and may benefit downstream applications which can incorporate sememe information.

**What happens before and after: Multi-Event Commonsense in Event Coreference Resolution**
*Sahithya Ravi, Chris Tanner, Raymond Ng and Vered Shwartz* 09:45-10:00 (Elafiti 3)
Event coreference models cluster event mentions pertaining to the same real-world event. Recent models rely on contextualized representations to recognize coreference among lexically or contextually similar mentions. However, models typically fail to leverage commonsense inferences, which is particularly limiting for resolving lexically-divergent mentions. We propose a model that extends event mentions with temporal commonsense inferences. Given a complex sentence with multiple events, e.g., "the man killed his wife and got arrested", with the target event "arrested", our model generates plausible events that happen before the target event – such as "the police arrived", and after it, such as "he was sentenced". We show that incorporating such inferences into an existing event coreference model improves its performance, and we analyze the coreferences in which such temporal knowledge is required.

**A simple but effective model for attachment in discourse parsing with multi-task learning for relation labeling**
*Zineb Bennis, Julie Hunter and Nicholas Asher* 10:00-10:15 (Elafiti 3)
In this paper, we present a discourse parsing model for conversation trained on the STAC. We fine-tune a BERT-based model to encode pairs of discourse units and use a simple linear layer to predict discourse attachments. We then exploit a multi-task setting to predict relation labels. The multitask approach effectively aids in the difficult task of relation type prediction; our f1 score of 57 surpasses the state of the art with no loss in performance for attachment, confirming the intuitive interdependence of these two tasks. Our method also improves over previous discourse parsing models in allowing longer input sizes and in permitting attachments in which one node has multiple parents, an important feature of multiparty conversation.

**Exploring Category Structure with Contextual Language Models and Lexical Semantic Networks**
*Joseph Renner, Pascal Denis, Remi Gilleron and Angèle Brunellière* 10:15-10:30 (Elafiti 3)
The psychological plausibility of word embeddings has been studied through different tasks such as word similarity, semantic priming, and lexical entailment. Recent work on predicting category structure with word embeddings report low correlations with human ratings. (Heyman and Heyman, 2019) showed that static word embeddings fail at predicting typicality using cosine similarity between category and exemplar words, while (Misra et al., 2021)obtain equally modest results for various contextual language models (CLMs) using a Cloze task formulation over hand-crafted taxonomic sentences. In this work, we test a wider array of methods for probing CLMs for predicting typicality scores. Our experiments, using BERT (Devlin et al., 2018), show the importance of using the right type of CLM probes, as our best BERT-based typicality prediction methods improve on previous works. Second, our results highlight the importance of polysemy in this task, as our best results are obtained when contextualization is paired with a disambiguation mechanism as in (Chronis and Erk, 2020). Finally, additional experiments and analyses reveal that Information Content-based WordNet (Miller, 1995) similarities with disambiguation match the performance of the best BERT-based method, and in fact capture complementary information, and when combined with BERT allow for enhanced typicality predictions.

## Session 10 Orals – Machine Translation and Multilinguality – Room A

09:00-10:30 (Elafiti 2)

**Automatic Evaluation and Analysis of Idioms in Neural Machine Translation**
*Christos Baziotis, Prashant Mathur and Eva Hasler* 09:00-09:15 (Elafiti 2)
A major open problem in neural machine translation (NMT) is the translation of idiomatic expressions, such as "under the weather". The meaning of these expressions is not composed by the meaning of their constituent words, and NMT models tend to translate them literally (i.e., word-by-word), which leads to confusing and nonsensical translations. Research on idioms in NMT is limited and obstructed by the absence of automatic methods for quantifying these errors. In this work, first, we propose a novel metric for automatically measuring the frequency of literal translation errors without human involvement. Equipped with this metric, we present controlled translation experiments with models trained in different conditions (with/without the test-set idioms) and across a wide range of (global and targeted) metrics and test sets. We explore the role of monolingual pretraining and find that it yields substantial targeted improvements, even without observing any translation examples of the test-set idioms. In our analysis, we probe the role of idiom context. We find that the randomly initialized models are more local or "myopic" as they are relatively unaffected by variations of the idiom context, unlike the pretrained ones.

**CTC Alignments Improve Autoregressive Translation**
*Brian Yan, Siddharth Dalmia, Yosuke Higuchi, Graham Neubig, Florian Metze, Alan W Black and Shinji Watanabe* 09:15-09:30 (Elafiti 2)
Connectionist Temporal Classification (CTC) is a widely used approach for automatic speech recognition (ASR) that performs conditionally independent monotonic alignment. However for translation, CTC exhibits clear limitations due to the contextual and non-monotonic nature of the task and thus lags behind attentional decoder approaches in terms of translation quality. In this work, we argue that CTC does in fact make sense for translation if applied in a joint CTC/attention framework wherein CTC's core properties can counteract several key weaknesses of pure-attention models during training and decoding. To validate this conjecture, we modify the Hybrid CTC/Attention model originally proposed for ASR to support text-to-text translation (MT) and speech-to-text translation (ST). Our proposed joint CTC/attention models outperform pure-attention baselines across six benchmark translation tasks.

**Exploring Segmentation Approaches for Neural Machine Translation of Code-Switched Egyptian Arabic-English Text**
*Marwa Gaser, Manuel Mager, Injy Hamed, Nizar Habash, Slim Abdennadher and Ngoc Thang Vu* 09:30-09:45 (Elafiti 2)
Data sparsity is one of the main challenges posed by code-switching (CS), which is further exacerbated in the case of morphologically rich languages. For the task of machine translation (MT), morphological segmentation has proven successful in alleviating data sparsity in monolingual contexts; however, it has not been investigated for CS settings. In this paper, we study the effectiveness of different segmentation approaches on MT performance, covering morphology-based and frequency-based segmentation techniques. We experiment on MT from code-switched Arabic-English to English. We provide detailed analysis, examining a variety of conditions, such as data size and sentences with different degrees of CS. Empirical results show that morphology-aware segmenters perform the best in segmentation tasks but underperform in MT. Nevertheless, we find that the choice of the segmentation setup to use for MT is highly dependent on the data size. For extreme low-resource scenarios, a combination of frequency and morphology-based segmentations is shown to perform the best. For more resourced settings, such a combination does not bring significant improvements over the use of frequency-based segmentation.

**Exploring Paracrawl for Document-level Neural Machine Translation**
*Yusser Al Ghussin, Jingyi Zhang and Josef Van Genabith* 09:45-09:55 (Elafiti 2)
Document-level neural machine translation (NMT) has outperformed sentence-level NMT on a number of datasets. However, document-level NMT is still not widely adopted in realworld translation systems mainly due to the lack of large-scale general-domain training data for document-level NMT. We examine the effectiveness of using Paracrawl for learning document-level translation. Paracrawl is a large-scale parallel corpus crawled from the Internet and contains data from various domains. The official Paracrawl corpus was released as parallel sentences (extracted from parallel webpages) and therefore previous works only used Paracrawl for learning sentence-level translation. In this work, we extract parallel paragraphs from Paracrawl parallel webpages using automatic sentence alignments and we use the extracted parallel paragraphs as parallel documents for training document-level translation models. We show that document-level NMT models trained with only parallel paragraphs from Paracrawl can be used to translate real documents from TED, News and Europarl, outperforming sentence-level NMT models. We also perform a targeted pronoun evaluation and show that document-level models trained with Paracrawl data can help context-aware pronoun translation.

**Robustification of Multilingual Language Models to Real-world Noise in Crosslingual Zero-shot Settings with Robust Contrastive Pretraining**
*Asa Cooper Stickland, Sailik Sengupta, Jason Krone, Saab Mansour and He He* 09:55-10:10 (Elafiti 2)
Advances in neural modeling have achieved state-of-the-art (SOTA) results on public natural language processing (NLP) benchmarks, at times surpassing human performance. However, there is a gap between public benchmarks and real-world applications where noise, such as typographical or grammatical mistakes, is abundant and can result in degraded performance. Unfortunately, works which evaluate the robustness of neural models on noisy data and propose improvements, are limited to the English language. Upon analyzing noise in different languages, we observe that noise types vary greatly across languages. Thus, existing investigations do not generalize trivially to multilingual settings. To benchmark the performance of pretrained multilingual language models, we construct noisy datasets covering five languages and four NLP tasks and observe a clear gap in the performance between clean and noisy data in the zero-shot cross-lingual setting. After investigating several ways to boost the robustness of multilingual models in this setting, we propose Robust Contrastive Pretraining (RCP). RCP combines data augmentation with a contrastive loss term at the pretraining stage and achieves large improvements on noisy (and original test data) across two sentence-level (+3.2%) and two sequence-labeling (+10 F1-score) multilingual classification tasks.

**Efficiently Upgrading Multilingual Machine Translation Models to Support More Languages**
*Simeng Sun, Maha Elbayad, Anna Sun and James Cross* 10:10-10:25 (Elafiti 2)
With multilingual machine translation (MMT) models continuing to grow in size and number of supported languages, it is natural to reuse and upgrade existing models to save computation as data becomes available in more languages. However, adding new languages requires updating the vocabulary, which complicates the reuse of embeddings. The question of how to reuse existing models while also making architectural changes to provide capacity for both old and new languages has also not been closely studied. In this work, we introduce three techniques that help speed up the effective learning of new languages and alleviate catastrophic forgetting despite vocabulary and architecture mismatches. Our results show that by (1) carefully initializing the network, (2) applying learning rate scaling, and (3) performing data up-sampling, it is possible to exceed the performance of a same-sized baseline model with 30\% computation and recover the performance of a larger model trained from scratch with over 50\% reduction in computation. Furthermore, our analysis reveals that the introduced techniques help learn new directions more effectively and alleviate catastrophic forgetting at the same time. We hope our work will guide research into more efficient approaches to growing languages for these MMT models and ultimately maximize the reuse of existing models.

# Parallel Session 11 - 11:15-12:45

## Session 11 Orals – Large Language Models – Room C

11:15-12:45 (Elafiti 4)

**A Discerning Several Thousand Judgments: GPT-3 Rates the Article + Adjective + Numeral + Noun Construction**
*Kyle Mahowald* 11:15-11:30 (Elafiti 4)
Knowledge of syntax includes knowledge of rare, idiosyncratic constructions. LLMs must overcome frequency biases in order to master such constructions. In this study, I prompt GPT-3 to give acceptability judgments on the English-language Article + Adjective + Numeral + Noun construction (e.g., "a lovely five days"). I validate the prompt using the CoLA corpus of acceptability judgments and then zero in on the AANN construction. I compare GPT- 3's judgments to crowdsourced human judgments on a subset of sentences. GPT-3's judgments are broadly similar to human judgments and generally align with proposed constraints in the literature but, in some cases, GPT-3's judgments and human judgments diverge from the literature and from each other.

**MiniALBERT: Model Distillation via Parameter-Efficient Recursive Transformers**
*Mohammadmahdi Nouriborji, Omid Rohanian, Samaneh Kouchaki and David A. Clifton* 11:30-11:45 (Elafiti 4)
Pre-trained Language Models (LMs) have become an integral part of Natural Language Processing (NLP) in recent years, due to their superior performance in downstream applications. In spite of this resounding success, the usability of LMs is constrained by computational and time complexity, along with their increasing size; an issue that has been referred to as overparameterisation. Different strategies have been proposed in the literature to alleviate these problems, with the aim to create effective compact models that nearly match the performance of their bloated counterparts with negligible performance losses. One of the most popular techniques in this area of research is model distillation. Another potent but underutilised technique is cross-layer parameter sharing. In this work, we combine these two strategies and present MiniALBERT, a technique for converting the knowledge of fully parameterised LMs (such as BERT) into a compact recursive student. In addition, we investigate the application of bottleneck adapters for layer-wise adaptation of our recursive student, and also explore the efficacy of adapter tuning for fine-tuning of compact models. We test our proposed models on a number of general and biomedical NLP tasks to demonstrate their viability and compare them with the state-of-the-art and other existing compact models. All the codes used in the experiments and the pre-trained compact models will be made publicly available.

**Methods for Measuring, Updating, and Visualizing Factual Beliefs in Language Models**
*Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal and Srinivasan Iyer* 11:45-12:00 (Elafiti 4)
Language models can memorize a considerable amount of factual information during pretraining that can be elicited through prompting or finetuning models on tasks like question answering. In this paper, we discuss approaches to measuring model factual beliefs, updating incorrect factual beliefs in models, and visualizing graphical relationships between factual beliefs. Our main contributions include: (1) new metrics for evaluating belief-updating methods focusing on the logical consistency of beliefs, (2) a training objective for Sequential, Local, and Generalizing updates (SLAG) that improves the performance of existing hypernetwork approaches, and (3) the introduction of the belief graph, a new form of visualization for language models that shows relationships between stored model beliefs. Our experiments suggest that models show only limited consistency between factual beliefs, but update methods can both fix incorrect model beliefs and greatly improve their consistency. Although off-the-shelf optimizers are surprisingly strong belief-updating baselines, our learned optimizers can outperform them in more difficult settings than have been considered in past work.

**How to Dissect a Muppet: The Structure of Transformer Embedding Spaces**
*Timothee Mickus, Denis Paperno and Mathieu Constant* 12:00-12:15 (Elafiti 4)
Pretrained embeddings based on the Transformer architecture have taken the NLP community by storm. We show that they can mathematically be reframed as a sum of vector factors and showcase how to use this reframing to study the impact of each component. We provide evidence that multi-head attentions and feed-forwards are not equally useful in all downstream applications, as well as a quantitative overview of the effects of finetuning on the overall embedding space. This approach allows us to draw connections to a wide range of previous studies, from vector space anisotropy to attention weights.

**Do Pretrained Contextual Language Models Distinguish between Hebrew Homograph Analyses?**
*Avi Shmidman, Cheyn Shmidman, Dan Bareket, Moshe Koppel and Reut Tsarfaty* 12:15-12:25 (Elafiti 4)
Semitic morphologically-rich languages (MRLs) are characterized by extreme word ambiguity. Because most vowels are omitted in standard texts, many of the words are homographs with multiple possible analyses, each with a different pronunciation and different morphosyntactic properties. This ambiguity goes {\em beyond} word-sense disambiguation (WSD), and may include token segmentation into multiple word units. Previous research on MRLs claimed that standardly trained pre-trained language models (PLMs) based on word-pieces may not sufficiently capture the internal structure of such tokens in order to distinguish between these analyses. Taking Hebrew as a case study, we investigate the extent to which Hebrew homographs can be disambiguated and analyzed using PLMs. We evaluate all existing models for contextualized Hebrew embeddings on a novel Hebrew homograph challenge sets that we deliver. Our empirical results demonstrate that contemporary Hebrew contextualized embeddings outperform non-contextualized embeddings; and that they are most effective for disambiguating segmentation and morphosyntactic features, less so regarding pure word-sense disambiguation. We show that these embeddings are more effective when the number of word-piece splits is limited, and they are more effective for 2-way and 3-way ambiguities than for 4-way ambiguity. We show that the embeddings are equally effective for homographs of both balanced and skewed distributions, whether calculated as masked or unmasked tokens. Finally, we show that these embeddings are as effective for homograph disambiguation with extensive supervised training as with a few-shot setup.

**WinoDict: Probing language models for in-context word acquisition**
*Julian Eisenschlos, Jeremy Cole, Fangyu Liu and William Cohen* 12:25-12:35 (Elafiti 4)
We introduce a new in-context learning paradigm to measure Large Language Models' (LLMs) ability to learn novel words during inference. In particular, we rewrite Winograd-style co-reference resolution problems by replacing the key concept word with a synthetic but plausible word that the model must understand to complete the task. Solving this task requires the model to make use of the dictionary definition of the new word given in the prompt. This benchmark addresses word acquisition, one important aspect of the diachronic degradation known to afflict LLMs. As LLMs are frozen in time at the moment they are trained, they are normally unable to reflect the way language changes over time. We show that the accuracy of LLMs compared to the original Winograd tasks decreases radically in our benchmark, thus identifying a limitation of current models and providing a benchmark to measure future improvements in LLMs ability to do in-context learning.

## Session 11 Orals – Question Generation and Answering – Room A

11:15-12:45 (Elafiti 2)

**Closed-book Question Generation via Contrastive Learning**
*Xiangjue Dong, Jiaying Lu, Jianling Wang and James Caverlee* 11:15-11:30 (Elafiti 2)
Question Generation (QG) is a fundamental NLP task for many downstream applications. Recent studies on open-book QG, where supportive answer-context pairs are provided to models, have achieved promising progress. However, generating natural questions under a more practical closed-book setting that lacks these supporting documents still remains a challenge. In this work, we propose a new QG model for this closed-book setting that is designed to better understand the semantics of long-form abstractive answers and store more information in its parameters through contrastive learning and an answer reconstruction module. Through experiments, we validate the proposed QG model on both public datasets and a new WikiCQA dataset. Empirical results show that the proposed QG model outperforms baselines in both automatic evaluation and human evaluation. In addition, we show how to leverage the proposed model to improve existing question-answering systems. These results further indicate the effectiveness of our QG model for enhancing closed-book question-answering tasks.

**CylE: Cylinder Embeddings for Multi-hop Reasoning over Knowledge Graphs**
*Chau Nguyen, Tim French, Wei Liu and Michael Stewart* 11:30-11:45 (Elafiti 2)
Recent geometric-based approaches have been shown to efficiently model complex logical queries (including the intersection operation) over Knowledge Graphs based on the natural representation of Venn diagram. Existing geometric-based models (using points, boxes embeddings), however, cannot handle the logical negation operation. Further, those using cones embeddings are limited to representing queries by two-dimensional shapes, which reduced their effectiveness in capturing entities query relations for correct answers. To overcome this challenge, we propose unbounded cylinder embeddings (namely CylE), which is a novel geometric-based model based on three-dimensional shapes. Our approach can handle a complete set of basic first-order logic operations (conjunctions, disjunctions and negations). CylE considers queries as Cartesian products of unbounded sector-cylinders and consider a set of nearest boxes corresponds to the set of answer entities. Precisely, the conjunctions can be represented via the intersections of unbounded sector-cylinders. Transforming queries to Disjunctive Normal Form can handle queries with disjunctions. The negations can be represented by considering the closure of complement for an arbitrary unbounded sector-cylinder. Empirical results show that the performance of multi-hop reasoning task using CylE significantly increases over state-of-the-art geometric-based query embedding models for queries without negation. For queries with negation operations, though the performance is on a par with the best performing geometric-based model, CylE significantly outperforms a recent distribution-based model.

**Retrieval Enhanced Data Augmentation for Question Answering on Privacy Policies**
*Md Rizwan Parvez, Jianfeng Chi, Wasi Uddin Ahmad, Yuan Tian and Kai-wei Chang* 11:45-11:55 (Elafiti 2)
Prior studies in privacy policies frame the question answering (QA) task as identifying the most relevant text segment or a list of sentences from a policy document given a user query. Existing labeled datasets are heavily imbalanced (only a few relevant segments), limiting the QA performance in this domain. In this paper, we develop a data augmentation framework based on ensembling retriever models that captures the relevant text segments from unlabeled policy documents and expand the positive examples in the training set. In addition, to improve the diversity and quality of the augmented data, we leverage multiple pre-trained language models (LMs) and cascaded them with noise reduction oracles. Using our augmented data on the PrivacyQA benchmark, we elevate the existing baseline by a large margin (10% F1) and achieve a new state-of-the-art F1 score of 50%. Our ablation studies provide further insights into the effectiveness of our approach.

**Question Generation Using Sequence-to-Sequence Model with Semantic Role Labels**
*Alireza Naeiji, Aijun An, Heidar Davoudi, Marjan Delpisheh and Muath Alzghool* 11:55-12:10 (Elafiti 2)
Automatic generation of questions from text has gained increasing attention due to its useful applications. We propose a novel question generation method that combines the benefits of rule-based and neural sequence-to-sequence (Seq2Seq) models. The proposed method can automatically generate multiple questions from an input sentence covering different views of the sentence as in rule-based methods, while more complicated "rules" can be learned via the Seq2Seq model. The method utilizes semantic role labeling to convert training examples into their semantic representations, and then trains a Seq2Seq model over the semantic representations. Our extensive experiments on three real-world data sets show that the proposed method significantly improves the state-of-the-art neural question generation approaches.

**Question-Answer Sentence Graph for Joint Modeling Answer Selection**
*Roshni Iyer, Thuy Vu, Alessandro Moschitti and Yizhou Sun* 12:10-12:25 (Elafiti 2)
This research studies graph-based approaches for Answer Sentence Selection (AS2), an essential component for retrieval-based Question Answering (QA) systems. During offline learning, our model constructs a small-scale relevant training graph per question in an unsupervised manner, and integrates with Graph Neural Networks. Graph nodes are question sentence to answer sentence pairs. We train and integrate state-of-the-art (SOTA) models for computing scores between question-question, question-answer, and answer-answer pairs, and use thresholding on relevance scores for creating graph edges. Online inference is then performed to solve the AS2 task on unseen queries. Experiments on two well-known academic benchmarks and a real-world dataset show that our approach consistently outperforms SOTA QA baseline models.

**Socratic Question Generation: A Novel Dataset, Models, and Evaluation**
*Beng Heng Ang, Sujatha Das Gollapalli and See-kiong Ng* 12:25-12:40 (Elafiti 2)
Socratic questioning is a form of reflective inquiry often employed in education to encourage critical thinking in students, and to elicit awareness of beliefs and perspectives in a subject during therapeutic counseling. Specific types of Socratic questions are employed for enabling reasoning and alternate views against the context of individual personal opinions on a topic. Socratic contexts are different from traditional question generation contexts where "answer-seeking" questions are generated against a given formal passage on a topic, narrative stories or conversations.

We present SocratiQ, the first large dataset of 110K (question, context) pairs for enabling studies on Socratic Question Generation (SoQG). We provide an in-depth study on the various types of Socratic questions and present models for generating Socratic questions against a given context through prompt tuning. Our automated and human evaluation results demonstrate that our SoQG models can produce realistic, type-sensitive, human-like Socratic questions enabling potential applications in counseling and coaching.

## Session 11 Orals – Semantics: Sentence level and Other areas – Room B

11:15-12:45 (Elafiti 3)

**A Kind Introduction to Lexical and Grammatical Aspect, with a Survey of Computational Approaches**
*Annemarie Friedrich, Nianwen Xue and Alexis Palmer* 11:15-11:30 (Elafiti 3)
Aspectual meaning refers to how the internal temporal structure of situations is presented. This includes whether a situation is described as a state or as an event, whether the situation is finished or ongoing, and whether it is viewed as a whole or with a focus on a particular phase. This survey gives an overview of computational approaches to modeling lexical and grammatical aspect along with intuitive explanations of the necessary linguistic concepts and terminology. In particular, we describe the concepts of stativity, telicity, habituality, perfective and imperfective, as well as influential inventories of eventuality and situation types. Aspect is a crucial component of semantics, especially for precise reporting of the temporal structure of situations, and future NLP approaches need to be able to handle and evaluate it systematically.

**End-to-end Case-Based Reasoning for Commonsense Knowledge Base Completion**
*Zonglin Yang, Xinya Du, Erik Cambria and Claire Cardie* 11:30-11:45 (Elafiti 3)
Pretrained language models have been shown to store knowledge in their parameters and have achieved reasonable performance in commonsense knowledge base completion (CKBC) tasks. However, CKBC is knowledge-intensive and it is reported that pretrained language models' performance in knowledge-intensive tasks are limited because of their incapability of accessing and manipulating knowledge. As a result, we hypothesize that providing retrieved passages that contain relevant knowledge as additional input to the CKBC task will improve performance. In particular, we draw insights from Case-Based Reasoning (CBR) – which aims to solve a new problem by reasoning with retrieved relevant cases, and investigate the direct application of it to CKBC. On two benchmark datasets, we demonstrate through automatic and human evaluations that our End-to-end Case-Based Reasoning Framework (ECBRF) generates more valid, informative, and novel knowledge than the state-of-the-art COMET model for CKBC in both the fully supervised and few-shot settings. We provide insights on why previous retrieval-based methods only achieve merely the same performance with COMET. From the perspective of CBR, our framework addresses a fundamental question on whether CBR methodology can be utilized to improve deep learning models.

**Identifying the limits of transformers when performing model-checking with natural language**
*Tharindu Madusanka, Riza Batista-navarro and Ian Pratt-hartmann* 11:45-12:00 (Elafiti 3)
Can transformers learn to comprehend logical semantics in natural language? Although many strands of work on natural language inference have focussed on transformer models' ability to perform reasoning on text, the above question has not been answered adequately. This is primarily because the logical problems that have been studied in the context of natural language inference have their computational complexity vary with the logical and grammatical constructs within the sentences. As such, it is difficult to access whether the difference in accuracy is due to logical semantics or the difference in computational complexity. A problem that is much suited to address this issue is that of the model-checking problem, whose computational complexity remains constant (for fragments derived from first-order logic). However, the model-checking problem remains untouched in natural language inference research. Thus, we investigated the problem of model-checking with natural language to adequately answer the question of how the logical semantics of natural language affects transformers' performance. Our results imply that the language fragment has a significant impact on the performance of transformer models. Furthermore, we hypothesise that a transformer model can at least partially understand the logical semantics in natural language but can not completely learn the rules governing the model-checking algorithm.

**Penguins Don't Fly: Reasoning about Generics through Instantiations and Exceptions**
*Emily Allaway, Jena D. Hwang, Chandra Bhagavatula, Kathleen Mckeown, Doug Downey and Yejin Choi* 12:00-12:15 (Elafiti 3)
Generics express generalizations about the world (e.g., birds can fly) that are not universally true (e.g., newborn birds and penguins cannot fly). Commonsense knowledge bases, used extensively in NLP, encode some generic knowledge but rarely enumerate such exceptions and knowing when a generic statement holds or does not hold true is crucial for developing a comprehensive understanding of generics. We present a novel framework informed by linguistic theory to generate exemplars—specific cases when a generic holds true or false. We generate ~19k exemplars for ~650 generics and show that our framework outperforms a strong GPT-3 baseline by 12.8 precision points. Our analysis highlights the importance of linguistic theory-based controllability for generating exemplars, the insufficiency of knowledge bases as a source of exemplars, and the challenges exemplars pose for the task of natural language inference.

**Retrieve-and-Fill for Scenario-based Task-Oriented Semantic Parsing**
*Akshat Shrivastava, Shrey Desai, Anchit Gupta, Ali Elkahky, Aleksandr Livshits, Alexander Zotov and Ahmed Aly* 12:15-12:30 (Elafiti 3)
Task-oriented semantic parsing models have achieved strong results in recent years, but unfortunately do not strike an appealing balance between model size, runtime latency, and cross-domain generalizability. We tackle this problem by introducing scenario-based semantic parsing: a variant of the original task which first requires disambiguating an utterance's "scenario" (an intent-slot template with variable leaf spans) before generating its frame, complete with ontology and utterance tokens. This formulation enables us to isolate coarse-grained and fine-grained aspects of the task, each of which we solve with off-the-shelf neural modules, also optimizing for the axes outlined above. Concretely, we create a Retrieve-and-Fill (RAF) architecture comprised of (1) a retrieval module which ranks the best scenario given an utterance and (2) a filling module which imputes spans into the scenario to create the frame. Our model is modular, differentiable, interpretable, and allows us to garner extra supervision from scenarios. RAF achieves strong results in high-resource, low-resource, and multilingual settings, outperforming recent approaches by wide margins despite, using base pre-trained encoders, small sequence lengths, and parallel decoding.

**Semantic Parsing for Conversational Question Answering over Knowledge Graphs**
*Laura Perez-beltrachini, Parag Jain, Emilio Monti and Mirella Lapata* 12:30-12:45 (Elafiti 3)
In this paper, we are interested in developing semantic parsers which understand natural language questions embedded in a conversation with a user and ground them to formal queries over definitions in a general purpose knowledge graph (KG) with very large vocabularies (covering thousands of concept names and relations, and millions of entities). To this end, we develop a dataset where user questions are annotated with Sparql parses and system answers correspond to execution results thereof. We present two different semantic parsing approaches and highlight the challenges of the task: dealing with large vocabularies, modelling conversation context, predicting queries with multiple entities, and generalising to new questions at test time. We hope our dataset will serve as useful testbed for the development of conversational semantic parsers. Our dataset and models are released at https://github.com/EdinburghNLP/SPICE.

## Session 11 Posters

11:15-12:45 (Exhibit Hall)

**Multi2Claim: Generating Scientific Claims from Multi-Choice Questions for Scientific Fact-Checking**

*Neset Tan, Trung Nguyen, Josh Bensemann, Alex Peng, Qiming Bao, Yang Chen, Mark Gahegan and Michael Witbrock* 11:15-12:45 (Exhibit Hall)
Training machine learning models to successfully perform scientific fact-checking tasks is challenging due to the expertise bottleneck that limits the availability of appropriate training datasets. In this task, models use textual evidence to confirm scientific claims, which requires data that contains extensive domain-expert annotation. Consequently, the number of existing scientific-fact-checking datasets and the sizes of those datasets are limited. However, these limitations do not apply to multiple-choice question datasets because of the necessity of domain exams in the modern education system. As one of the first steps towards addressing the fact-checking dataset scarcity problem in scientific domains, we propose a pipeline for automatically converting multiple-choice questions into fact-checking data, which we call Multi2Claim. By applying the proposed pipeline, we generated two large-scale datasets for scientific-fact-checking tasks: Med-Fact and Gsci-Fact for the medical and general science domains, respectively. These two datasets are among the first examples of large-scale scientific-fact-checking datasets. We developed baseline models for the verdict prediction task using each dataset. Additionally, we demonstrated that the datasets could be used to improve performance with respect to the F 1 weighted metric on existing fact-checking datasets such as SciFact, HEALTHVER, COVID-Fact, and CLIMATE-FEVER. In some cases, the improvement in performance was up to a 26% increase.

**SwitchPrompt: Learning Domain-Specific Gated Soft Prompts for Classification in Low-Resource Domains**
*Koustava Goswami, Lukas Lange, Jun Araki and Heike Adel* 11:15-12:45 (Exhibit Hall)
Prompting pre-trained language models leads to promising results across natural language processing tasks but is less effective when applied in low-resource domains, due to the domain gap between the pre-training data and the downstream task. In this work, we bridge this gap with a novel and lightweight prompting methodology called SwitchPrompt for the adaptation of language models trained on datasets from the general domain to diverse low-resource domains. Using domain-specific keywords with a trainable gated prompt, SwitchPrompt offers domain-oriented prompting, that is, effective guidance on the target domains for general-domain language models. Our few-shot experiments on three text classification benchmarks demonstrate the efficacy of the general-domain pre-trained language models when used with SwitchPrompt. They often even outperform their domain-specific counterparts trained with baseline state-of-the-art prompting methods by up to 10.7% performance increase in accuracy. This result indicates that SwitchPrompt effectively reduces the need for domain-specific language model pre-training.

**Parameter-efficient Modularised Bias Mitigation via AdapterFusion**
*Deepak Kumar, Oleg Lesota, George Zerveas, Daniel Cohen, Carsten Eickhoff, Markus Schedl and Navid Rekabsaz* 11:15-12:45 (Exhibit Hall)
Large pre-trained language models contain societal biases and carry along these biases to downstream tasks. Current in-processing bias mitigation approaches (like adversarial training) impose debiasing by updating a model's parameters, effectively transferring the model to a new, irreversible debiased state. In this work, we propose a novel approach to develop stand-alone debiasing functionalities separate from the model, which can be integrated into the model on-demand, while keeping the core model untouched. Drawing from the concept of AdapterFusion in multi-task learning, we introduce DAM (Debiasing with Adapter Modules) – a debiasing approach to first encapsulate arbitrary bias mitigation functionalities into separate adapters, and then add them to the model on-demand in order to deliver fairness qualities. We conduct a large set of experiments on three classification tasks with gender, race, and age as protected attributes. Our results show that DAM improves or maintains the effectiveness of bias mitigation, avoids catastrophic forgetting in a multi-attribute scenario, and maintains on-par task performance, while granting parameter-efficiency and easy switching between the original and debiased models.

**The StatCan Dialogue Dataset: Retrieving Data Tables through Conversations with Genuine Intents**
*Xing Han Lu, Siva Reddy and Harm De Vries* 11:15-12:45 (Exhibit Hall)
We introduce the StatCan Dialogue Dataset consisting of 19,379 conversation turns between agents working at Statistics Canada and online users looking for published data tables. The conversations stem from genuine intents, are held in English or French, and lead to agents retrieving one of over 5000 complex data tables. Based on this dataset, we propose two tasks: (1) automatic retrieval of relevant tables based on a on-going conversation, and (2) automatic generation of appropriate agent responses at each turn. We investigate the difficulty of each task by establishing strong baselines. Our experiments on a temporal data split reveal that all models struggle to generalize to future conversations, as we observe a significant drop in performance across both tasks when we move from the validation to the test set. In addition, we find that response generation models struggle to decide when to return a table. Considering that the tasks pose significant challenges to existing models, we encourage the community to develop models for our task, which can be directly used to help knowledge workers find relevant tables for live chat users.

**Probabilistic Robustness for Data Filtering**
*Yu Yu, Abdul Khan, Shahram Khadivi and Jia Xu* 11:15-12:45 (Exhibit Hall)
We introduce our probabilistic robustness rewarded data optimization (PRoDO) approach as a framework to enhance the model's generalization power by selecting training data that optimizes our probabilistic robustness metrics. We use proximal policy optimization (PPO) reinforcement learning to approximately solve the computationally intractable training subset selection problem. The PPO's reward is defined as our ($\alpha,\epsilon, \gamma$)-Robustness that measures performance consistency over multiple domains by simulating unknown test sets in real-world scenarios using a leaving-one-out strategy. We demonstrate that our PRoDO effectively filters data that lead to significantly higher prediction accuracy and robustness on unknown-domain test sets. Our experiments achieve up to +17.2\% increase of accuracy (+25.5\% relatively) in sentiment analysis, and -28.05 decrease of perplexity (-32.1\% relatively) in language modeling. In addition, our probabilistic ($\alpha,\epsilon, \gamma$)-Robustness definition serves as an evaluation metric with higher levels of agreement with human annotations than typical performance-based metrics.

**Contextual Dynamic Prompting for Response Generation in Task-oriented Dialog Systems**
*Sandesh Swamy, Narges Tabari, Chacha Chen and Rashmi Gangadharaiah* 11:15-12:45 (Exhibit Hall)
Response generation is one of the critical components in task-oriented dialog systems. Existing studies have shown that large pre-trained language models can be adapted to this task. The typical paradigm of adapting such extremely large language models would be by fine-tuning on the downstream tasks which is not only time-consuming but also involves significant resources and access to fine-tuning data. Prompting (Schick and Schütze, 2020) has been an alternative to fine-tuning in many NLP tasks. In our work, we explore the idea of using prompting for response generation in task-oriented dialog systems. Specifically, we propose an approach that performs contextual dynamic prompting where the prompts are learnt from dialog contexts. We aim to distill useful prompting signals from the dialog context. On experiments with MultiWOZ 2.2 dataset (Zang et al., 2020), we show that contextual dynamic prompts improve response generation in terms of combined score (Mehri et al., 2019) by 3 absolute points, and an additional 17 points when dialog states are incorporated. Furthermore, we carried out human annotation on these conversations and found that agents which incorporate context are preferred over agents with vanilla prefix-tuning.

**Models Teaching Models: Improving Model Accuracy with Slingshot Learning**
*Lachlan O'neill, Nandini Anantharama, Satya Borgohain and Simon Angus* 11:15-12:45 (Exhibit Hall)
One significant obstacle to the successful application of machine learning to real-world data is that of labeling: it is often prohibitively expensive to pay an ethical amount for the human labor required to label a dataset successfully. Human-in-the-loop techniques such as active

learning can reduce the cost, but the required human time is still significant and many fixed costs remain. Another option is to employ pre-trained transformer models as labelers at scale, which can yield reasonable accuracy and significant cost savings. However, such models can still be expensive to use due to their high computational requirements, and the opaque nature of these models is not always suitable in applied social science and public use contexts. We propose a novel semi-supervised method, named Slingshot Learning, in which we iteratively and selectively augment a small human-labeled dataset with labels from a high-quality "teacher" model to slingshot the performance of a "student" model in a cost-efficient manner. This reduces the accuracy trade-off required to use these simpler algorithms without disrupting their benefits, such as lower compute requirements, better interpretability, and faster inference. We define and discuss the slingshot learning algorithm and demonstrate its effectiveness on several benchmark tasks, using ALBERT to teach a simple Naive Bayes binary classifier. We experimentally demonstrate that Slingshot learning effectively decreases the performance gap between the teacher and student models. We also analyze its performance in several scenarios and compare different variants of the algorithm.

**Learning the Legibility of Visual Text Perturbations**
*Dev Seth, Rickard Stureborg, Danish Pruthi and Bhuwan Dhingra* 11:15-12:45 (Exhibit Hall)
Many adversarial attacks in NLP perturb text in puts to produce visually similar strings which are legible to humans but degrade model performance. Although preserving legibility is a necessary condition for text perturbation, little work has been done to systematically characterize it; instead, legibility is typically loosely enforced via intuitions around the nature and extent of perturbations. Particularly, it is unclear to what extent can inputs be perturbed while preserving legibility, or how to quantify the legibility of a perturbed string. In this work, we address this gap by learning models that predict the legibility of a perturbed string, and rank candidate perturbations based on their legibility. To do so, we collect and release LEGIT, a human-annotated dataset comprising the legibility of visually perturbed text. Using this dataset, we build both text- and vision-based models which achieve up to 0.91 F score in predicting whether an input is legible, and an accuracy of 0.86 in predicting which of two given perturbations is more legible. Additionally, we discover that legible perturbations from the LEGIT dataset are more effective at lowering the performance of NLP models than best-known attack strategies, suggesting that current models may be vulnerable to a broad range of perturbations beyond what is captured by existing visual attacks.

**DyLoRA: Parameter-Efficient Tuning of Pre-trained Models using Dynamic Search-Free Low-Rank Adaptation**
*Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev and Ali Ghodsi* 11:15-12:45 (Exhibit Hall)
With the ever-growing size of pretrained models (PMs), fine-tuning them has become more expensive and resource-hungry. As a remedy, low-rank adapters (LoRA) keep the main pretrained weights of the model frozen and just introduce some learnable truncated SVD modules (so-called LoRA blocks) to the model. While LoRA blocks are parameter-efficient, they suffer from two major problems: first, the size of these blocks is fixed and cannot be modified after training (for example, if we need to change the rank of LoRA blocks, then we need to re-train them from scratch); second, optimizing their rank requires an exhaustive search and effort. In this work, we introduce a dynamic low-rank adaptation (DyLoRA) technique to address these two problems together. Our DyLoRA method trains LoRA blocks for a range of ranks instead of a single rank by sorting the representation learned by the adapter module at different ranks during training. We evaluate our solution on different natural language understanding (GLUE benchmark) and language generation tasks (E2E, DART and WebNLG) using different pretrained models such as RoBERTa and GPT with different sizes. Our results show that we can train dynamic search-free models with DyLoRA at least 4 to 7 times (depending to the task) faster than LoRA without significantly compromising performance. Moreover, our models can perform consistently well on a much larger range of ranks compared to LoRA.

**A weakly supervised textual entailment approach to zero-shot text classification**
*Marc Pàmies, Joan Llop, Francesco Multari, Nicolau Duran-silva, César Parra-rojas, Aitor Gonzalez-agirre, Francesco Alessandro Massucci and Marta Villegas* 11:15-12:45 (Exhibit Hall)
Zero-shot text classification is a widely studied task that deals with a lack of annotated data. The most common approach is to reformulate it as a textual entailment problem, enabling classification into unseen classes. This work explores an effective approach that trains on a weakly supervised dataset generated from traditional classification data. We empirically study the relation between the performance of the entailment task, which is used as a proxy, and the target zero-shot text classification task. Our findings reveal that there is no linear correlation between both tasks, to the extent that it can be detrimental to lengthen the fine-tuning process even when the model is still learning, and propose a straightforward method to stop training on time. As a proof of concept, we introduce a domain-specific zero-shot text classifier that was trained on Microsoft Academic Graph data. The model, called SCIroShot, achieves state-of-the-art performance in the scientific domain and competitive results in other areas. Both the model and evaluation benchmark are publicly available on HuggingFace and GitHub.

**KGVL-BART: Knowledge Graph Augmented Visual Language BART for Radiology Report Generation**
*Kaveri Kale, Pushpak Bhattacharyya, Milind Gune, Aditya Shetty and Rustom Lawyer* 11:15-12:45 (Exhibit Hall)
Timely generation of radiology reports and diagnoses is a challenge worldwide due to the enormous number of cases and shortage of radiology specialists. In this paper, we propose a Knowledge Graph Augmented Vision Language BART (KGVL-BART) model that takes as input two chest X-ray images- one frontal and the other lateral- along with tags which are diagnostic keywords, and outputs a report with the patient-specific findings. Our system development effort is divided into 3 stages: i) construction of the Chest X-ray KG (referred to as chestX-KG), ii) image feature extraction, and iii) training a KGVL-BART model using the visual, text, and KG data. The dataset we use is the well-known Indiana University Chest X-ray reports with the train, validation, and test split of 3025 instances, 300 instances, and 500 instances respectively. We construct a Chest X-Ray knowledge graph from these reports by extracting entity1-relation-entity2 triples; the triples get extracted by a rule-based tool of our own. Constructed KG is verified by two experienced radiologists (with experience of 30 years and 8 years, respectively). We demonstrate that our model- KGVL-BART- outperforms State-of-the-Art transformer-based models on standard NLG scoring metrics. We also include a qualitative evaluation of our system by experienced radiologist (with experience of 30 years) on the test data, which showed that 73% of the reports generated were fully correct, only 5.5% are completely wrong and 21.5% have important missing details though overall correct. To the best of our knowledge, ours is the first system to make use of multi-modality and domain knowledge to generate X-ray reports automatically.

**How Far Can It Go? On Intrinsic Gender Bias Mitigation for Text Classification**
*Ewoenam Tokpo, Pieter Delobelle, Bettina Berendt and Toon Calders* 11:15-12:45 (Exhibit Hall)
To mitigate gender bias in contextualized language models, different intrinsic mitigation strategies have been proposed, alongside many bias metrics. Considering that the end use of these language models is for downstream tasks like text classification, it is important to understand how these intrinsic bias mitigation strategies actually translate to fairness in downstream tasks and the extent of this. In this work, we design a probe to investigate the effects that some of the major intrinsic gender bias mitigation strategies have on downstream text classification tasks. We discover that instead of resolving gender bias, intrinsic mitigation techniques and metrics are able to hide it in such a way that significant gender information is retained in the embeddings. Furthermore, we show that each mitigation technique is able to hide the bias from some of the intrinsic bias measures but not all, and each intrinsic bias measure can be fooled by some mitigation techniques, but not all. We confirm experimentally, that none of the intrinsic mitigation techniques used without any other fairness intervention is able to consistently impact extrinsic bias. We recommend that intrinsic bias mitigation techniques should be combined with other fairness interventions for downstream tasks.

**Concept-based Persona Expansion for Improving Diversity of Persona-Grounded Dialogue**
*Donghyun Kim, Youbin Ahn, Chanhee Lee, Wongyu Kim, Kyong-ho Lee, Donghoon Shin and Yeonsoo Lee* 11:15-12:45 (Exhibit Hall)
A persona-grounded dialogue model aims to improve the quality of responses to promote user engagement. However, because the given personas are mostly short and limited to only a few informative words, it is challenging to utilize them to generate diverse responses. To tackle this problem, we propose a novel persona expansion framework, Concept-based Persona eXpansion (CPX). CPX takes the original persona as input and generates expanded personas that contain conceptually rich content. We constitute CPX with two task modules: 1) Concept Extractor and 2) Sentence Generator. To train these modules, we exploit the duality of two tasks with a commonsense dataset consisting of a concept set and the corresponding sentences which contain the given concepts. Extensive experiments on persona expansion and response generation show that our work sufficiently contributes to improving the quality of responses in diversity and richness.

**Improving the Generalizability of Collaborative Dialogue Analysis With Multi-Feature Embeddings**
*Ayesha Enayet and Gita Sukthankar* 11:15-12:45 (Exhibit Hall)
Conflict prediction in communication is integral to the design of virtual agents that support successful teamwork by providing timely assistance. The aim of our research is to analyze discourse to predict collaboration success. Unfortunately, resource scarcity is a problem that teamwork researchers commonly face since it is hard to gather a large number of training examples. To alleviate this problem, this paper introduces a multi-feature embedding (MFeEmb) that improves the generalizability of conflict prediction models trained on dialogue sequences. MFeEmb leverages textual, structural, and semantic information from the dialogues by incorporating lexical, dialogue acts, and sentiment features. The use of dialogue acts and sentiment features reduces performance loss from natural distribution shifts caused mainly by changes in vocabulary. This paper demonstrates the performance of MFeEmb on domain adaptation problems in which the model is trained on discourse from one task domain and applied to predict team performance in a different domain. The generalizability of MFeEmb is quantified using the similarity measure proposed by Bontonou et al. (2021). Our results show that MFeEmb serves as an excellent domain-agnostic representation for meta-pretraining a few-shot model on collaborative multiparty dialogues.

**Representation biases in sentence transformers**
*Dmitry Nikolaev and Sebastian Padó* 11:15-12:45 (Exhibit Hall)
Variants of the BERT architecture specialised for producing full-sentence representations often achieve better performance on downstream tasks than sentence embeddings extracted from vanilla BERT. However, there is still little understanding of what properties of inputs determine the properties of such representations. In this study, we construct several sets of sentences with pre-defined lexical and syntactic structures and show that SOTA sentence transformers have a strong nominal-participant-set bias: cosine similarities between pairs of sentences are more strongly determined by the overlap in the set of their noun participants than by having the same predicates, lengthy nominal modifiers, or adjuncts. At the same time, the precise syntactic-thematic functions of the participants are largely irrelevant.

**Counter-GAP: Counterfactual Bias Evaluation through Gendered Ambiguous Pronouns**
*Zhongbin Xie, Vid Kocijan, Thomas Lukasiewicz and Oana-maria Camburu* 11:15-12:45 (Exhibit Hall)
Bias-measuring datasets play a critical role in detecting biased behavior of language models and in evaluating progress of bias mitigation methods. In this work, we focus on evaluating gender bias through coreference resolution, where previous datasets are either hand-crafted or fail to reliably measure an explicitly defined bias. To overcome these shortcomings, we propose a novel method to collect diverse, natural, and minimally distant text pairs via counterfactual generation, and construct Counter-GAP, an annotated dataset consisting of 4008 instances grouped into 1002 quadruples. We further identify a bias cancellation problem in previous group-level metrics on Counter-GAP, and propose to use the difference between inconsistency across genders and within genders to measure bias at a quadruple level. Our results show that four pre-trained language models are significantly more inconsistent across different gender groups than within each group, and that a name-based counterfactual data augmentation method is more effective to mitigate such bias than an anonymization-based method.

**The NLP Task Effectiveness of Long-Range Transformers**
*Guanghui Qin, Yukun Feng and Benjamin Van Durme* 11:15-12:45 (Exhibit Hall)
Transformer models cannot easily scale to long sequences due to their O(N^2) time and space complexity. This has led to Transformer variants seeking to lower computational complexity, such as Longformer and Performer. While such models have theoretically greater efficiency, their effectiveness on real NLP tasks has not been well studied. We benchmark 7 variants of Transformer models on 5 difficult NLP tasks and 7 datasets. We design experiments to isolate the effect of pretraining and hyperparameter settings, to focus on their capacity for long-range attention. Moreover, we present various methods to investigate attention behaviors to illuminate model details beyond metric scores. We find that the modified attention in long-range transformers has advantages on content selection and query-guided decoding, but they come with previously unrecognized drawbacks such as insufficient attention to distant tokens and accumulated approximation error.

**Semi-supervised New Event Type Induction and Description via Contrastive Loss-Enforced Batch Attention**
*Carl Edwards and Heng Ji* 11:15-12:45 (Exhibit Hall)
Most event extraction methods have traditionally relied on an annotated set of event types. However, creating event ontologies and annotating supervised training data are expensive and time-consuming. Previous work has proposed semi-supervised approaches which leverage seen (annotated) types to learn how to automatically discover new event types. State-of-the-art methods, both semi-supervised or fully unsupervised, use a form of reconstruction loss on specific tokens in a context. In contrast, we present a novel approach to semi-supervised new event type induction using a masked contrastive loss, which learns similarities between event mentions by enforcing an attention mechanism over the data minibatch. We further disentangle the discovered clusters by approximating the underlying manifolds in the data, which allows us to achieve an adjusted rand index score of 48.85%. Building on these clustering results, we extend our approach to two new tasks: predicting the type name of the discovered clusters and linking them to FrameNet frames.

**DiscoScore: Evaluating Text Generation with BERT and Discourse Coherence**
*Wei Zhao, Michael Strube and Steffen Eger* 11:15-12:45 (Exhibit Hall)
Recently, there has been a growing interest in designing text generation systems from a discourse coherence perspective, e.g., modeling the interdependence between sentences. Still, recent BERT-based evaluation metrics are weak in recognizing coherence, and thus are not reliable in a way to spot the discourse-level improvements of those text generation systems. In this work, we introduce DiscoScore, a parametrized discourse metric, which uses BERT to model discourse coherence from different perspectives, driven by Centering theory. Our experiments encompass 16 non-discourse and discourse metrics, including DiscoScore and popular coherence models, evaluated on summarization and document-level machine translation (MT). We find that (i) the majority of BERT-based metrics correlate much worse with human rated coherence than early discourse metrics, invented a decade ago; (ii) the recent state-of-the-art BARTScore is weak when operated at system level—which is particularly problematic as systems are typically compared in this manner. DiscoScore, in contrast, achieves strong system-level correlation with human ratings, not only in coherence but also in factual consistency and other aspects, and surpasses BARTScore by over 10 correlation points on average. Further, aiming to understand DiscoScore, we provide justifications to the importance of discourse coherence for evaluation metrics, and explain the superiority of one variant over another. Our code is available at \url{https://github.com/AIPHES/DiscoScore}.

**DiTTO: A Feature Representation Imitation Approach for Improving Cross-Lingual Transfer**
*Shanu Kumar, Soujanya Abbaraju, Sandipan Dandapat, Sunayana Sitaram and Monojit Choudhury* 11:15-12:45 (Exhibit Hall)
Zero-shot cross-lingual transfer is promising, however has been shown to be sub-optimal, with inferior transfer performance across low-resource languages. In this work, we envision languages as domains for improving zero-shot transfer by jointly reducing the feature incongruity between the source and the target language and increasing the generalization capabilities of pre-trained multilingual transformers. We show that our approach, DiTTO, significantly outperforms the standard zero-shot fine-tuning method on multiple datasets across all languages using solely unlabeled instances in the target language. Empirical results show that jointly reducing feature incongruity for multiple target languages is vital for successful cross-lingual transfer. Moreover, our model enables better cross-lingual transfer than standard fine-tuning methods, even in the few-shot setting.

**Efficient Encoders for Streaming Sequence Tagging**
*Ayush Kaushal, Aditya Gupta, Shyam Upadhyay and Manaal Faruqui* 11:15-12:45 (Exhibit Hall)
A naive application of state-of-the-art bidirectional encoders for streaming sequence tagging would require encoding each token from scratch for each new token in an incremental streaming input (like transcribed speech). The lack of re-usability of previous computation leads to a higher number of Floating Point Operations (or FLOPs) and higher number of unnecessary label flips. Increased FLOPs consequently lead to higher wall-clock time and increased label flipping leads to poorer streaming performance. In this work, we present a Hybrid Encoder with Adaptive Restart (HEAR) that addresses these issues while maintaining the performance of bidirectional encoders over the offline (or complete) and improving streaming (or incomplete) inputs. HEAR has a Hybrid unidirectional-bidirectional encoder architecture to perform sequence tagging, along with an Adaptive Restart Module (ARM) to selectively guide the restart of bidirectional portion of the encoder. Across four sequence tagging tasks, HEAR offers FLOP savings in streaming settings upto 71.1% and also outperforms bidirectional encoders for streaming predictions by upto +10% streaming exact match.

**Patient Outcome and Zero-shot Diagnosis Prediction with Hypernetwork-guided Multitask Learning**
*Shaoxiong Ji and Pekka Marttinen* 11:15-12:45 (Exhibit Hall)
Multitask deep learning has been applied to patient outcome prediction from text, taking clinical notes as input and training deep neural networks with a joint loss function of multiple tasks. However, the joint training scheme of multitask learning suffers from inter-task interference, and diagnosis prediction among the multiple tasks has the generalizability issue due to rare diseases or unseen diagnoses. To solve these challenges, we propose a hypernetwork-based approach that generates task-conditioned parameters and coefficients of multitask prediction heads to learn task-specific prediction and balance the multitask learning. We also incorporate semantic task information to improve the generalizability of our task-conditioned multitask model. Experiments on early and discharge notes extracted from the real-world MIMIC database show our method can achieve better performance on multitask patient outcome prediction than strong baselines in most cases. Besides, our method can effectively handle the scenario with limited information and improve zero-shot prediction on unseen diagnosis categories.

**Shironaam: Bengali News Headline Generation using Auxiliary Information**
*Abu Ubaida Akash, Mir Tafseer Nayeem, Faisal Tareque Shohan and Tanvir Islam* 11:15-12:45 (Exhibit Hall)
Automatic headline generation systems have the potential to assist editors in finding interesting headlines to attract visitors or readers. However, the performance of headline generation systems remains challenging due to the unavailability of sufficient parallel data for low-resource languages like Bengali and the lack of ideal approaches to develop a system for headline generation using pre-trained language models, especially for long news articles. To address these challenges, we present Shironaam, a large-scale dataset in Bengali containing over 240K news article-headline pairings with auxiliary data such as image captions, topic words, and category information. Unlike other headline generation models, this paper uses this auxiliary information to better model this task. Furthermore, we utilize the contextualized language models to design encoder-decoder model for Bengali news headline generation and follow a simple yet cost-effective coarse-to-fine approach using topic-words to retrieve important sentences considering the fixed length requirement of the pre-trained language models. Finally, we conduct extensive experiments on our dataset containing news articles of 13 different categories to demonstrate the effectiveness of incorporating auxiliary information and evaluate our system on a wide range of metrics. The experimental results demonstrate that our methods bring significant improvements (i.e., 3 to 10 percentage points across all evaluation metrics) over the baselines. Also to illustrate the utility and robustness, we report experimental results in few-shot and non-few-shot settings.

**ViHOS: Hate Speech Spans Detection for Vietnamese**
*Phu Gia Hoang, Canh Luu, Khanh Tran, Kiet Nguyen and Ngan Nguyen* 11:15-12:45 (Exhibit Hall)
The rise in hateful and offensive language directed at other users is one of the adverse side effects of the increased use of social networking platforms. This could make it difficult for human moderators to review tagged comments filtered by classification systems. To help address this issue, we present the ViHOS (Vietnamese Hate and Offensive Spans) dataset, the first human-annotated corpus containing 26k spans on 11k comments. We also provide definitions of hateful and offensive spans in Vietnamese comments as well as detailed annotation guidelines. Besides, we conduct experiments with various state-of-the-art models. Specifically, XLM-R-Large achieved the best F1-scores in Single span detection and All spans detection, while PhoBERT-Large obtained the highest in Multiple spans detection. Finally, our error analysis demonstrates the difficulties in detecting specific types of spans in our data for future research. Our dataset is released on GitHub.

**Vote'n'Rank: Revision of Benchmarking with Social Choice Theory**
*Mark Rofin, Vladislav Mikhailov, Mikhail Florinsky, Andrey Kravchenko, Tatiana Shavrina, Elena Tutubalina, Daniel Karabekyan and Ekaterina Artemova* 11:15-12:45 (Exhibit Hall)
The development of state-of-the-art systems in different applied areas of machine learning (ML) is driven by benchmarks, which have shaped the paradigm of evaluating generalisation capabilities from multiple perspectives. Although the paradigm is shifting towards more fine-grained evaluation across diverse tasks, the delicate question of how to aggregate the performances has received particular interest in the community. In general, benchmarks follow the unspoken utilitarian principles, where the systems are ranked based on their mean average score over task-specific metrics. Such aggregation procedure has been viewed as a sub-optimal evaluation protocol, which may have created the illusion of progress. This paper proposes Vote'n'Rank, a framework for ranking systems in multi-task benchmarks under the principles of the social choice theory. We demonstrate that our approach can be efficiently utilised to draw new insights on benchmarking in several ML sub-fields and identify the best-performing systems in research and development case studies. The Vote'n'Rank's procedures are more robust than the mean average while being able to handle missing performance scores and determine conditions under which the system becomes the winner.

**Parameter-Efficient Tuning with Special Token Adaptation**
*Xiaocong Yang, James Y. Huang, Wenxuan Zhou and Muhao Chen* 11:15-12:45 (Exhibit Hall)
Parameter-efficient tuning aims at updating only a small subset of parameters when adapting a pretrained model to downstream tasks. In this work, we introduce PASTA, in which we only modify the special token representations (e.g., [SEP] and [CLS] in BERT) before the self-attention module at each layer in Transformer-based models. PASTA achieves comparable performance to fine-tuning in natural language understanding tasks including text classification and NER with up to only 0.029% of total parameters trained. Our work not only provides a simple yet effective way of parameter-efficient tuning, which has a wide range of practical applications when deploying finetuned models for multiple tasks, but also demonstrates the pivotal role of special tokens in pretrained language models.

**Conclusion-based Counter-Argument Generation**
*Milad Alshomary and Henning Wachsmuth* 11:15-12:45 (Exhibit Hall)
In real-world debates, the most common way to counter an argument is to reason against its main point, that is, its conclusion. Existing work on the automatic generation of natural language counter-arguments does not address the relation to the conclusion, possibly because many arguments leave their conclusion implicit. In this paper, we hypothesize that the key to effective counter-argument generation is to explicitly model the argument's conclusion and to ensure that the stance of the generated counter is opposite to that conclusion. In particular, we propose a multitask approach that jointly learns to generate both the conclusion and the counter of an input argument. The approach employs a stance-based ranking component that selects the counter from a diverse set of generated candidates whose stance best opposes the generated conclusion. In both automatic and manual evaluation, we provide evidence that our approach generates more relevant and stance-adhering counters than strong baselines.

**Annotation Error Detection: Analyzing the Past and Present for a More Coherent Future**
*Jan-Christoph Klie, Bonnie Webber and Iryna Gurevych* 11:15-12:45 (Exhibit Hall)
Annotated data is an essential ingredient in natural language processing for training and evaluating machine learning models. It is therefore very desirable for the annotations to be of high quality. Recent work, however, has shown that several popular datasets contain a surprising amount of annotation errors or inconsistencies. To alleviate this issue, many methods for annotation error detection have been devised over the years. While researchers show that their approaches work well on their newly introduced datasets, they rarely compare their methods to previous work or on the same datasets. This raises strong concerns on methods' general performance and makes it difficult to asses their strengths and weaknesses. We therefore reimplement 18 methods for detecting potential annotation errors and evaluate them on 9 English datasets for text classification as well as token and span labeling. In addition, we define a uniform evaluation setup including a new formalization of the annotation error detection task, evaluation protocol and general best practices. To facilitate future research and reproducibility, we release our datasets and implementations in an easy-to-use and open source software package.

**Discontinuous Combinatory Constituency Parsing**
*Zhousi Chen* 11:15-12:45 (Exhibit Hall)
We extend a pair of continuous combinator-based constituency parsers (one binary and one multi-branching) into a discontinuous pair. Our parsers iteratively compose constituent vectors from word embeddings without any grammar constraints. Their empirical complexities are subquadratic. Our extension includes 1) a swap action for the orientation-based binary model and 2) biaffine attention for the chunker-based multi-branching model. In tests conducted with the Discontinuous Penn Treebank and TIGER Treebank, we achieved state-of-the-art discontinuous accuracy with a significant speed advantage.

**What Clued the AI Doctor In? On the Influence of Data Source and Quality for Transformer-Based Medical Self-Disclosure Detection**
*Mina Valizadeh, Xing Qian, Pardis Ranjbar-noiey, Cornelia Caragea and Natalie Parde* 11:15-12:45 (Exhibit Hall)
Recognizing medical self-disclosure is important in many healthcare contexts, but it has been under-explored by the NLP community. We conduct a three-pronged investigation of this task. We (1) manually expand and refine the only existing medical self-disclosure corpus, resulting in a new, publicly available dataset of 3,919 social media posts with clinically validated labels and high compatibility with the existing task-specific protocol. We also (2) study the merits of pretraining task domain and text style by comparing Transformer-based models for this task, pretrained from general, medical, and social media sources. Our BERTweet condition outperforms the existing state of the art for this task by a relative F1 score increase of 16.73%. Finally, we (3) compare data augmentation techniques for this task, to assess the extent to which medical self-disclosure data may be further synthetically expanded. We discover that this task poses many challenges for data augmentation techniques, and we provide an in-depth analysis of identified trends.

**Exploring Paracrawl for Document-level Neural Machine Translation**
*Yusser Al Ghussin, Jingyi Zhang and Josef Van Genabith* 11:15-12:45 (Exhibit Hall)
Document-level neural machine translation (NMT) has outperformed sentence-level NMT on a number of datasets. However, document-level NMT is still not widely adopted in realworld translation systems mainly due to the lack of large-scale general-domain training data for document-level NMT. We examine the effectiveness of using Paracrawl for learning document-level translation. Paracrawl is a large-scale parallel corpus crawled from the Internet and contains data from various domains. The official Paracrawl corpus was released as parallel sentences (extracted from parallel webpages) and therefore previous works only used Paracrawl for learning sentence-level translation. In this work, we extract parallel paragraphs from Paracrawl parallel webpages using automatic sentence alignments and we use the extracted parallel paragraphs as parallel documents for training document-level translation models. We show that document-level NMT models trained with only parallel paragraphs from Paracrawl can be used to translate real documents from TED, News and Europarl, outperforming sentence-level NMT models. We also perform a targeted pronoun evaluation and show that document-level models trained with Paracrawl data can help context-aware pronoun translation.

**Shorten the Long Tail for Rare Entity and Event Extraction**
*Pengfei Yu and Heng Ji* 11:15-12:45 (Exhibit Hall)
The distribution of knowledge elements such as entity types and event types is long-tailed in natural language. Hence information extraction datasets naturally conform long-tailed distribution. Although imbalanced datasets can teach the model about the useful real-world bias, deep learning models may learn features not generalizable to rare or unseen expressions of entities or events during evaluation, especially for rare types without sufficient training instances. Existing approaches for the long-tailed learning problem seek to manipulate the training data by re-balancing, augmentation or introducing extra prior knowledge. In comparison, we propose to handle the generalization challenge by making the evaluation instances closer to the frequent training cases. We design a new transformation module that transforms infrequent candidate mention representation during evaluation with the average mention representation in the training dataset. Experimental results on classic benchmarks on three entity or event extraction datasets demonstrates the effectiveness of our framework.

**Metaphor Detection with Effective Context Denoising**
*Shun Wang, Yucheng Li, Chenghua Lin, Loic Barrault and Frank Guerin* 11:15-12:45 (Exhibit Hall)
We propose a novel RoBERTa-based model, RoPPT, which introduces a target-oriented parse tree structure in metaphor detection. Compared to existing models, RoPPT focuses on semantically relevant information and achieves the state-of-the-art on several main metaphor datasets. We also compare our approach against several popular denoising and pruning methods, demonstrating the effectiveness of our approach in context denoising. Our code and dataset can be found at https://github.com/MajiBear000/RoPPT.

**Selective In-Context Data Augmentation for Intent Detection using Pointwise V-Information**
*Yen Ting Lin, Alexandros Papangelis, Seokhwan Kim, Sungjin Lee, Devamanyu Hazarika, Mahdi Namazifar, Di Jin, Yang Liu and Dilek Hakkani-tur* 11:15-12:45 (Exhibit Hall)
This work focuses on in-context data augmentation for intent detection. Having found that augmentation via in-context prompting of large

pre-trained language models (PLMs) alone does not improve performance, we introduce a novel approach based on PLMs and pointwise V-information (PVI), a metric that can measure the usefulness of a datapoint for training a model. Our method first fine-tunes a PLM on a small seed of training data and then synthesizes new datapoints - utterances that correspond to given intents. It then employs intent-aware filtering, based on PVI, to remove datapoints that are not helpful to the downstream intent classifier. Our method is thus able to leverage the expressive power of large language models to produce diverse training data. Empirical results demonstrate that our method can produce synthetic training data that achieve state-of-the-art performance on three challenging intent detection datasets under few-shot settings (1.28% absolute improvement in 5-shot and 1.18% absolute in 10-shot, on average) and perform on par with the state-of-the-art in full-shot settings (within 0.01% absolute, on average).

**A Systematic Search for Compound Semantics in Pretrained BERT Architectures**
*Filip Miletic and Sabine Schulte Im Walde* 11:15-12:45 (Exhibit Hall)
To date, transformer-based models such as BERT have been less successful in predicting compositionality of noun compounds than static word embeddings. This is likely related to a suboptimal use of the encoded information, reflecting an incomplete grasp of how the models represent the meanings of complex linguistic structures. This paper investigates variants of semantic knowledge derived from pretrained BERT when predicting the degrees of compositionality for 280 English noun compounds associated with human compositionality ratings. Our performance strongly improves on earlier unsupervised implementations of pretrained BERT and highlights beneficial decisions in data preprocessing, embedding computation, and compositionality estimation. The distinct linguistic roles of heads and modifiers are reflected by differences in BERT-derived representations, with empirical properties such as frequency, productivity, and ambiguity affecting model performance. The most relevant representational information is concentrated in the initial layers of the model architecture.

**Augmenting Pre-trained Language Models with QA-Memory for Open-Domain Question Answering**
*Wenhu Chen, Pat Verga, Michiel De Jong, John Wieting and William Cohen* 11:15-12:45 (Exhibit Hall)
Existing state-of-the-art methods for open-domain question-answering (ODQA) use an open book approach in which information is first retrieved from a large text corpus or knowledge base (KB) and then reasoned over to produce an answer. A recent alternative is to retrieve from a collection of previously-generated question-answer pairs; this has several practical advantages including being more memory and compute-efficient. Question-answer pairs are also appealing in that they can be viewed as an intermediate between text and KB triples: like KB triples, they often concisely express a single relationship, but like text, have much higher coverage than traditional KBs. In this work, we describe a new QA system that augments a text-to-text model with a large memory of question-answer pairs, and a new pre-training task for the latent step of question retrieval. The pre-training task substantially simplifies training and greatly improves performance on smaller QA benchmarks. Unlike prior systems of this sort, our QA system can also answer multi-hop questions that do not explicitly appear in the collection of stored question-answer pairs.

**Gold Doesn't Always Glitter: Spectral Removal of Linear and Nonlinear Guarded Attribute Information**
*Shun Shao, Yftah Ziser and Shay B. Cohen* 11:15-12:45 (Exhibit Hall)
We describe a simple and effective method (Spectral Attribute removaL; SAL) to remove private or guarded information from neural representations. Our method uses matrix decomposition to project the input representations into directions with reduced covariance with the guarded information rather than maximal covariance as factorization methods normally use. We begin with linear information removal and proceed to generalize our algorithm to the case of nonlinear information removal using kernels. Our experiments demonstrate that our algorithm retains better main task performance after removing the guarded information compared to previous work. In addition, our experiments demonstrate that we need a relatively small amount of guarded attribute data to remove information about these attributes, which lowers the exposure to sensitive data and is more suitable for low-resource scenarios.

**Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions**
*Mihir Parmar, Swaroop Mishra, Mor Geva and Chitta Baral* 11:15-12:45 (Exhibit Hall)
In recent years, progress in NLU has been driven by benchmarks. These benchmarks are typically collected by crowdsourcing, where annotators write examples based on annotation instructions crafted by dataset creators. In this work, we hypothesize that annotators pick up on patterns in the crowdsourcing instructions, which bias them to write many similar examples that are then over-represented in the collected data. We study this form of bias, termed instruction bias, in 14 recent NLU benchmarks, showing that instruction examples often exhibit concrete patterns, which are propagated by crowdworkers to the collected data. This extends previous work (Geva et al., 2019) and raises a new concern of whether we are modeling the dataset creator's instructions, rather than the task. Through a series of experiments, we show that, indeed, instruction bias can lead to overestimation of model performance, and that models struggle to generalize beyond biases originating in the crowdsourcing instructions. We further analyze the influence of instruction bias in terms of pattern frequency and model size, and derive concrete recommendations for creating future NLU benchmarks.

**ZELDA: A Comprehensive Benchmark for Supervised Entity Disambiguation**
*Marcel Milich and Alan Akbik* 11:15-12:45 (Exhibit Hall)
Entity disambiguation (ED) is the task of disambiguating named entity mentions in text to unique entries in a knowledge base. Due to its industrial relevance, as well as current progress in leveraging pre-trained language models, a multitude of ED approaches have been proposed in recent years. However, we observe a severe lack of uniformity across experimental setups in current ED work,rendering a direct comparison of approaches based solely on reported numbers impossible: Current approaches widely differ in the data set used to train, the size of the covered entity vocabulary, and the usage of additional signals such as candidate lists. To address this issue, we present ZELDA , a novel entity disambiguation benchmark that includes a unified training data set, entity vocabulary, candidate lists, as well as challenging evaluation splits covering 8 different domains. We illustrate its design and construction, and present experiments in which we train and compare current state-of-the-art approaches on our benchmark. To encourage greater direct comparability in the entity disambiguation domain, we make our benchmark publicly available to the research community.

**GLADIS: A General and Large Acronym Disambiguation Benchmark**
*Lihu Chen, Gael Varoquaux and Fabian Suchanek* 11:15-12:45 (Exhibit Hall)
Acronym Disambiguation (AD) is crucial for natural language understanding on various sources, including biomedical reports, scientific papers, and search engine queries. However, existing acronym disambiguation benchmarks and tools are limited to specific domains, and the size of prior benchmarks is rather small. To accelerate the research on acronym disambiguation, we construct a new benchmark with three components: (1) a much larger acronym dictionary with 1.5M acronyms and 6.4M long forms; (2) a pre-training corpus with 160 million sentences; (3) three datasets that cover the general, scientific, and biomedical domains. We then pre-train a language model, \emph{AcroBERT}, on our constructed corpus for general acronym disambiguation, and show the challenges and values of our new benchmark.

**Do Neural Topic Models Really Need Dropout? Analysis of the Effect of Dropout in Topic Modeling**
*Suman Adhya, Avishek Lahiri and Debarshi Kumar Sanyal* 11:15-12:45 (Exhibit Hall)
Dropout is a widely used regularization trick to resolve the overfitting issue in large feedforward neural networks trained on a small dataset, which performs poorly on the held-out test subset. Although the effectiveness of this regularization trick has been extensively studied for

convolutional neural networks, there is a lack of analysis of it for unsupervised models and in particular, VAE-based neural topic models. In this paper, we have analyzed the consequences of dropout in the encoder as well as in the decoder of the VAE architecture in three widely used neural topic models, namely, contextualized topic model (CTM), ProdLDA, and embedded topic model (ETM) using four publicly available datasets. We characterize the dropout effect on these models in terms of the quality and predictive performance of the generated topics.

**Efficient CTC Regularization via Coarse Labels for End-to-End Speech Translation**
*Biao Zhang, Barry Haddow and Rico Sennrich* 11:15-12:45 (Exhibit Hall)
For end-to-end speech translation, regularizing the encoder with the Connectionist Temporal Classification (CTC) objective using the source transcript or target translation as labels can greatly improve quality. However, CTC demands an extra prediction layer over the vocabulary space, bringing in non-negligible model parameters and computational overheads, although this layer becomes useless at inference. In this paper, we re-examine the need for genuine vocabulary labels for CTC for regularization and explore strategies to reduce the CTC label space, targeting improved efficiency without quality degradation. We propose coarse labeling for CTC (CoLaCTC), which merges vocabulary labels via simple heuristic rules, such as using truncation, division or modulo (MOD) operations. Despite its simplicity, our experiments on 4 source and 8 target languages show that CoLaCTC with MOD particularly can compress the label space aggressively to 256 and even further, gaining training efficiency yet still delivering comparable or better performance than the CTC baseline. We also show that CoLaCTC successfully generalizes to CTC regularization regardless of using transcript or translation for labeling.

**Instruction Clarification Requests in Multimodal Collaborative Dialogue Games: Tasks, and an Analysis of the CoDraw Dataset**
*Brielen Madureira and David Schlangen* 11:15-12:45 (Exhibit Hall)
In visual instruction-following dialogue games, players can engage in repair mechanisms in face of an ambiguous or underspecified instruction that cannot be fully mapped to actions in the world. In this work, we annotate Instruction Clarification Requests (iCRs) in CoDraw, an existing dataset of interactions in a multimodal collaborative dialogue game. We show that it contains lexically and semantically diverse iCRs being produced self-motivatedly by players deciding to clarify in order to solve the task successfully. With 8.8k iCRs found in 9.9k dialogues, CoDraw-iCR (v1) is a large spontaneous iCR corpus, making it a valuable resource for data-driven research on clarification in dialogue. We then formalise and provide baseline models for two tasks: Determining when to make an iCR and how to recognise them, in order to investigate to what extent these tasks are learnable from data.

**Opportunities and Challenges in Neural Dialog Tutoring**
*Jakub Macina, Nico Daheim, Lingzhi Wang, Tanmay Sinha, Manu Kapur, Iryna Gurevych and Mrinmaya Sachan* 11:15-12:45 (Exhibit Hall)
Designing dialog tutors has been challenging as it involves modeling the diverse and complex pedagogical strategies employed by human tutors. Although there have been significant recent advances in neural conversational systems using large language models and growth in available dialog corpora, dialog tutoring has largely remained unaffected by these advances. In this paper, we rigorously analyze various generative language models on two dialog tutoring datasets for language learning using automatic and human evaluations to understand the new opportunities brought by these advances as well as the challenges we must overcome to build models that would be usable in real educational settings. We find that although current approaches can model tutoring in constrained learning scenarios when the number of concepts to be taught and possible teacher strategies are small, they perform poorly in less constrained scenarios. Our human quality evaluation shows that both models and ground-truth annotations exhibit low performance in terms of equitable tutoring, which measures learning opportunities for students and how engaging the dialog is. To understand the behavior of our models in a real tutoring setting, we conduct a user study using expert annotators and find a significantly large number of model reasoning errors in 45% of conversations. Finally, we connect our findings to outline future work.

**Modeling Complex Event Scenarios via Simple Entity-focused Questions**
*Mahnaz Koupaee, Greg Durrett, Nathanael Chambers and Niranjan Balasubramanian* 11:15-12:45 (Exhibit Hall)
Event scenarios are often complex and involve multiple event sequences connected through different entity participants. Exploring such complex scenarios requires an ability to branch through different sequences, something that is difficult to achieve with standard event language modeling. To address this, we propose a question-guided generation framework that models events in complex scenarios as answers to questions about participants. At any step in the generation process, the framework uses the previously-generated events as context, but generates the next event as an answer to one of three questions: what else a participant did, what else happened to a participant, or what else happened. The participants and the questions themselves can be sampled or be provided as input from a user, allowing for controllable exploration. Our empirical evaluation shows that this question-guided generation provides better coverage of participants, diverse events within a domain, comparable perplexities for modeling event sequences, and more effective control for interactive schema generation.

**MAPL: Parameter-Efficient Adaptation of Unimodal Pre-Trained Models for Vision-Language Few-Shot Prompting**
*Oscar Mañas, Pau Rodriguez Lopez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal and Aishwarya Agrawal* 11:15-12:45 (Exhibit Hall)
Large pre-trained models have proved to be remarkable zero- and (prompt-based) few-shot learners in unimodal vision and language tasks. We propose MAPL, a simple and parameter-efficient method that reuses frozen pre-trained unimodal models and leverages their strong generalization capabilities in multimodal vision-language (VL) settings. MAPL learns a lightweight mapping between the representation spaces of unimodal models using aligned image-text data, and can generalize to unseen VL tasks from just a few in-context examples. The small number of trainable parameters makes MAPL effective at low-data and in-domain learning. Moreover, MAPL's modularity enables easy extension to other pre-trained models. Extensive experiments on several visual question answering and image captioning benchmarks show that MAPL achieves superior or competitive performance compared to similar methods while training orders of magnitude fewer parameters. MAPL can be trained in just a few hours using modest computational resources and public datasets. We release our code and pre-trained model weights at https://github.com/oscmansan/mapl.

**How Many and Which Training Points Would Need to be Removed to Flip this Prediction?**
*Jinghan Yang, Sarthak Jain and Byron Wallace* 11:15-12:45 (Exhibit Hall)

# 9 Workshops

## Overview

During the days of the workshops, **Registration** will be held from 08:00.

### Friday, May 5, 2023

| | | |
|---|---|---|
| Elafiti 4 | W1 - Fourth Workshop on Insights from Negative Results in NLP | p.96 |
| Divona 1 | W2 - Tenth Workshop on NLP for Similar Languages, Varieties and Dialects | p.97 |
| Elafiti 3 | W3 - Cross-Cultural Considerations in NLP | p.98 |
| Bokar | W4 - The Second Ukrainian Natural Language Processing Workshop | p.99 |
| Divona 2 | W5 - Sixth Workshop on Fact Extraction and VERification | p.100 |

### Saturday, May 6, 2023

| | | |
|---|---|---|
| Asimon | W6 - The 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature | p.102 |
| Elafiti 4 | W7 - Workshop Proposal: The Sixth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2023) | p.103 |
| Elafiti 3 | W8 - The 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP) | p.105 |
| Divona 1 | W9 - SlavNLP-2023: The 9th Biennial Workshop on Slavic NLP | p.107 |
| Divona 2 | W10 - The fourth workshop on Resources for African Indigenous Languages (RAIL) | p.108 |
| Bokar | W11 - 19th Workshop on Multiword Expressions | p.110 |
| Asimon | W12 - The Second Workshop on NLP Applications to Field Linguistics | p.112 |

# W1 - Fourth Workshop on Insights from Negative Results in NLP

**Organizers:**
Shabnam Tafreshi, Arjun Reddy Akula, João Sedoc, Anna Rogers, Aleksandr Drozd, Anna Rumshisky

https://insights-workshop.github.io/
Venue: Elafiti 4
**Friday, May 5, 2023**

Insights from Negative Results in NLP workshop invites both practical and theoretical unexpected or negative results that have important implications for future research, highlight methodological issues with existing approaches, and/or point out pervasive misunderstandings or bad practices. In particular, the most successful NLP models currently rely on different kinds of pretrained meaning representations (from word embeddings to Transformer-based models like BERT and GPT-3). To complement all the success stories, it would be insightful to see where and possibly why they fail.

| | |
|---|---|
| 09:00 - 09:15 | ***Opening Remarks*** |
| 09:15 - 10:00 | ***Thematic Session 1: Text Generation*** |
| 10:00 - 10:45 | ***Invited Talk: Vered Shwartz*** |
| 10:45 - 11:15 | ***Coffee Break*** |
| 11:15 - 11:45 | ***Thematic Session 2: Text Classification & Comprehension*** |
| 11:45 - 12:15 | ***Thematic Session 3: Representation Learning & Pre-training*** |
| 12:15 - 14:00 | ***Lunch*** |
| 14:00 - 14:30 | ***Invited Talk: Mohit Iyyer*** |
| 14:30 - 15:00 | ***Thematic Session 4: Robustness & Error Analysis*** |
| 15:00 - 15:30 | ***Invited Talk: Rachel Rudinger*** |
| 15:30 - 16:00 | ***Coffee Break*** |
| 16:00 - 16:30 | ***Invited Talk: Hila Gonen*** |
| 16:30 - 18:00 | ***Poster Session*** |
| 18:00 - 18:10 | ***Closing Remarks*** |

# W2 - Tenth Workshop on NLP for Similar Languages, Varieties and Dialects

**Organizers:**
Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, Marcos Zampieri

https://sites.google.com/view/vardial-2023
Venue: Divona 1
**Friday, May 5, 2023**

VarDial is a well-established series of workshops promoting a forum for scholars working on various topics related to the study of diatopic language variation from a computational perspective. The workshop deals with computational methods and language resources for closely related languages, language varieties, and dialects. VarDial also welcomes papers dealing with diachronic language variation (e.g., phylogenetic methods and historical dialects). This edition marks VarDial's ten-year anniversary. We are pleased to see that the workshop continues to serve the community as the main venue for researchers interested in the computational processing of closely related languages, language varieties and dialects. The papers accepted this year address a wide range of topics, such as corpus building, part-of-speech tagging, and machine translation. This volume once again showcases the great linguistic diversity that VarDial embodies, including work on dialects and varieties of many different languages, such as Arabic, Cantonese, Croatian, Finnish, German, Irish, Italian, Mandarin, Occitan, Serbian, and Spanish. The VarDial evaluation campaign continues to be an essential part of the workshop. This year, three shared tasks were organized: Slot and intent detection for low-resource language varieties (SID4LR), Discriminating Between Similar Languages – True Labels (DSL-TL), and Discriminating Between Similar Languages – Speech (DSL-S).

| | |
|---|---|
| 09:00 - 09:10 | ***Opening remarks*** |
| 09:10 - 10:30 | ***Oral presentations*** |
| 10:30 - 11:00 | ***Coffee break*** |
| 11:00 - 12:15 | ***Oral presentations*** |
| 12:15 - 12:40 | ***Poster Boosters I*** |
| 12:40 - 14:00 | ***Lunch break*** |
| 14:00 - 14:50 | ***Keynote Talk by Ivan Vulić: Bridging the Dialect Gap with Modular Transfer Learning?*** |
| 14:50 - 15:40 | ***Round Table: VarDial in the Era of Large Language Models*** |
| 15:40 - 16:15 | ***Coffee break*** |
| 16:15 - 16:40 | ***Poster Boosters II*** |
| 16:40 - 18:00 | ***Poster Session*** |

# W3 - Cross-Cultural Considerations in NLP

**Organizers:**
Vinodkumar Prabhakaran, Sunipa Dev, Dirk Hovy, Luciana Benotti, David Adelani

https://sites.google.com/view/c3nlp/home
Venue: Elafiti 3
**Friday, May 5, 2023**

Natural Language Processing has seen impressive gains in recent years. This research includes the demonstration by NLP models to have turned into useful technologies with improved capabilities, measured in terms of how well they match human behavior captured in web-scale language data or through annotations. However, human behavior is inherently shaped by the cultural contexts humans are embedded in, the values and beliefs they hold, and the social practices they follow, part of which will be reflected in the data used to train NLP models, and the behavior these NLP models exhibit. This workshop will bring together NLP researchers invested in this work, along with a community of scholars with multi-disciplinary expertise spanning linguistics, social sciences, and cultural anthropology.

| | |
|---|---|
| 09:00 - 09:15 | ***Opening Remarks*** |
| 09:15 - 10:00 | ***Keynote*** |
| 10:00 - 10:30 | ***Morning talks*** |
| 10:30 - 11:15 | ***Coffee break*** |
| 11:15 - 12:00 | ***In person panel*** |
| 12:00 - 12:45 | ***Contributed Talks*** |
| 12:45 - 14:15 | ***Lunch Break*** |
| 14:15 - 15:45 | ***Contributed Talks*** |
| 15:45 - 16:30 | ***Coffee break*** |
| 16:30 - 17:15 | ***Virtual panel*** |
| 17:15 - 17:55 | ***Contributed Talks*** |

# W4 - The Second Ukrainian Natural Language Processing Workshop

**Organizers:**
Andrii Hlybovets, Oleksii Ignatenko, Oleksii Molchanovskii, Mariana Romanyshyn, Oleksii Syvokon

https://unlp.org.ua/
Venue: Bokar
**Friday, May 5, 2023**

The UNLP workshop brings together academics, researchers, and practitioners in the fields of natural language processing and computational linguistics who work with the Ukrainian language or do cross-Slavic research that can be applied to the Ukrainian language. The Ukrainian NLP community has only started forming in recent years, with most of the projects done by isolated groups of researchers. The UNLP workshop provides a platform for discussion and sharing of ideas, encourages collaboration between different research groups, and improves the visibility of the Ukrainian research community.

| | |
|---|---|
| 09:00 - 09:10 | ***Opening Remarks*** |
| 09:10 - 09:55 | ***Keynote Speech: Mona Diab*** |
| 09:55 - 10:50 | ***Morning Session: New Datasets*** |
| 10:50 - 11:20 | ***Morning Break*** |
| 11:20 - 12:05 | ***Keynote Speech: Gulnara Muratova*** |
| 12:05 - 12:55 | ***Morning Session: New Directions*** |
| 12:55 - 14:25 | ***Lunch*** |
| 14:25 - 16:00 | ***Afternoon Session: Shared Task*** |
| 15:50 - 16:00 | ***Best Paper and Thank You*** |
| 16:00 - 16:30 | ***Afternoon Break*** |
| 16:30 - 18:00 | ***Afternoon Session: UberText*** |
| 18:00 - 18:10 | ***Closing Words*** |

# W5 - Sixth Workshop on Fact Extraction and VERification

**Organizers:**
Mubashara Akhtar, Rami Aly, Christos Christodoulopoulos, Oana Cocarascu,
Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, Andreas Vlachos

`https://fever.ai/workshop.html`
Venue: Divona 2
**Friday, May 5, 2023**

With billions of individual pages on the web providing information on almost every conceivable topic, we should have the ability to collect facts that answer almost every conceivable question. However, only a small fraction of this information is contained in structured sources (Wikidata, Freebase, etc.) – we are therefore limited by our ability to transform free-form text to structured knowledge. There is, however, another problem that has become the focus of a lot of recent research and media coverage: false information coming from unreliable sources. [1] [2] The FEVER workshops are a venue for work in verifiable knowledge extraction and to stimulate progress in this direction.

| | |
|---|---|
| 09:00 - 09:45 | ***Keynote Talk: Iryna Gurevych*** |
| 09:45 - 10:30 | ***Contributed Talks*** |
| 09:45-10:00 | *Hierarchical Representations in Dense Passage Retrieval for Question-Answering*<br>Philipp Ennen, Federica Freddi, Chyi-Jiunn Lin, Po-Nien Kung, RenChu Wang, Chien-Yi Yang, Da-shan Shiu and Alberto Bernacchia |
| 10:00-10:15 | *Enhancing Information Retrieval in Fact Extraction and Verification*<br>Daniel Guzman Olivares, Lara Quijano and Federico Liberatore |
| 10:15-10:30 | *BEVERS: A General, Simple, and Performant Framework for Automatic Fact Verification*<br>Mitchell DeHaven and Stephen Scott |
| 10:30 - 11:15 | ***Coffee break*** |
| 11:15 - 12:00 | ***Keynote Talk: Lucy Lu Wang*** |
| 12:00 - 12:45 | ***Poster session*** |
| | *Rethinking the Event Coding Pipeline with Prompt Entailment*<br>Clément Lefebvre and Niklas Stoehr |
| | *An Entity-based Claim Extraction Pipeline for Real-world Biomedical Fact-checking*<br>Amelie Wuehrl, Lara Grimminger and Roman Klinger |
| | *"World Knowledge" in Multiple Choice Reading Comprehension*<br>Adian Liusie, Vatsal Raina and Mark Gales |
| | *An Effective Approach for Informational and Lexical Bias Detection*<br>Iffat Maab, Edison Marrese-Taylor and Yutaka Matsuo |
| 12:45 - 14:15 | ***Lunch Break*** |
| 14:15 - 15:00 | ***Keynote Talk: Dirk Hovy*** |

| | |
|---|---|
| 15:00 - 15:45 | ***Keynote Talk: Tom Stafford*** |
| 15:45 - 16:30 | ***Coffee break*** |
| 16:30 - 18:00 | ***Panel Discussion: 6 years of FEVER workshops - how far have we come? With Isabelle Augenstein, Mohit Bansal, Christopher Guess, Preslav Nakov, and Tom Stafford.*** |

# W6 - The 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature

**Organizers:**
Stefania Degaetano-Ortlieb, Anna Kazantseva, Nils Reiter, Stan Szpakowicz

https://sighum.wordpress.com/events/latech-clfl-2023/
Venue: Asimon
**Saturday, May 6, 2023**

NLP methods for semantic and structural annotation, intelligent linking, discovery, querying, cleaning and visualization of primary and secondary data for the Humanities, Social Sciences, Cultural Heritage and literary communities.

| | |
|---|---|
| 09:00 - 10:30 | ***Contributed Talks*** |
| 10:30 - 11:15 | ***Coffee break*** |
| 11:15 - 12:45 | ***Contributed Talks*** |
| 12:45 - 14:15 | ***Lunch Break*** |
| 14:15 - 15:45 | ***Poster session*** |
| 15:45 - 16:30 | ***Coffee break*** |
| 16:30 - 18:00 | ***Contributed Talks*** |

# W7 - Workshop Proposal: The Sixth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2023)

**Organizers:**
Atul Kr. Ojha, Chao-Hong Liu, Ekaterina Vylomova, Jade Abbott, Jonathan Washington, Nathaniel Oco, Tommi A Pirinen, Valentin Malykh, Varvara Logacheva, Xiaobing Zhao

https://sites.google.com/view/loresmt/
Venue: Elafiti 4
**Saturday, May 6, 2023**

The LoResMT workshop aims to improve machine translation (MT) coverage for low-resource and under-represented languages through the development of comparable MT systems with relatively small datasets and the evaluation of supplementary natural language processing (NLP) tools' impact on MT output quality.

| | |
|---|---|
| 09:00 - 09:15 | ***Opening Remarks*** |
| 09:15 - 10:05 | ***Invited Talk 1*** |
| 10:05 - 10:30 | ***Session 1: Finding Papers*** |
| 10:30 - 11:15 | ***COFFEE/TEA BREAK*** |
| 11:15 - 12:45 | ***Session 2: Scientific Research Papers*** |
| 11:15-11:35 | *Train Global, Tailor Local: Minimalist Multilingual Translation into Endangered Languages*<br>Zhong Zhou, Jan Niehues and Alexander Waibel |
| 11:35-11:55 | *Measuring the Impact of Data Augmentation Methods for Extremely Low-Resource NMT*<br>Annie Lamar and Zeyneb Kaya |
| 11:55-12:15 | *Language-Family Adapters for Low-Resource Multilingual Neural Machine Translation*<br>Alexandra Chronopoulou, Dario Stojanovski and Alexander Fraser |
| 12:15-12:45 | *Multilingual Bidirectional Unsupervised Translation through Multilingual Finetuning and Back-Translation*<br>Bryan Li, Mohammad Sadegh Rasooli, Ajay Patel and Chris Callison-burch |
| 12:45 - 14:15 | ***Lunch*** |
| 14:15 - 15:00 | ***Invited Talk 2*** |
| 15:00 - 15:30 | ***Session 3: Finding Papers*** |
| 15:30-15:45 | *A Simplified Training Pipeline for Low-Resource and Unsupervised Machine Translation*<br>Àlex R. Atrio, Alexis Allemann, Ljiljana Dolamic and Andrei Popescu-belis |
| 15:45 - 16:30 | ***COFFEE/TEA BREAK*** |
| 16:30 - 18:05 | ***Session 4: Scientific Research Papers*** |
| 16:30-15:16 | *Improving Neural Machine Translation of Indigenous Languages with Multilingual Transfer Learning*<br>Wei-rui Chen and Muhammad Abdul-mageed |

16:50-17:10 *PEACH: Pre-Training Sequence-to-Sequence Multilingual Models for Translation with Semi-Supervised Pseudo-Parallel Document Generation*
Alireza Salemi, Amirhossein Abaskohi, Sara Tavakoli, Azadeh Shakery and Yadollah Yaghoobzadeh

17:10-17:30 *Investigating Lexical Replacements for Arabic-English Code-Switched Data Augmentation*
Injy Hamed, Nizar Habash, Slim Abdennadher and Ngoc Thang Vu

17:30-17:50 *Evaluating Sentence Alignment Methods in a Low-Resource Setting: An English-YorùBá Study Case*
Edoardo Signoroni and Pavel Rychlý

17:50-18:05 *Findings from the Bambara - French Machine Translation Competition (BFMT 2023)*
Ninoh Agostinho Da Silva, Tunde Ajayi, Alex Antonov, Panga Azazia Kamate, Moussa Coulibaly, Mason Del Rio, Yacouba Diarra, Sebastian Diarra, Chris Emezue and Joel Hamilcaro

18:05 - 18:15 ***Closing remarks***

# W8 - The 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP)

**Organizers:**
Koustava Goswami, Alexey Sorokin, Ritesh Kumar, Andrey Shcherbakov, Edoardo M. Ponti, Saliha Muradoğlu, Lisa Beinborn, Ryan Cotterell, Ekaterina Vylomova

https://sigtyp.github.io/workshop.html
Venue: Elafiti 3
**Saturday, May 6, 2023**

The SIGTYP workshop acts as a platform and a forum for the exchange of information between typology-related research, multilingual NLP, and other research areas that can lead to the development of truly multilingual NLP methods. The workshop is specifically aimed at raising awareness of linguistic typology and its potential in supporting and widening the global reach of multilingual NLP, as well as at introducing computational approaches to linguistic typology. It will foster research and discussion on open problems, not only within the active community working on cross- and multilingual NLP but also inviting input from leading researchers in linguistic typology.

| | |
|---|---|
| 08:50 - 09:00 | ***Opening Remarks*** |
| 09:00 - 09:50 | ***Keynote by Ella Rabinovich*** |
| 09:50 - 10:20 | ***Cross-lingual Transfer*** |
| 09:50-10:00 | *Identifying the Correlation Between Language Distance and Cross-Lingual Transfer in a Multilingual Representation Space*<br>Fred Philippy, Siwen Guo and Shohreh Haddadan |
| 10:00-10:10 | *Gradual Language Model Adaptation Using Fine-Grained Typology*<br>Marcell Richard Fekete and Johannes Bjerva |
| 10:10-10:20 | *Cross-lingual Transfer Learning with Persian*<br>Sepideh Mollanorozy, Marc Tanti and Malvina Nissim |
| 10:20 - 10:35 | ***Cross-Lingual Transfer of Cognitive Complexity (Findings)*** |
| 10:35 - 11:15 | ***Break*** |
| 11:15 - 11:30 | ***Does Transliteration Help Multilingual Language Modeling Long (Findings)*** |
| 11:30 - 12:30 | ***Multilinguality*** |
| 11:30-11:40 | *Multilingual BERT has an Accent: Evaluating English Influences on Fluency in Multilingual Models*<br>Isabel Papadimitriou, Kezia Lopez and Dan Jurafsky |
| 11:40-11:55 | *The Denglisch Corpus of German-English Code-Switching*<br>Doreen Osmelak and Shuly Wintner |
| 11:55-12:10 | *Trimming Phonetic Alignments Improves the Inference of Sound Correspondence Patterns from Multilingual Wordlists*<br>Frederic Blum and Johann-Mattis List |
| 12:10-12:20 | *On the Nature of Discrete Speech Representations in Multilingual Self-supervised Models*<br>Badr M. Abdullah, Mohammed Maqsood Shaik and Dietrich Klakow |

| | |
|---|---|
| 12:20-12:30 | *You Can Have Your Data and Balance It Too: Towards Balanced and Efficient Multilingual Models*<br>Tomasz Limisiewicz, Dan Malkin and Gabriel Stanovsky |
| 12:30 - 12:45 | ***Evaluating the Diversity, Equity and Inclusion of NLP Technology: A Case Study for Indian Languages (Findings)*** |
| 12:45 - 13:00 | ***A Large-Scale Multilingual Study of Visual Constraints on Linguistic Selection of Descriptions (Findings)*** |
| 13:00 - 14:15 | ***Lunch (with Linguistic Trivia at 13:45–14:15)*** |
| 14:15 - 15:05 | ***Keynote by Natalia Levshina*** |
| 15:05 - 15:50 | ***Linguistic Complexity*** |
| 15:05-15:20 | *Information-Theoretic Characterization of Vowel Harmony: A Cross-Linguistic Study on Word Lists*<br>Julius Steuer, Johann-Mattis List, Badr M. Abdullah and Dietrich Klakow |
| 15:20-15:35 | *A Crosslinguistic Database for Combinatorial and Semantic Properties of Attitude Predicates*<br>Deniz Özyıldız, Ciyang Qing, Floris Roelofsen, Maribel Romero and Wataru Uegaki |
| 15:35-15:50 | *Revisiting Dependency Length and Intervener Complexity Minimisation on a Parallel Corpus in 35 Languages*<br>Andrew Thomas Dyer |
| 15:50 - 16:10 | ***Break*** |
| 16:10 - 16:40 | ***Shared task on Cognate and Derivative Detection For Low-Resourced Languages*** |
| 16:10-16:20 | *Findings of the SIGTYP 2023 Shared task on Cognate and Derivative Detection For Low-Resourced Languages*<br>Priya Rani, Koustava Goswami, Adrian Doyle, Theodorus Fransen, Bernardo Stearns and John P. McCrae |
| 16:20-16:30 | *ÚFAL Submission for SIGTYP Supervised Cognate Detection Task*<br>Tomasz Limisiewicz |
| 16:30-16:40 | *CoToHiLi at SIGTYP 2023: Ensemble Models for Cognate and Derivative Words Detection*<br>Liviu P. Dinu, Ioan-Bogdan Iordache and Ana Sabina Uban |
| 16:40 - 16:45 | ***Break*** |
| 16:45 - 18:05 | ***Syntax and Morphology*** |
| 16:45-16:55 | *Grambank's Typological Advances Support Computational Research on Diverse Languages*<br>Hannah J. Haynie, Damián Blasi, Hedvig Skirgård, Simon J. Greenhill, Quentin D. Atkinson and Russell D. Gray |
| 16:55-17:05 | *Language-Agnostic Measures Discriminate Inflection and Derivation*<br>Coleman Haley, Edoardo M. Ponti and Sharon Goldwater |
| 17:05-17:20 | *Does Topological Ordering of Morphological Segments Reduce Morphological Modeling Complexity? A Preliminary Study on 13 Languages*<br>Andreas Shcherbakov and Ekaterina Vylomova |
| 17:20-17:35 | *Multilingual End-to-end Dependency Parsing with Linguistic Typology knowledge*<br>Chinmay Choudhary and Colm O'riordan |
| 17:35-17:50 | *Using Modern Languages to Parse Ancient Ones: a Test on Old English*<br>Luca Brigada Villa and Martina Giarda |
| 17:50-18:05 | *Corpus-based Syntactic Typological Methods for Dependency Parsing Improvement*<br>Diego Alves, Božo Bekavac, Daniel Zeman and Marko Tadić |
| 18:05 - 18:10 | ***Best Paper Awards, Closing*** |

# W9 - SlavNLP-2023: The 9th Biennial Workshop on Slavic NLP

**Organizers:**
Jakub Piskorski, Michał Marcińczuk, Preslav Nakov, Maciej Ogrodniczuk, Senja Pollak, Pavel Přibáň, Piotr Rybak, Josef Steinberger, Roman Yangarber

http://bsnlp.cs.helsinki.fi/index.html
Venue: Divona 1
**Saturday, May 6, 2023**

Slavic NLP 2023 is a Workshop that addresses Natural Language Processing (NLP) for the Slavic languages. The goal of this Workshop is to bring together researchers from academia and industry working on NLP for Slavic languages. In particular, the Workshop will serve to stimulate research and foster the creation of tools and resources for these languages.

# W10 - The fourth workshop on Resources for African Indigenous Languages (RAIL)

**Organizers:**
Rooweither Mabuya, Don Mthobela, Mmasibidi Setaka, Menno van Zaanen

https://sadilar.org/index.php/en/2/374
Venue: Divona 2
**Saturday, May 6, 2023**

The Resources for African Indigenous Languages (RAIL) workshop is an interdisciplinary platform for researchers working on resources (data collections, tools, etc.) specifically targeted towards African indigenous languages. In particular, it aims to create the conditions for the emergence of a scientific community of practice that focuses on data, as well as computational linguistic tools specifically designed for or applied to indigenous languages found in Africa.

| | |
|---|---|
| 08:30 - 09:00 | ***Registration and opening remarks*** |
| 09:00 - 10:15 | ***Morning Session 1*** |
| 09:00-09:25 | *IsiXhosa Intellectual Traditions Digital Archive: Digitizing isiXhosa texts from 1870-1914*<br>Jonathan Schoots, Amandla Ngwendu, Jacques De Wet and Sanjin Muftic |
| 09:25-09:50 | *Preparing the Vuk'uzenzele and ZA-gov-multilingual South African multilingual corpora*<br>Richard Lastrucci, Jenalea Rajab, Matimba Shingange, Daniel Njini and Vukosi Marivate |
| 09:50-10:15 | *Automatic Spell Checker and Correction for Under-represented Spoken Languages: Case Study on Wolof*<br>Thierno Ibrahima Cissé and Fatiha Sadat |
| 10:15 - 10:55 | ***Morning tea break*** |
| 10:55 - 12:30 | ***Morning Session 2*** |
| 10:55-11:20 | *Discourse reporting database: annotated corpora of West African traditional narratives*<br>Ekaterina Aplonova, Izabela Jordanoska, Timofey Arkhangelskiy and Tatiana Nikitina |
| 11:20-11:45 | *Analyzing political formation through historical isiXhosa text analysis: Using frequency analysis to examine emerging African Nationalism in South Africa*<br>Jonathan Schoots |
| 11:45-12:05 | *Unsupervised Cross-lingual Word Embedding Representation for English-isiZulu*<br>Derwin Ngomane, Rooweither Mabuya, Jade Abbott and Vukosi Marivate |
| 12:05-12:30 | *Investigating Sentiment-Bearing Words- and Emoji-based Distant Supervision Approaches for Sentiment Analysis*<br>Ronny Mabokela, Mpho Roborife and Turguy Celik |
| 12:30 - 14:00 | ***Lunch break*** |
| 14:00 - 15:40 | ***Afternoon Session 1*** |
| 14:00-14:25 | *Towards a Swahili Universal Dependency Treebank: Leveraging the Annotations of the Helsinki Corpus of Swahili*<br>Kenneth Steimel, Sandra Kübler and Daniel Dakota |
| 14:25-14:50 | *Evaluating the Sesotho rule-based syllabification system on Sepedi and Setswana words* |

| | |
|---|---|
| | Johannes Sibeko and Mmasibidi Setaka |
| 14:50-15:15 | *Deep learning and low-resource languages: How much data is enough? A case study of three linguistically distinct South African languages*<br>Roald Eiselen and Tanja Gaustad |
| 15:15-15:40 | *Comparing methods of orthographic conversion for Bàsàá, a language of Cameroon*<br>Alexandra O'neil, Daniel Swanson, Robert Pugh, Francis Tyers and Emmanuel Ngue Um |
| 15:40 - 16:20 | ***Afternoon tea break*** |
| 16:20 - 18:00 | ***Afternoon Session 2*** |
| 16:45-17:10 | *Natural Language Processing in Ethiopian Languages: Current State, Challenges, and Opportunities*<br>Atnafu Lambebo Tonja, Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Moges Ahmed Mehamed, Olga Kolesnikova and Seid Muhie Yimam |
| 17:10-17:35 | *A Corpus-Based List of Frequently Used Words in Sesotho*<br>Johannes Sibeko and Orphée De Clercq |
| 17:35-18:00 | *Vowels and the Igala Language Resources*<br>Mahmud Momoh |
| 18:00 - 18:05 | ***Closing statements*** |

# W11 - 19th Workshop on Multiword Expressions

**Organizers:**
Marcos Garcia, Voula Giouli, Shiva Taslimipoor, Lifeng Han, Archna Bhatia, Kilian Evang

https://multiword.org/mwe2023/
Venue: Bokar
**Saturday, May 6, 2023**

The MWE 2023 Workshop, organized and sponsored by the Special Interest Group on the Lexicon (SIGLEX) for ACL, focuses on processing, identification, and/or interpretation of MWEs, be it in specialized domains, low-resource languages, for end-user-applications or for understanding their representation in and modeling using pre-trained language models. The topic of the workshop, multiword expressions (MWEs), presents an interesting research area due to the lexical, syntactic, semantic, pragmatic, and/or statistical idiosyncrasies MWEs exhibit. Given their irregular nature, they pose complex problems in linguistic modeling (e.g. annotation), NLP tasks (e.g. parsing), and end-user applications (e.g. natural language understanding and MT). For the past two decades, modeling and processing MWEs for NLP has been the topic of the MWE workshop. Impressive progress has been made in the field, but our understanding of MWEs still requires much research considering their need and usefulness in NLP applications. This is also relevant to domain-specific NLP pipelines that need to tackle terminologies that often manifest as MWEs. Therefore, for this 19th edition of the MWE workshop, we collaborated with the Clinical NLP Workshop (ACL 2023) and organized a special track on “MWEs in Clinical NLP” to draw interest to domain-specific NLP pipelines that tackle terminologies involving MWEs.

| | |
|---|---|
| 08:30 - 09:00 | ***Registration*** |
| 09:00 - 09:10 | ***Opening*** |
| 09:10 - 10:30 | ***Oral long paper presentations*** |
| 09:10-09:30 | *Are Frequent Phrases Directly Retrieved like Idioms? An Investigation with Self-Paced Reading and Language Models*<br>Giulia Rambelli, Emmanuele Chersoni, Marco S. G. Senaldi, Philippe Blache and Alessandro Lenci |
| 09:30-09:50 | *A Survey of MWE Identification Experiments: The Devil is in the Details*<br>Carlos Ramisch, Abigail Walsh, Thomas Blanchard and Shiva Taslimipoor |
| 09:50-10:10 | *PARSEME corpus release 1.3*<br>Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze and Abigail Walsh |
| 10:10-10:30 | *Predicting Compositionality of Verbal Multiword Expressions in Persian*<br>Mahtab Sarlak, Yalda Yarandi and Mehrnoush Shamsfard |
| 10:30 - 11:15 | ***Morning coffee break*** |
| 11:15 - 12:15 | ***Keynote Talk, Leo Wanner: Lexical collocations: Explored a lot, still a lot more to explore*** |
| 12:15 - 12:45 | ***Oral short paper presentations*** |

12:15-12:30 *Romanian Multiword Expression Detection Using Multilingual Adversarial Training and Lateral Inhibition*
Andrei Avram, Verginica Barbu Mititelu and Dumitru-Clementin Cercel

12:30-12:45 *The Better Your Syntax, the Better Your Semantics? Probing Pretrained Language Models for the English Comparative Correlative*
Leonie Weissweiler, Valentin Hofmann, Abdullatif Koksal and Hinrich Schütze

12:45 - 14:15 ***Lunch Break***

14:15 - 14:45 ***Keynote Talk, Asma Abacha and Goran Nenadic: MWEs in ClinicalNLP and Healthcare Text Analytics***

14:45 - 15:15 ***Oral short paper presentations***

14:45-15:00 *Detecting Idiomatic Multiword Expressions in Clinical Terminology using Definition-Based Representation Learning*
François Remy, Alfiya Khabibullina and Thomas Demeester

15:00-15:15 *Investigating the Effects of MWE Identification in Structural Topic Modelling*
Dimitrios Kokkinakis, Ricardo Sánchez, Sebastianus Bruinsma and Mia-Marie Hammarlin

15:15 - 15:45 ***Panel discussion: Multiword Expressions in Knowledge-intensive Domains: Clinical Text as a Case Study***

15:45 - 16:30 ***Afternoon coffee break***

16:30 - 17:15 ***Poster session***

*Idioms, Probing and Dangerous Things: Towards Structural Probing for Idiomaticity in Vector Space*
Filip Klubička, Vasudevan Nedumpozhimana and John Kelleher

*Simple and Effective Multi-Token Completion from Masked Language Models*
Oren Kalinsky, Guy Kushilevitz, Alexander Libov and Yoav Goldberg

*Annotation of lexical bundles with discourse functions in a Spanish academic corpus*
Eleonora Guzzi, Margarita Alonso-Ramos, Marcos Garcia and Marcos García Salido

*Enriching Multiword Terms in Wiktionary with Pronunciation Information*
Lenka Bajcetic, Thierry Declerck and Gilles Sérasset

*Automatic Generation of Vocabulary Lists with Multiword Expressions*
John Lee and Adilet Uvaliyev

*A MWE lexicon formalism optimised for observational adequacy*
Adam Lion-Bouton, Agata Savary and Jean-Yves Antoine

17:15 - 17:45 ***Oral short paper presentations***

17:15-17:30 *Token-level Identification of Multiword Expressions using Pre-trained Multilingual Language Models*
Raghuraman Swaminathan and Paul Cook

17:30-17:45 *Graph-based multi-layer querying in Parseme Corpora*
Bruno Guillaume

17:45 - 18:00 ***Closing***

# W12 - The Second Workshop on NLP Applications to Field Linguistics

**Organizers:**
Oleg Serikov, Elena Klyachko, Francis Tyers, Tatiana Shavrina, Ekaterina Vylomova, Éric Le Ferrand, Ekaterina Voloshina, Anna Postnikova, Ekaterina Neminova

`https://field-matters.github.io/`
Venue: Asimon
**Saturday, May 6, 2023**

Field linguistics plays a crucial role in the development of linguistic theory and universal language modelling, as it provides uncontested, the only way to obtain structural data about the rapidly diminishing diversity of natural languages. The Field matters workshop aims to bring together the urgent needs of field linguists and the vast community of NLP practitioners, developing up-to-date NLP tools for easier, faster, more reliable data collection and annotation.

| | |
|---|---|
| 09:00 - 10:30 | ***Invited talk. Lane Schwartz.*** |
| 11:15 - 12:45 | ***Invited talk. Emmanuel Schang.*** |
| 12:45 - 14:15 | ***lunch break*** |
| 14:15 - 15:45 | ***Presentations*** |
| 15:45 - 16:30 | ***Coffee Break*** |
| 16:30 - 18:00 | ***Presentations*** |

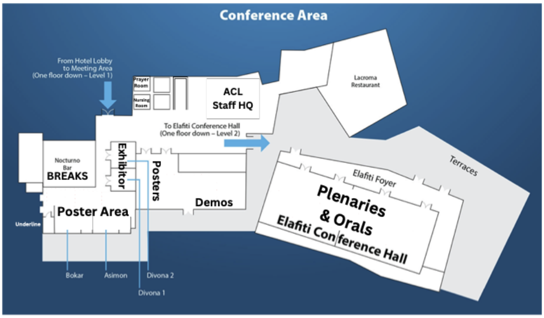
# 10

# Venue Map

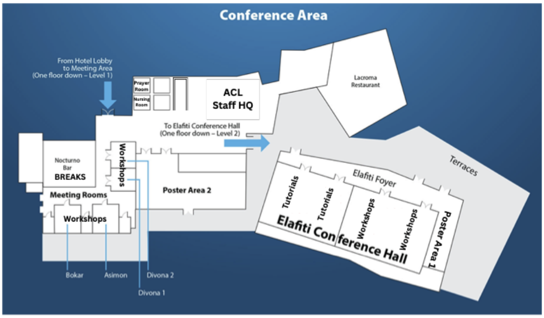
## Valamar Lacroma Dubrovnik Hotel Floor Plan

The Valamar Lacroma Dubrovnik Hotel is located on the tranquil Babin Kuk peninsula, surrounded by lush greenery and soothing white pebble beaches, only 15 minutes from the old town of Dubrovnik.

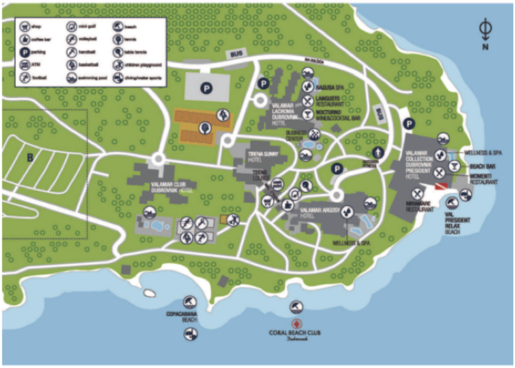
Main Conference Layout



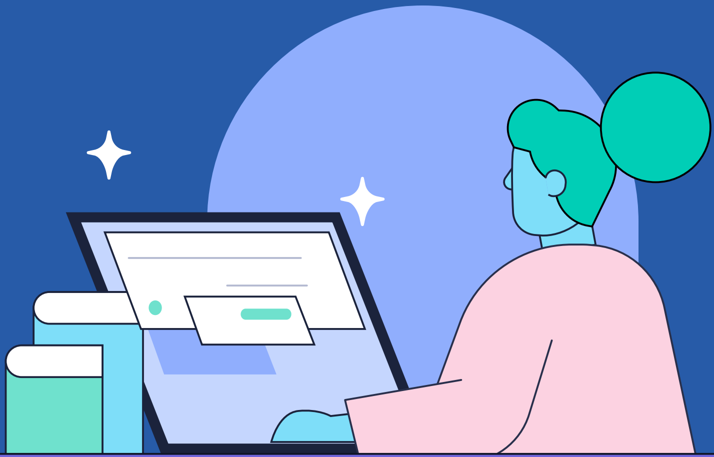
## Workshops & Tutorials Layout

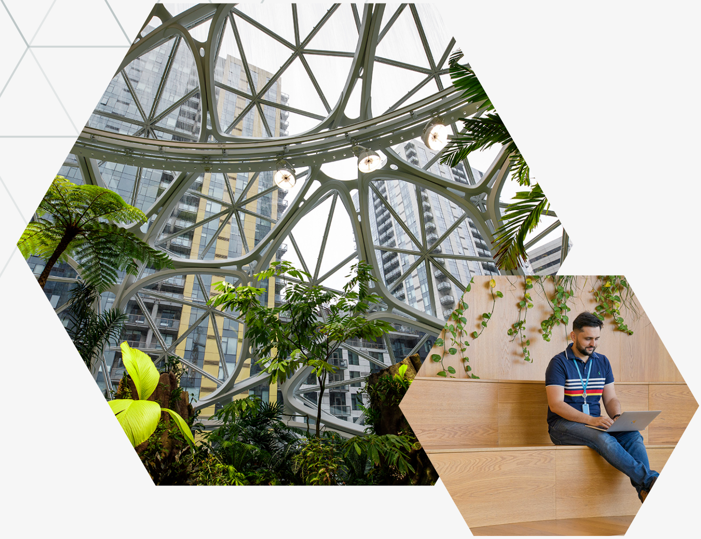


## Valamar Hotels Map



# Index





LIVEPERSON®

# 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/

Welcome Event Sponsor and Diamond Sponsor

grammarly

Diamond Sponsor

LIVEPERSON

Platinum Sponsor and D&I Ally

amazon | science

Platinum Sponsor

Bloomberg

Engineering

Silver Sponsor

duolingo

Bronze Sponsors

Adobe

Babelscape