
EMNLP 2022
Dec 7–11, Abu Dhabi

# Contents

# 1 Conference Information

## Message from the General Chair

I am delighted to welcome you to EMNLP 2022! I believe this conference has been complicated beyond any precedent. Over the past year, it's been thrilling to see the organization team approach each new puzzle with creativity and enthusiasm. We hope that those participating in Abu Dhabi as well as those joining remotely will leave the conference feeling newly inspired by the program and newly connected to our ever-growing community. Following EMNLP 2021 and major NLP conferences since, EMNLP 2022 is "hybrid," serving both virtual and in-person participants.

Our key innovations for EMNLP 2022 include:

- EMNLP 2022 is "hybrid" in a second sense, as well: we allowed both direct and rolling review paper submissions, building on the pilot experiment of EMNLP 2021, which considered a small number of ARR submissions.
- Familiar from NAACL but new to EMNLP, we've added an industry track.
- During the conference, "portals" will link virtual poster sessions to in-person conference participants during poster sessions each day.
- The first *ACL-family conference in the United Arab Emirates.

My sincere thanks go to all the members of our organization team; here I list by name the leaders but recognize with gratitude that they represent a much larger population of volunteers who have made EMNLP 2022 possible.

- The program chairs — Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang — who made the bold decision to take the training wheels off of rolling review, making their jobs much harder but taking an important step for the community.
- The senior area chairs, area chairs, and reviewers whose collective work improved not only the papers in the EMNLP 2022 proceedings and findings volumes, but also papers to appear in future venues.

- The diversity and inclusion chairs — Meriem Beloucif, Thamar Solorio and Andreas Vlachos — who were tireless advocates for the inclusive culture our conference aspires to.

- The demonstration chairs — Wanxiang Che and Ekaterina Shutova — who selected an exciting set of demos to enliven the program.

- The workshop chairs — Asli Celikyilmaz and Daniel Hershcovich — who shepherded a diverse collection of satellite events to complement the main conference.

- The tutorial chairs — Samhaa R. El-Beltagy and Xipeng Qiu — who curated an exciting set of tutorials for the benefit of attendees.

- The publications chairs — Ryan Cotterell, Steffen Eger, and Sam Wiseman — who not only ensured a legacy of high-quality artifacts for EMNLP 2022, but also continued to improve the tools and workflow to serve future meetings.

- The student volunteer chairs — Houda Bouamor (who is also thanked for her role on the local team and chairing WANLP), Hanan AlDarmaki, and Ashutosh Modi — who brought exceptional enthusiasm to recruiting and organizing the student volunteer team.

- The student volunteers themselves, who are critical to the success of our mostly volunteer-run conference.

- The virtual infrastructure chairs — Wassim El Hajj and Hao Fang — who worked to ensure that the virtual experience will be as smooth and engaging as possible.

- In a new position, the poster session chair — Jordan Boyd-Graber — has pushed us to improve the virtual experience and make it more unified with the in-person one.

- The industry track chairs — Angeliki Lazaridou and Yunyao Li — who brought EMNLP our first industry track and set the bar high for the future.

- The ethics chairs — Lea Frermann and Margot Mieskes — who oversaw the important task of ethical review.

- The publicity chairs — Eunsol Choi and Wajdi Zaghouani — and website chairs — Zhaochun Ren and Fajie Yuan — who kept our community abreast of important developments as our plans for the conference unfolded.

- The reviewer mentoring chairs — David Mimno and Yanyan Lan — who worked to help initiate newcomers to our peer review processes.

- The sponsorship chairs — Mingxuan Wang and Imed Zitouni — who worked alongside ACL's sponsorship director Chris Callison-Burch to ensure that the conference was in a strong fiscal position.

- Last but not least, the local team, led by Nizar Habash, whose tireless efforts would warrant a *1,001 Nights* literary treatment if not for confidentiality concerns, and who deserves the deepest gratitude of the community and a very long vacation. The local NLP community, both officially on the local committee and not, including Eric Xing, Tim Baldwin, Bashar Alhafni, Go Inoue, and many more, deserve our thanks as well.

I also want to express special thanks to Priscilla Rasmussen and Jenn Rachford of the ACL business office for their endless guidance and attention to detail on all aspects of organizing this huge event. In Abu Dhabi, Thembi Kuchena (ADNEC) and Zenab Mohamed (DCT) were instrumental from the initial bid and throughout the process. The Underline team, led by Damira Mršić, was immensely helpful in keeping us on schedule, and in many other ways.

Finally, thanks to our sponsors, without whom this conference would not be possible:

**Supporting partner**: Abu Dhabi Convention and Exhibition Bureau.

**Diamond sponsors**: Amazon, Apple, Bloomberg, Google, Meta, Mohamed bin Zayed University of Artificial Intelligence, and New York University Abu Dhabi.

**Platinum sponsors**: Baidu, ByteDance, Megagon, and Microsoft.

**Gold sponsors**: Beyond Limits, Cohere.ai, Huawei, ServiceNow, and the Technology Innovation Institute in Abu Dhabi.

**Silver sponsors**: Duolingo, Naver Labs, and Translated.

**Bronze sponsors**: Adobe, aiXplain, Babelscape, CAIR, Grammarly, HLTCOE, and NEC Laboratories Europe.

**Diversity and Inclusion Champions**: Google, Microsoft and New York University Abu Dhabi. ACL SIGDAT has also contributed to supporting scholarships for attending the conference.

Noah A. Smith
University of Washington and Allen Institute for AI, Seattle, Washington, USA
EMNLP 2022 General Chair

# Message from the Program Chairs

Welcome to EMNLP 2022, which is one of the most-attended conferences in the field of natural language processing, held in "hybrid" mode this year serving both virtual and in-person participants in Abu Dhabi!

**Submission and Acceptance** EMNLP 2022 received 4190 full paper submissions, the largest number to date. This number includes 275 ARR papers that were committed to EMNLP (see further discussion of ARR below). 225 papers were desk rejected for various reasons (missing limitation section, anonymity policy, multiple submission policy or formatting violations), leaving us with 3965 submissions that were fully reviewed. Despite the record-breaking number of submissions, based on the reviewers, area chairs and senior area chair comments, we kept the EMNLP 2022 acceptance rate similar to previous events, and accepted **829** papers to the main conference. Out of these, 175 are oral presentations and 654 poster presentations. Similarly to prior years, we also accepted 549 papers for "Findings of EMNLP". The EMNLP 2022 program also features 39 papers from the Transactions of the Association for Computational Linguistics (TACL) and Computational Linguistics (CL) journals. More statistics on the accepted papers can be found below.

| | Long | Short | Total |
|---|---|---|---|
| Submitted (Including ARR commits) | 3242 | 948 | 4190 |
| Accepted as Oral | 163 | 12 | 175 |
| Accepted as Poster | 552 | 102 | 654 |
| Acceptance Rate (main conference) | 22% | 12% | 20% |
| Accepted to Findings | 453 | 96 | 549 |
| Acceptance Rate (Findings) | 14% | 10% | 13% |
| Presented TACL papers | – | – | 27 |
| Presented CL papers | – | – | 12 |

**Limitations Section** One innovation of EMNLP 2022 is the requirement that each submitted paper must include an explicitly named Limitations section, discussing the limitations of the work. This discussion does not count towards the page limit, and we asked reviewers to not use the mentioned limitations as reasons to reject the paper, unless there is a really good reason to. The effect was overall positive: most papers included a limitations section, and many of them were informative. We hope to see this requirement continue in the future conferences.

**Tracks** To ensure a smooth process, the submissions to EMNLP 2022 were divided into 26 tracks. The tracks mostly followed these of previous EMNLP conferences, reflecting the "standard" divisions in the field. We did however make the following changes: the "Machine Translation and Multilinguality" track was split into two separate tracks ("Machine Translation" and "Multilinguality"); the "Syntax: Tagging, Chunking and Parsing" track was renamed to "Syntax, Parsing and their Applications"; and we added 4 additional tracks, reflecting upcoming trends in the research landscape: Commonsense Reasoning; Language Models and Analysis of Language Models; Efficient Methods for NLP; and Semi-supervised and Weakly-supervised Methods. Finally, we also solicited papers for a "Theme Track", discussing Open questions, major obstacles, and unresolved issues in NLP. Of the 26 tracks, the Resources and Evaluation, NLP Applications, Machine Learning for NLP, Dialogue and Interactive Systems, Language Modeling and Analysis of Language Models, Speech, Vision, Robotics, Multimodal Grounding and Information Extraction tracks were the most popular with over 200 submissions per track.

**Program committee structure & reviewing** Similar to prior NLP conferences, we adopted the hierarchical program committee structure, where for each area we invited between 1 to 5 Senior Area Chairs (SACs), who worked with a team of Area Chairs (ACs), and an army of reviewers. We relied on statistics from prior years to estimate how many SACs, ACs and reviewers will be needed and ended up with 70 SACs and 297 ACs. For the reviewers, we used the reviewer lists from prior EMNLP conferences, solicited volunteer reviewers, and also invited all EMNLP 2022 authors to serve as reviewers. To this end, we attempted to recruit the most competent and matching reviewers by making a Google Form in reviewer recruitment, which was publicized through different channels and is linked to the author information page for all the submission authors to fill. We then provided the resulting information to the program committee for making reviewer assignments. This resulted in a reviewer pool of 4647 reviewers, of which 3828 reviewers were assigned at least one paper to review. For each submission, we assigned three reviewers and an AC. The initial paper assignment was made using an automatic algorithm that matches the abstracts with ACs/reviewers' past publication records, then the assignments were further refined by the SACs/PCs. We tried to avoid the Toronto matching score which had its limitations. In the end, most reviewers were assigned less than 6 papers, with a few reviewers working on 10 assignments and a large number of reviewers working on 1 assignment only. In the Google Form, we also asked whether the reviewer will volunteer for emergency review, which turned out to be very useful. We adapted the review forms from NAACL 2022, and ACL 2022 and EMNLP 2021. Besides the overall recommendation, reviewers were asked to evaluate how reproducible the results in the paper were, and whether there was any ethical concern. To ensure the review quality, we provided detailed guidelines about what reviewers should and shouldn't do in a review. We made final decisions according to the rankings and SAC recommendations. Our final decisions were made not just on the review scores, but also took into account the reviews, author responses, discussions among reviewers, meta-reviews and SAC/AC recommendations.

**Ethics committee** We also formed an Ethics Committee (EC) dedicated to ethical issues. After the technical reviews, but before the author-response and discussion phases, the ethics committee considered 150 papers that were flagged by the technical reviewing committee for ethical concerns. The two EC chairs went over the papers, to determine whether a full EC review would be required. As a result, 22 papers received two dedicated ethics reviews from a committee of 41 reviewers recruited by the EC chairs. An additional 22 papers raised one or more recurring issues (e.g., no information on annotator payment), which did not require a full ethics review, but were raised to the authors by the EC chairs. For any paper that was recommended to be accepted based on technical reviews and that had been referred to the EC, the EC chairs recommended one of the following to the PC chairs: (a) accept with comments (authors should take concerns raised in the ethical review into account in the camera-ready version; 29 papers); (b) conditionally accept (the ethical issues must be addressed in the camera-ready version, to be verified by the EC prior to final acceptance; 15 papers); and (c) reject due to ethical issues (0 papers). 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 meta-reviews. The EC chairs 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. The EC chairs thank their committee for the excellent work.

**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 EMNLP 2022, we decided to run a process which is separate from ARR, but allows for ARR submissions. Specifically, authors could either submit papers to EMNLP 2022 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 provides when committing their ARR submission to EMNLP. The ARR commission 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 of the given

tracks, together with the direct submissions in the track, and based on the reviews and meta-reviews from ARR. Overall, EMNLP had 275 papers committed from ARR, of these 97 were accepted to the main conference and 73 were accepted to Findings of EMNLP.

**Best paper selection** Based on the nominations from SACs and ACs, we identified 11 candidates for the best papers and outstanding papers award. These papers are assessed by the Best Paper Award Committee. The award winners will be announced and present their works at the closing ceremony.

**Presentation Mode** We attempted the decision for oral vs poster presentations not to be made based on the quality/merit of the papers, but rather on the authors' interest in the presentation mode, and our understanding of what would be the best format for presentation of each individual paper.

**Keynote talks** Another highlight of our program is the three exciting keynote talks, presented by prof. Gary Marcus, NYU (Emeritus) on "Towards a Foundation for AGI"; prof. Neil Cohn, Tilburg University, Department of Communication and Cognition on "The multimodal language faculty and the visual languages of comics" & Dr. Mona Diab from Meta Responsible AI on "Towards a Responsible NLP: Walking the walk".

**Gratitude** We would like to thank the following people for their support & contributions:

- Our General Chair, Noah Smith, who led the whole organizing team, and helped with many of the decision processes;
- 70 SACs who helped us throughout the entire review process, from assigning papers, checking review quality, making final recommendation, suggesting presentation formats to recommending best paper candidates;
- 297 ACs who checked the initial submissions, led paper discussions, wrote meta reviews, ensured review quality and suggested best paper candidates;
- 3828 reviewers who reviewed the papers and actively participated in paper discussions; special thanks to those who stepped in at the last minute to serve as emergency reviewers;
- Reviewing Mentoring Chairs, David Mimno and Yanyan Lan and the team of mentors they recruited, for mentoring new reviewers.
- 36 Ethics Committee members, chaired by Lea Frermann and Margot Mieskes, for their hard work to provide ethical reviews and meta-reviews for all papers with serious ethical issues, and ensure that all the conditionally accepted papers have addressed the ethical issues appropriately;
- Best Paper Award Committee: Claire Cardie, Ellen Riloff, Hal Daume III, Rada Mihalcea, Raymond Mooney and Preslav Nakov for selecting the best papers;
- Jordan Boyd-Graber, the Virtual Posters chair, for organizing and managing the virtual poster sessions.
- Publication Chairs Ryan Cotterell, Steffen Eger and Sam Wiseman for completing the final proceedings within a short period;
- ACL Anthology Director Matt Post and his team, for his help in the production of the conference proceedings and maintenance of the ACL Anthology;
- TACL editor-in-chief Brian Roark and TACL Editorial Assistant Cindy Robinson, and CL Editor-in-Chief Hwee Tou Ng for coordinating TACL and CL presentations with us;

- The ARR team for their continued effort in running ARR, and for coordination with us. Particular thanks to Harold Rubio and Goran Glavas for multiple rounds of technical help in extracting data from the ARR OpenReview system.
- Rich Gerber at SoftConf, for setting up EMNLP 2022 conference site and quickly responding to our emails and resolving any issues we encountered with the START system;
- Website chairs Fajie Yuan and Zhaochun Ren and their team for continued effort in prompt updates to the website.
- Publicity chairs Eunsol Choi and Wajdi Zaghouani for publicizing the conference and handling communications on social media.
- Sol Rosenberg, Damira Mršić and the whole Underline team, for helping to manage the logistics of both the virtual and online conference.
- Jenn Rachford and Priscilla Rasmussen for their professional and very valuable help in organizing the logistics of the conference.
- Nizar Habash and the rest of the Local Organizing Committee, for various discussions on organizing EMNLP, and making the local arrangements.
- 11854 authors for submitting their work to EMNLP 2022.

We hope that you will enjoy this year's program and hybrid conference!

Yoav Goldberg, Bar Ilan University and Allen Institute for AI
Zornitsa Kozareva, SliceX AI
Yue Zhang, Westlake University
EMNLP 2022 Program Co-Chairs

# Organizing Committee

**General Chair**

Noah Smith, University of Washington and Allen Institute for AI

**Program Chairs**

Yoav Goldberg, Bar Ilan University and Allen Institute for AI
Zornitsa Kozareva, SliceX AI
Yue Zhang, Westlake University

**Industry Track Chairs**

Angeliki Lazaridou, DeepMind
Yunyao Li, Apple

**Poster Session Chair**

Jordan Boyd-Graber, University of Maryland

**Workshop Chairs**

Asli Celikyilmaz, Meta AI
Daniel Hershcovich, University of Copenhagen

**Tutorials Chairs**

Samhaa R. El-Beltagy, Newgiza University and Optomatica
Xipeng Qiu, Fudan University

**Ethics Chairs**

Lea Frermann, University of Melbourne
Margot Mieskes, Darmstadt University of Applied Sciences

**Demonstrations Chairs**

Wanxiang Che, Harbin Institute of Technology
Ekaterina Shutova, University of Amsterdam

**Publications Chairs**

Ryan Cotterell, ETH Zürich
Steffen Eger, Bielefeld University
Sam Wiseman, Duke University

**Publicity Chairs**

Eunsol Choi, University of Texas at Austin
Wajdi Zaghouani, Hamad Bin Khalifa University

**Student Volunteer Chairs**

Houda Bouamor, Carnegie Mellon University in Qatar
Ashutosh Modi, Indian Institute of Technology Kanpur
Hanan AlDarmaki, Mohamed bin Zayed University of Artificial Intelligence

**Virtual Infrastructure Chairs**

Wassim El Hajj, American University of Beirut
Hao Fang, Microsoft

**Website Chairs**

Zhaochun Ren, Shandong University
Fajie Yuan, Westlake University

**Diversity, Inclusion, and Outreach Chairs**

Thamar Solorio, University of Houston
Andreas Vlachos, University of Cambridge
Meriem Beloucif, Uppsala University

**Reviewer Mentoring Chairs**

David Mimno, Cornell University
Yanyan Lan, Tsinghua University

**Sponsorship Chairs**

Mingxuan Wang, ByteDance AI Lab
Imed Zitouni, Google

**Local Arrangements**

Nizar Habash, New York University Abu Dhabi
Eric Xing, Mohamed bin Zayed University of Artificial Intelligence

# COVID-19 Safety

EMNLP 2022 is adhering to the ACL precaution and safety measure guidelines to make sure this event is safe for everyone. See https://www.2022.aclweb.org/covid-19-safety.

**Before you leave**
We strongly encourage all participants to test before traveling to the conference. You should not travel if your test is positive.

Know the current Government Policies for Vaccinated and Unvaccinated Travel to the UAE: https://ae.usembassy.gov/covid-19-information/

Please ensure that you have checked with your airline to understand their policies regarding COVID-19 and have adequate travel insurance in place if required.

**At the conference**
The ACL has worked with the Covid Testing facilities inside of the ADNEC for a rapid antigen test. We strongly ask that you test before the day before the start of the workshops and tutorials and again before the start of the Main conference.

To ensure everyone's safety, and with the purpose of making everyone comfortable at the conference, please carefully consider the following if you are attending in person:

- The ACL strongly encourages our attendees to bring and wear high-quality, well-fitting masks while attending indoor functions at ACL events, removing their masks only the minimal amount necessary to eat and drink.
- The ACL strongly encourages attendees to be cautious with masking in indoor space outside the conference venue (e.g., on public transit or while eating away from the venue) to avoid bringing COVID to the venue.
- If you start to feel ill or test positive during the conference, please leave the venue immediately, and do not come to the venue anymore.
- To avoid crowded situations, we recommend making optimal use of the available spaces, particularly during coffee breaks.
- Speakers and session chairs are not required to wear masks while presenting, as there is enough distance between speakers and audience.

**Space and distancing**
We come from different countries and different personal experiences about COVID-19, and for many of us EMNLP 2022 is the first hybrid conference after the relaxation of the Covid protocols. In order to favor interpersonal relations, we will use a "traffic light system" with stickers to attach to your badge to indicate your feelings on social distancing:

- Red: keep a distance
- Amber: being cautious, no touching
- Green: handshakes and hugs ok

**Plan ahead to return home**
Please ensure you know the COVID-19 testing requirements for return travel back into your country of origin. The requirements are available on your government website.

2

# Anti-Harassment Policy

EMNLP 2022 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 Social Events

We have planned the following events for EMNLP 2022. Please follow ACL's code of conduct and COVID policy when you are attending these events.

**Welcome Reception** - **Thursday, December 8th, 2022**
Venue: **Aloft Abu Dhabi**
Time: **19.00 - 22.00**

A Welcome Reception will be held on Thursday December 8, 2022 from 7:00pm to 10:00pm. Join us at the Aloft Abu Dhabi Hotel where you can meet and make new friends and catch up with your colleagues over a drink (alcoholic & nonalcoholic drink ticket to be provided) and light canapes.

Admission to the welcome reception is included in the main conference in-person registration. Additional tickets can be bought on site.
Additional Welcome Reception Ticket - $50.
Welcome Reception - Under 10 - $25.

**Social Event** - **Saturday, December 10th, 2022**
Venue: **Ritz-Carlton Abu Dhabi, Grand Canal Hotel**
Time: **19.00 - 22.00**

The EMNLP Social Event will be held on Saturday December 10, 2022 from 7:00pm – 10:00pm at the Ritz-Carlton Abu Dhabi, Grand Canal Hotel. With its captivating sunsets and extraordinary view of the Sheikh Zayed Grand Mosque, you will be sure to enjoy an evening out at our Social Event with luxurious foods and libations (alcoholic and nonalcoholic drink ticket provided), along with Abu Dhabi's local entertainment compliments of the Department of Tourism, and a DJ for some evening dancing.

Transportation to the event:
Charter buses will be available from ADNEC to the Ritz Carlton beginning at 6:30pm to 10:30pm. Departures from the Ritz will stop by Andaz, Aloft and Adnec.

Admission to the social event is included in the main conference in-person registration. Additional tickets can be bought on site.
Social Event Ticket Additional Guest - $125.
Social Event Ticket - Under 10 - $65.
Social Event Ticket - Exhibitor - $125.

# 4 Keynotes

## The Multimodal Language Faculty and the Visual Languages of Comics

**Neil Cohn**
Tilburg University, Department of Communication and Cognition

**Friday, December 9, 2022** - Room: **Hall B** - Time: **9:00-10:30**

**Abstract:** Contrary to the notions of language as an amodal system, natural human communication is multimodal and combines speech, gestures, writing, and pictures. To account for this, recent work has proposed that our vocal, bodily, and graphic modalities persist in parallel in a multimodal language faculty, and both unimodal and multimodal expressions arise out of emergent states of a shared architecture. Such a model carries different expectations for the ways in which modalities may be similar or different from each other, and how they may interact. I will highlight these properties specifically for our graphic modality, which I argue can manifest in full visual languages when displaying both a systematic lexicon and complex grammar. I will use analysis of a corpus of several hundred annotated comics to show distinctive patterns that suggest they are drawn in different visual languages. Yet, I will also show that consistent

"universal" linguistic principles persist across this structural diversity. Finally, I will argue that a multimodal language faculty requires us to change our conception of linguistic relativity, and I will show how subtle structures of spoken languages permeate across to visual languages. Altogether, this work argues for a multimodal basis of linguistic structure, and heralds a reconsideration of what constitutes the language system.

**Bio:** Neil Cohn is an American cognitive scientist best known for his research on the overlap in structure and cognition between language and graphic communication like comics and emoji. He is the author of 80+ academic papers, 4 academic books, and 2 graphic novels. He received his PhD in cognitive psychology at Tufts University and is currently an associate professor at the Department of Cognition and Communication at Tilburg University in The Netherlands. His work can be found online at `https://www.visuallanguagelab.com/`.

# Takeaways from a Systematic Study of 75K Models on Hugging Face

**Nazneen Rajani**
Hugging Face

**Friday, December 9, 2022** - Room: **Hall B** - Time: **17:30-18:30**

**Abstract:** Language models trained using transformers dominate the NLP model landscape, making Hugging Face (HF) the defacto hub for sharing, benchmarking, and evaluating NLP models. The HF hub provides a rich resource for understanding how language models evolved, opening up research questions such as 'Is there a correlation between model documentation and its usage?', 'How have the models evolved?', 'What do users document about their models?'. In the first part of my talk, I'll give a macro-level view of how the NLP model landscape has evolved based on our systematic study of 75K HF models.

In the second part, I'll discuss advances, challenges and opportunities in evaluating and documenting NLP models developed in an industry setting. Based on the results, do we see a paradigm shift from model-centric to data-centric evaluation and documentation?

**Bio:** Nazneen is a Research Lead at Hugging Face, a startup with a mission to democratize ML, leading data-centric ML research which involves systematically analyzing, curating, and automatically annotating data. Before HF, she worked at Salesforce Research with Richard Socher and led a team of researchers focused on building robust natural language generation systems based on LLMs. She completed her Ph.D. in CS at UT-Austin with Prof. Ray Mooney.

Nazneen has over 30 papers accepted at ACL, EMNLP, NAACL, NeurIPs, and ICLR and has her research covered by Quanta magazine, VentureBeat, SiliconAngle, ZDNet, and Datanami. She is also teaching a course on interpreting ML models with Corise – `http://corise.com/go/nazneen`. More details about her work can be found here `https://www.nazneenrajani.com/`.

# Towards a Foundation for AGI

**Gary Marcus**
Robust.AI, New York University (Emeritus)

**Saturday, December 10, 2022** - Room: **Hall B** - Time: **14:00-15:00**

**Abstract:** Large pretrained language models like GPT-3 and PaLM have generated enormous enthusiasm, and are capable of producing remarkably fluent language. But they have also been criticized on many grounds, and described as "stochastic parrots." Are they adequate as a basis for artificial general intelligence (AGI), and if not, what would a better foundation for general intelligence look like?

**Bio:** Gary Marcus is a leading voice in artificial intelligence. He is a scientist, best-selling author, and serial entrepreneur (Founder of Robust.AI and Geometric.AI, acquired by Uber). He is well-known for his challenges to contemporary AI, anticipating many of the current problems decades in advance, and for his research in human language development and cognitive neuroscience. An Emeritus Professor of Psychology and Neural Science at NYU, he is the author of five books, including, *The Algebraic Mind*, *Kluge*, *The Birth of the Mind*, and the *New York Times* Bestseller *Guitar Zero*. He has often contributed to *The New Yorker*, *Wired*, and *The New York Times*. His most recent book, *Rebooting AI*, with Ernest Davis, is one of *Forbes*'s 7 Must Read Books in AI.

# Towards a Responsible NLP: Walking the Walk

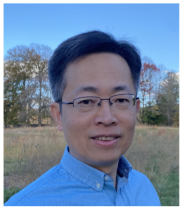
**Mona Diab**
Meta Responsible AI and The George Washington University

**Sunday, December 11, 2022** - Room: **Hall B** - Time: **14:00-15:00**

**Abstract:** In a world of racing to get the best systems on leaderboards, winning best shared tasks, building the largest LM, are we losing our soul as a scientific enterprise? Do we need to re-orient and re-pivot NLP? If so, what is needed to make this happen? Can we chart together a program where we ensure that science is the pivotal ingredient in CL/NLP? Could Responsible NLP be an avenue that could lead us back towards that goal? In this talk, in the spirit of Empirical NLP, I will explore some "practical" ideas around framing a Responsible NLP vision hoping to achieve a higher scientific standard for our field, addressing issues from the "how" we conduct our research and venturing into the "what" we work on and produce using tenets from responsible mindset perspective. I will pose more questions than answers. This is a call to action, an invitation to start a real global community conversation, hopefully engaging all stakeholders: academia, industry, government and civic society.

**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. Mona has served the community in several capacities: Elected President of SIGLEX and SIGSemitic, and she currently serves as the elected VP for ACL SIGDAT, the board supporting EMNLP conferences. She has delivered tutorials and organized numerous workshops and panels around Arabic processing, Responsible NLP, Code Switching, etc. She is a cofounder of CADIM (Consortium on Arabic Dialect Modeling, previously known as Columbia University Arabic Dialects Modeling Group), in 2005, which served as a world renowned reference point on Arabic Language Technologies. Moreover she helped establish two research trends in NLP, namely computational approaches to Code Switching and Semantic Textual Similarity. She is also a founding member of the *SEM conference, one of the top tier conferences in NLP. Mona has published more than 250 peer reviewed articles.

# Panel: "Careers in NLP"


**Saturday, December 10, 2022** - Room: **Hall B** - Time: **17:00-18:00**

The Careers in NLP Panel is a new feature of the EMNLP Industry Track. The panel is addressed to graduate students and junior researchers as well as their supervisors and mentors, although all EMNLP participants are welcome. The panelists will discuss the diversity of career paths in NLP: from more research-oriented NLP scientist roles to careers in product.

**Asli Celikyilmaz, Meta AI**
**Bio:** Asli Celikyilmaz is a Research Science Manager at FAIR Labs in Seattle. Formerly, she was Senior Principal Researcher at Microsoft Research (MSR) in Redmond, Washington. She is also an Affiliate Associate Member at the University of Washington. She has received a Ph.D. Degree in Information Science from University of Toronto, Canada, and later continued her Postdoc study at the Computer Science Department of the University of California, Berkeley. Her research interests are mainly in deep learning and natural language, specifically on language generation with long-term coherence, language understanding, language grounding with vision, and building intelligent agents for human-computer interaction She is serving as the co-Editor-in-Chief of the Transactions of the ACL (TACL) and as area editor on Open Journal of Signal Processing (OJSP) as Associate Editor. She has received several "best of" awards including Semantic Computing in 2009, and CVPR in 2019.

**Bing Xiang, Amazon AWS**
**Bio:** Bing Xiang is currently a Director of Applied Science at Amazon Web Services, leading a global science organization in AWS AI Labs. He oversees the science work powering dozens of AWS AI services and products that leverage machine learning and deep learning for search, question answering, information extraction, program synthesis, recommendation, forecasting, anomaly detection, and business analytics. Before joining Amazon in 2017, he was a Principal Research Staff Member and Science Manager at IBM Watson Research Center, leading a research team developing algorithms for multiple NLP services. Prior to IBM, he worked at BBN Technologies as a key contributor to several DARPA projects on speech recognition, speech-to-speech translation, and machine translation. He has published over 100 papers and served as an Area Chair and Program Committee Member at top NLP conferences like ACL, NAACL and EMNLP. He holds a PhD degree from Cornell University and BS/MS degrees from Peking University.

**Hatem Haddad, iCompass**
**Bio:** Hatem Haddad is Co-Founder and CTO of iCompass. He received a doctorate in Computer Science and Information Systems from University Grenoble Alpes, France. He was a Postdoctoral Fellow at VTT Technical Research Center of Finland and at Norwegian University of Science and Technology. He occupied assistant professor positions at Grenoble Alpes university (France), at UAEU (EAU), at Sousse university (Tunisia), at Mevlana university (Turkey) and at ULB (Belgium). He worked for industrial corporations in R&D at VTT Technical Research Centre of Finland and Institute for Infocomm Research, Image

Processing and Applications Lab of Singapore. He was an invited researcher at Leibniz-Fachhochschule School of Business (Germany) and Polytechnic Institute of Coimbra (Portugal). His current research interests include Natural Language Processing, Machine Learning and Deep Learning. He is a Program Chair in various global conferences and serves as a reviewer for relevant journals and conferences in the Artificial Intelligence field.

5

# D&I Events and Initiatives

## Statement by the EMNLP 2022 Diversity & Inclusion Committee

Dear attendees of EMNLP 2022,

Greetings from the co-chairs of the Diversity and Inclusion (D&I) Committee. In this short statement, we would like to highlight the following items related to our efforts toward a more inclusive and diverse EMNLP 2022.

**Subsidies** We received 164 applications for D&I subsidies and awarded a total of $56K to 51 applicants. The funding for this came from the D&I sponsors Google Research, NYU Abu Dhabi, and Microsoft, as well as ACL.

**Anti-harassment policy** We want to remind participants of our anti-harassment policy, which can be found here: `https://2022.emnlp.org/participants/anti_harassment/`.

**Events**

- There will be two D&I lunches for affinity groups. One lunch will feature an invited talk by Prof. Hannah Elsisi (`https://www.hist.cam.ac.uk/people/dr-hannah-elsisi`) of NYU Abu Dhabi on "Intersectional Identities in the Arab World." This event is sponsored by Snap Inc.
- There will be several Birds of a Feather sessions. We thank the organizers of these sessions for facilitating a space where people sharing similar interests can come together during the conference. The table below shows more details about the sessions and schedule.

Wishing you a fun, safe, and productive time in Abu Dhabi,

Meriem Beloucif, Thamar Solorio, and Andreas Vlachos

| Session | Hosts | Location | Date, Time |
|---|---|---|---|
| SIGARAB Social | Salam Khalifa, Go Inoue, Bashar Alhafni | Capital Suite 8 | Dec. 7, 14:00 |
| Challenges in Hate Speech Detection for African Languages | Idris Abdulmumin, Shamsuddeen Hassan Mohammed, Nedjma Ousidhoum | Capital Suite 8 & Zoom | Dec. 7, 16:00 |
| Biomedical NLP | Kirk Roberts | Zoom | Dec. 7, 18:00 |
| Language Models | Mark Dredze | Zoom | Dec. 7, 20:30 |
| Ethics in NLP | Fatemehsadat Mireshghallah, Luciana Benotti, Patrick Blackburn | Capital Suite 7 | Dec. 8, 14:00 |
| NLP in the time of social media: challenges and applications | Jose Camacho-Collados, Luis Espinosa-Anke | Capital Suite 7 | Dec. 8, 16:00 |
| LatinX in AI Social at EMNLP 2022 | Diana Galvan | Zoom | Dec. 8, 19:00 |
| Code-Switching and Multilinguality NLP | Genta Winata, Marina Zhukova, Sudipta Kar | Capital Suite 7 & Zoom | Dec. 10, 11:00 |
| Expanding horizons for female researchers in AI/NLP | Hanan Salam, Vinutha Magal Shreenath | Capital Suite 7 & Zoom | Dec. 10, 15:30 |
| We need to talk about random seeds | Steven Bethard, Xueqing Liu | Zoom | Dec. 10, 19:00 |

Birds of a Feather Sessions at EMNLP 2022

6

# Tutorials: Wednesday, December 7, 2022

## Overview

| | | |
|---|---|---|
| 07:30 - 16:30 | ***Day 1 Registration*** | |
| 08:00 - 08:45 | ***Extra Q&A 1 - Morning Tutorials*** | |
| | *Tutorial 1 – Meaning Representations for Natural Languages: Design, Models and Applications*<br>Jeffrey Flanigan, Ishan Jindal, Yunyao Li, Tim O'Gorman, Martha Palmer, Nianwen Xue | *Capital Suite 7* |
| | *Tutorial 2 – Arabic Natural Language Processing*<br>Nizar Habash | *Capital Suite 9* |
| 09:00 - 12:30 | ***Morning Tutorials*** | |
| | *Tutorial 1 – Meaning Representations for Natural Languages: Design, Models and Applications*<br>Jeffrey Flanigan, Ishan Jindal, Yunyao Li, Tim O'Gorman, Martha Palmer, Nianwen Xue | *Capital Suite 7* |
| | *Tutorial 2 – Arabic Natural Language Processing*<br>Nizar Habash | *Capital Suite 9* |
| 12:30 - 13:00 | ***Extra Q&A 2 - Morning Tutorials*** | |
| | *Tutorial 1 – Meaning Representations for Natural Languages: Design, Models and Applications*<br>Jeffrey Flanigan, Ishan Jindal, Yunyao Li, Tim O'Gorman, Martha Palmer, Nianwen Xue | *Capital Suite 7* |
| | *Tutorial 2 – Arabic Natural Language Processing*<br>Nizar Habash | *Capital Suite 9* |
| 13:30 - 14:00 | ***Extra Q&A 1 - Afternoon Tutorials*** | |
| | *Tutorial 1 – Meaning Representations for Natural Languages: Design, Models and Applications* | *Capital Suite 7* |

| | | |
|---|---|---|
| | Jeffrey Flanigan, Ishan Jindal, Yunyao Li, Tim O'Gorman, Martha Palmer, Nianwen Xue | |
| | *Tutorial 3 – Emergent Language-Based Coordination In Deep Multi-Agent Systems*<br>Marco Baroni, Roberto Dessì, Angeliki Lazaridou | *Capital Suite 9* |
| 14:00 - 17:30 | ***Afternoon tutorials*** | |
| | *Tutorial 1 – Meaning Representations for Natural Languages: Design, Models and Applications*<br>Jeffrey Flanigan, Ishan Jindal, Yunyao Li, Tim O'Gorman, Martha Palmer, Nianwen Xue | *Capital Suite 7* |
| | *Tutorial 3 – Emergent Language-Based Coordination In Deep Multi-Agent Systems*<br>Marco Baroni, Roberto Dessì, Angeliki Lazaridou | *Capital Suite 9* |
| 18:00 - 18:45 | ***Extra Q&A 2 - Afternoon Tutorials*** | |
| | *Tutorial 1 – Meaning Representations for Natural Languages: Design, Models and Applications*<br>Jeffrey Flanigan, Ishan Jindal, Yunyao Li, Tim O'Gorman, Martha Palmer, Nianwen Xue | *Capital Suite 7* |
| | *Tutorial 3 – Emergent Language-Based Coordination In Deep Multi-Agent Systems*<br>Marco Baroni, Roberto Dessì, Angeliki Lazaridou | *Capital Suite 9* |

7

# Tutorials: Thursday, December 8, 2022

## Overview

| | | |
|---|---|---|
| 08:00 - 16:30 | ***Day 2 Registration*** | |
| 08:00 - 08:45 | ***Extra Q&A 1 - Morning Tutorials*** | |
| | *Tutorial 4 – CausalNLP Tutorial: An Introduction to Causality for Natural Language Processing*<br>Zhijing Jin, Amir Feder, Kun Zhang | *Capital Suite 9* |
| | *Tutorial 5 – Modular and Parameter-Efficient Fine-Tuning for NLP Models*<br>Sebastian Ruder, Jonas Pfeiffer, Ivan Vulić | *Capital Suite 7* |
| 09:00 - 12:30 | ***Morning Tutorials*** | |
| | *Tutorial 4 – CausalNLP Tutorial: An Introduction to Causality for Natural Language Processing*<br>Zhijing Jin, Amir Feder, Kun Zhang | *Capital Suite 9* |
| | *Tutorial 5 – Modular and Parameter-Efficient Fine-Tuning for NLP Models*<br>Sebastian Ruder, Jonas Pfeiffer, Ivan Vulić | *Capital Suite 7* |
| 12:30 - 13:00 | ***Extra Q&A 2 - Morning Tutorials*** | |
| | *Tutorial 4 – CausalNLP Tutorial: An Introduction to Causality for Natural Language Processing*<br>Zhijing Jin, Amir Feder, Kun Zhang | *Capital Suite 9* |
| | *Tutorial 5 – Modular and Parameter-Efficient Fine-Tuning for NLP Models*<br>Sebastian Ruder, Jonas Pfeiffer, Ivan Vulić | *Capital Suite 7* |
| 13:30 - 14:00 | ***Extra Q&A 1 - Afternoon Tutorials*** | |
| | *Tutorial 6 – Non-Autoregressive Models for Fast Sequence Generation* | *Capital Suite 9* |

| | | |
|---|---|---|
| | Yang Feng, Chenze Shao | |
| 14:00 - 17:30 | ***Afternoon tutorials*** | |
| | *Tutorial 6 – Non-Autoregressive Models for Fast Sequence Generation*<br>Yang Feng, Chenze Shao | *Capital Suite 9* |
| 18:00 - 18:45 | ***Extra Q&A 2 - Afternoon Tutorials*** | |
| | *Tutorial 6 – Non-Autoregressive Models for Fast Sequence Generation*<br>Yang Feng, Chenze Shao | *Capital Suite 9* |

Welcome to the Tutorials Session of EMNLP 2022

The EMNLP 2022 tutorials session provides an in depth coverage of a variety of topics reflecting recent advances in Natural Language Processing methods and applications, presented by experts from academia and ranging from introductory to cutting-edge.

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, NAACL, COLING and EMNLP. A review committee consisting of ACL, NAACL, COLING and EMNLP tutorial chairs as well as 23 external reviewers (see Program Committee for the full list), was formed. The committee followed a review process that ensured that each of the 47 submitted tutorial proposals, received 3 reviews. The selection criteria included clarity and preparedness, novelty or timely character of the topic, instructors' experience, likely audience interest, open access of the tutorial instructional material, and diversity and inclusion.
The six tutorials selected for EMNLP include 2 introductory tutorials and 4 cutting-edge tutorials. The two introductory tutorials address Arabic natural language processing (T2) and causal inference for natural language processing(T4) while the cutting-edge tutorials address meaning representations for natural languages (T1), emergent language-based coordination in deep Multi-Agent Systems (T3), modular and parameter-efficient fine-tuning for NLP models (T5), and non-autoregressive models for fast sequence generation (T6).

We would like to thank the ACL, NAACL, and COLING tutorial chairs and the 23 external reviewers for their effective collaboration and their efforts to ensure a smooth selection process as well as their invaluable assistance in the decision process. We would also like to thank EMNLP's general chair Noah Smith for his readiness to extend support whenever requested. We are very grateful for tutorial organizers for their valuable contributions.

As has been the case last year, tutorial presentations will be a mixture of online, on-site and hybrid presentations. We hope you all benefit from and enjoy the tutorial program at EMNLP 2022!

EMNLP 2022 Tutorial Co-chairs
Samhaa R. El-Beltagy
Xipeng Qiu

# T1 - Meaning Representations for Natural Languages: Design, Models and Applications

Jeffrey Flanigan, Tim O'Gorman, Ishal Jindal, Yunyao Li, Martha Palmer, Nianwen Xue

Wednesday, December 7, 2022 - 9:00-17:30. Extra Q&A Sessions: 8:00-8:45, 12:30-13:00, 13:30-14:00, 18:00-18:45 (Capital Suite 7)

This tutorial reviews the design of common meaning representations, SoTA models for predicting meaning representations, and the applications of meaning representations in a wide range of downstream NLP tasks and real-world applications. Reporting by a diverse team of NLP researchers from academia and industry with extensive experience in designing, building and using meaning representations, our tutorial has three components: (1) an introduction to common meaning representations, including basic concepts and design challenges; (2) a review of SoTA methods on building models for meaning representations; and (3) an overview of applications of meaning representations in downstream NLP tasks and real-world applications. We will also present qualitative comparisons of common meaning representations and a quantitative study on how their differences impact model performance. Finally, we will share best practices in choosing the right meaning representation for downstream tasks.

---

**Jeffrey Flanigan**, University of California Santa Cruz
**Website:** `https://jflanigan.github.io/`
Jeffrey Flanigan is an Assistant Professor in the Computer Science and Engineering Department at University of California Santa Cruz. He research interests are in semantic parsing and generation, with a focus on AMR, and using semantic representations in downstream applications such as summarization and machine translation. Previously he has given a tutorial in AMR at NAACL 2015.

**Tim O'Gorman**, Thorn
**Website:** `https://timjogorman.github.io/`
Tim O'Gorman is a Senior Research Scientist at Thorn. He was involved in AMR 2.0 and 3.0 annotations, the Multi-sentence AMR corpus, and updates to PropBank. He co-organized the CoNLL'19 and '20 Meaning Representation Parsing shared task. His interests are in the extensions of meaning representations to cross-sentence phenomena.

**Ishan Jindal**, IBM
**Website:** `https://ijindal.github.io/`
Ishan Jindal is a Research Staff Member with IBM Research - Almaden. His research interest lies at the intersection of machine learning and NLP, primarily in semantic parsing and model analysis for enterprise use cases. He regularly publishes papers at ML and NLP conferences.

**Yunyao Li**, Apple Knowledge Platform
**Website:** https://yunyaoli.github.io/
Yunyao Li is the Head of Machine Learning, Apple Knowledge Platform. Until very recently, she was a Distinguished Research Staff Member and Senior Research Manager at IBM Research - Almaden where she built and managed the Scalable Knowledge Intelligence department. She is particularly known for her work in scalable NLP, enterprise search, and database usability. She has built systems, developed solutions, and delivered core technologies to over 20 IBM products under brands such as Watson, InfoSphere, and Cognos. She has published over 80 articles and a book. She was a IBM Master Inventor, with over 50 patents filed/granted. Her technical contributions have been recognized by prestigious awards within and outside of IBM on regular basis. She is an ACM Distinguished Member. She was a member of the inaugural New Voices program of the American National Academies (1 out of 18 selected nationwide) and represented US young scientists at World Laureates Forum Young Scientists Forum in 2019 (1 of 4 selected nationwide). Yunyao obtained her Ph.D degree in Computer Science & Engineering and dual-degrees of M.S.E in Computer Science & Engineering and M.S in Information from the University of Michigan. She went to college at Tsinghua University, Beijing, China, and graduated with dual-degrees of B.E in Automation and B.S in Economics.

**Martha Palmer**, University of Colorado
**Website:** https://www.colorado.edu/faculty/palmer-martha/
Martha Palmer is the Helen & Hubert Croft Professor of Engineering in the Computer Science Department, and Arts & Sciences Professor of Distinction for Linguistics, at the University of Colorado, with over 300 peer-reviewed publications. Her research is focused on capturing elements of the meanings of words that can comprise automatic representations of complex sentences and documents in many languages. She is a co-Director of CLEAR, an ACL Fellow, and an AAAI Fellow.

**Nianwen Xue**, Brandeis University
**Website:** https://www.cs.brandeis.edu/~xuen/
Nianwen Xue is a Professor in the Computer Science Department and the Language & Linguistics Program at Brandeis University. His core research interests include developing linguistic corpora annotated with syntactic, semantic, and discourse structures, as well as machine learning approaches to syntactic, semantic and discourse parsing. He is an action editor for Computational Linguistics.

# T2 - Arabic Natural Language Processing

Nizar Habash

Wednesday, December 7, 2022 - 9:00-12:30. Extra Q&A Sessions: 8:00-8:45 and 12:30-13:00 (Capital Suite 9)

The Arabic language continues to be the focus of an increasing number of projects in natural language processing (NLP) and computational linguistics (CL). This tutorial provides NLP/CL system developers and researchers (computer scientists and linguists alike) with the necessary background information for working with Arabic in its various forms: Classical, Modern Standard and Dialectal. We discuss various Arabic linguistic phenomena and review the state-of-the-art in Arabic processing from enabling technologies and resources, to common tasks and applications. The tutorial will explain important concepts, common wisdom, and common pitfalls in Arabic processing. Given the wide range of possible issues, we invite tutorial attendees to bring up interesting challenges and problems they are working on to discuss during the tutorial.

---

**Nizar Habash**, New York University Abu Dhabi
**Website:** `www.nizarhabash.com`
Nizar Habash is a Professor of Computer Science at New York University Abu Dhabi (NYUAD). He is also the director of the Computational Approaches to Modeling Language (CAMeL) Lab. Professor Habash specializes in natural language processing and computational linguistics. Before joining NYUAD in 2014, he was a research scientist at Columbia University's Center for Computational Learning Systems. He received his PhD in Computer Science from the University of Maryland College Park in 2003. He has two bachelors degrees, one in Computer Engineering and one in Linguistics and Languages. His research includes extensive work on machine translation, morphological analysis, and computational modeling of Arabic and its dialects. Professor Habash has been a principal investigator or co-investigator on over 25 research grants. And he has over 250 publications including a book entitled "Introduction to Arabic Natural Language Processing" (Habash, 2010).

# T3 - Emergent Language-Based Coordination In Deep Multi-Agent Systems



Marco Baroni, Roberto Dessì, Angeliki Lazaridou

Wednesday, December 7, 2022 - 14:00-17:30. Extra Q&A Sessions: 13:30-14:00 and 18:00-18:45 (Capital Suite 9)

Large pre-trained deep networks are the standard building blocks of modern AI applications. This raises fundamental questions about how to control their behaviour and how to make them efficiently interact with each other. Deep net emergent communication tackles these challenges by studying how to induce communication protocols between neural network agents, and how to include humans in the communication loop. Traditionally, this research had focussed on relatively small-scale experiments where two networks had to develop a discrete code from scratch for referential communication. However, with the rise of large pre-trained language models that can work well on many tasks, the emphasis is now shifting on how to let these models interact through a language-like channel to engage in more complex behaviors. By reviewing several representative papers, we will provide an introduction to deep net emergent communication, we will cover various central topics from the present and recent past, as well as discussing current shortcomings and suggest future directions. The presentation is complemented by a hands-on section where participants will implement and analyze two emergent communications setups from the literature. The tutorial should be of interest to researchers wanting to develop more flexible AI systems, but also to cognitive scientists and linguists interested in the evolution of communication systems.

---

**Marco Baroni**, Universitat Pompeu Fabra
**Website:** `https://marcobaroni.org/`
Marco Baroni is ICREA research professor at Universitat Pompeu Fabra. He co-authored one of the earliest and most influential papers on emergent communication among deep net agents (Lazaridou et al., 2017) as well as a recent survey of the area (Lazaridou and Baroni, 2020). Marco has extensive teaching experience, including interdisciplinary classes aimed at computer scientists, linguists and cognitive scientists, and lectures and tutorials in international venues such as ESSLLI, ACL and the CIFAR Deep Learning Summer School (where he presented an introduction to deep net emergent communication). He was recently awarded an ERC Advanced Grant to work on emergent communication.

**Roberto Dessì**, FAIR Paris and Universitat Pompeu Fabra
**Website:** `https://robertodessi.github.io/`
Roberto Dessì is a 3rd-year PhD student at Facebook AI Research and Universitat Pompeu Fabra. His work focuses on scaling up emergent communication research, including a paper on the topic appearing at NeurIPS 2021. Roberto was a co-organizer of the last two Emergent Communication workshops and is currently the maintainer of the EGG toolkit for emergent communication simulations.

**Angeliki Lazaridou**, DeepMind
**Website:** `http://angelikilazaridou.github.io/`
Angeliki Lazaridou is staff research scientist at DeepMind. Angeliki co-authored one of the earliest and most influential papers on emergent communication among deep net agents (Lazaridou et al., 2017) as well as a recent survey of the area (Lazaridou and Baroni, 2020). Angeliki's work in the area was recognized with a 2019 ICML best-paper mention (Jaques et al., 2019). She co-initiated the Emergent Communication NeurIPS Workshop series (which ran successfully for 6 years).

# T4 - CausalNLP Tutorial: An Introduction to Causality for Natural Language Processing



Zhijing Jin, Amir Feder, Kun Zhang

Thursday, December 8, 2022 - 9:00-12:30. Extra Q&A Sessions: 8:00-8:45 and 12:30-13:00 (Capital Suite 9)

Causal inference is becoming an increasingly important topic in deep learning, with the potential to help with critical deep learning problems such as model robustness, interpretability, and fairness. In addition, causality is naturally widely used in various disciplines of science, to discover causal relationships among variables and estimate causal effects of interest. In this tutorial, we introduce the fundamentals of causal discovery and causal effect estimation to the natural language processing (NLP) audience, provide an overview of causal perspectives to NLP problems, and aim to inspire novel approaches to NLP further. This tutorial is inclusive to a variety of audiences and is expected to facilitate the community's developments in formulating and addressing new, important NLP problems in light of emerging causal principles and methodologies.

---

**Zhijing Jin**, Max Planck Institute and ETH Zürich
**Website:** `https://zhijing-jin.com/`
Zhijing Jin (she/her) is a PhD at Max Planck Institute and ETH Zürich supervised by Prof Bernhard Schölkopf. Her research aims to (1) improve NLP models by connecting NLP with causal inference (Jin et al., 2021c,b; Ni et al., 2022), and (2) expand the impact of NLP by promoting NLP for social good (Jin et al., 2021a; Field et al., 2021; Gonzalez et al., 2022). She has published at many NLP and AI venues (e.g., AAAI, ACL, EMNLP, NAACL, COLING, AISTATS), and NLP for healthcare venues (e.g., AAHPM, JPSM). To foster the causality research community, she serves as the Publications Chair for the 1st conference on Causal Learning and Reasoning (CLeaR) (Schölkopf et al., 2022). She is also actively involved in AI for social good, as the organizer of NLP for Positive Impact Workshop at ACL 2021 (Field et al., 2021) and EMNLP 2022, and RobustML workshop at ICLR 2021. To support the NLP research community, she organizes the ACL Year-Round Mentorship program from 2021.

**Amir Feder**, Columbia University
**Website:** `https://amirfeder.github.io/`
Amir Feder (he/him) is a postdoc at Columbia University, working with Prof David Blei. Amir develops methods that integrate causality into natural language processing to generate more robust and interpretable models. He is also interested in investigating and developing linguistically informed algorithms for predicting and understanding human behavior. Amir is currently also a visiting researcher (part time) at Google Research's Medical Brain Team, where he works on methods that leverage causal methodology for medical language models. He is a co-organizer of the First Workshop on Causal Inference and NLP (CI+NLP) at EMNLP 2021 (Feder et al., 2021a).

**Kun Zhang**, Carnegie Mellon University and MBZUAI
**Website:** `https://www.cmu.edu/dietrich/philosophy/people/faculty/zhang.html`
Kun Zhang (he/him) is an associate professor at Carnegie Mellon University and MBZUAI. His research interests lie in causal discovery and causality-based learning. He develops methods for automated causal discovery from various kinds of data, investigates learning problems including transfer learning and deep learning from a causal view, and studies philosophical foundations of causation and machine learning. He co-authored a best student paper for the Conference on Uncertainty in Artificial Intelligence (UAI) and a best finalist paper for the Conference on Computer Vision and Pattern Recognition (CVPR), and received the best benchmark award of the 2nd causality challenge. He has taken essential roles at many events of causal inference, including the general and program co-chair of the 1st Conference on Causal Learning and Reasoning (CLeaR 2022), program co-chair of the UAI 2022, co-organizer of the 9th Causal Inference Workshop at UAI 2021, co-organizer of NeurIPS 2020 Workshop on Causal Discovery and Causality-Inspired Machine Learning 2020, co-editor of a number of journal special issues on causality, and many others.

# T5 - Modular and Parameter-Efficient Fine-Tuning for NLP Models



Sebastian Ruder, Jonas Pfeiffer, Ivan Vulić

Thursday, December 8, 2022 - 9:00-12:30. Extra Q&A Sessions: 8:00-8:45 and 12:30-13:00 (Capital Suite 7)

State-of-the-art language models in NLP perform best when fine-tuned even on small datasets, but due to their increasing size, finetuning and downstream usage have become extremely compute-intensive. Being able to efficiently and effectively fine-tune the largest pretrained models is thus key in order to reap the benefits of the latest advances in NLP. In this tutorial, we provide a comprehensive overview of parameter-efficient fine-tuning methods. We highlight their similarities and differences by presenting them in a unified view. We explore the benefits and usage scenarios of a neglected property of such parameter-efficient models — modularity — such as composition of modules to deal with previously unseen data conditions. We finally highlight how both properties — parameter efficiency and modularity — can be useful in the real-world setting of adapting pre-trained models to under-represented languages and domains with scarce annotated data for several downstream applications.

---

**Sebastian Ruder**, Google Research
**Website:** `http://ruder.io`
Sebastian is a research scientist at Google Research where he works on transfer and cross-lingual learning and on parameter-efficient models. He was the Program Co-Chair for EurNLP 2019 and has co-organized the 4th Workshop on Representation Learning for NLP at ACL 2019 and the First Workshop on Multilingual Representation Learning at EMNLP 2021 and 2022. He has taught tutorials on "Transfer learning in natural language processing", "Unsupervised Cross-lingual Representation Learning", and "Multi-domain Multilingual Question Answering" at NAACL 2019, ACL 2019, and EMNLP 2021 respectively. He has also co-organized and taught at the NLP Session at the Deep Learning Indaba 2018, 2019, and 2022.

**Jonas Pfeiffer**, Google Research
**Website:** `https://pfeiffer.ai`
Jonas is a research scientist at Google Research. He is interested in modular and compositional representation learning in multi-task, multilingual, and multi-modal contexts. Jonas has received the IBM PhD Research Fellowship award in 2020. He has given invited talks in academia (e.g. University of Cambridge, ETH, EPFL, NYU), industry (e.g. Facebook AI Research, IBM Research), as well as at Machine Learning Summer/Winter Schools (e.g. Lisbon ML Summer School (LxMLS) 2021, Advanced Language Processing Winter School (ALPS) 2022).

**Ivan Vulić**, University of Cambridge and PolyAI

**Website:** `https://sites.google.com/site/ivanvulic/`
Ivan is a Principal Research Associate and a Royal Society University Research Fellow in the Language Technology Lab at the University of Cambridge, and a Senior Scientist at PolyAI. His research interests are in multilingual and multimodal representation learning, and transfer learning for low-resource languages and applications such as task-oriented dialogue systems. He has extensive experience giving invited and keynote talks, and co-organising tutorials (e.g., ECIR 2013, WSDM 2014, EMNLP 2017, NAACL-HLT 2018, ESSLLI 2018, ACL 2019, 2 tutorials at EMNLP 2019, AILC Lectures 2021, ACL 2022) and workshops in areas relevant to the tutorial proposal (e.g., VL'15, SIGTYP 2019-2021, DeeLIO 2020-2022, RepL4NLP 2021, MML 2022, publication chair of ACL 2019, program chair of *SEM 2021, tutorial co-chair of EMNLP 2021).

# T6 - Non-Autoregressive Models for Fast Sequence Generation


Yang Feng, Chenze Shao

Thursday, December 8, 2022 - 14:00-17:30. Extra Q&A Sessions: 13:30-14:00 and 18:00-18:45 (Capital Suite 9)

Autoregressive (AR) models have achieved great success in various sequence generation tasks. However, AR models can only generate target sequence word-by-word due to the AR mechanism and hence suffer from slow inference. Recently, non-autoregressive (NAR) models, which generate all the tokens in parallel by removing the sequential dependencies within the target sequence, have received increasing attention in sequence generation tasks such as neural machine translation (NMT), automatic speech recognition (ASR), and text to speech (TTS). In this tutorial, we will provide a comprehensive introduction to non-autoregressive sequence generation.

---

**Yang Feng**, Institute of Computing Technology, Chinese Academy of Sciences
**Website:** `https://yangfengyf.github.io/`
Yang Feng is a professor in Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS). She got her PhD degree in ICT/CAS and then worked in University of Sheffield and Information Sciences Institute, University of Southern California, and now leads the natural language processing group in ICT/CAS. Her research interests are natural language process, mainly focusing on machine translation and dialogue. She was the recipient of the Best Long Paper Award of ACL 2019. She served as a senior area chair of EMNLP 2021 and area chairs of ACL, EMNLP, COLING etc., and she is serving as an Action Editor of ACL Roling Review and an editorial board member of the Northern European Journal of Language Technology. She has given a tutorial in the 10th CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC2021) and has been invited to give talks in NLPCC, CCL(China National Conference on Computational Linguistics) etc.

**Chenze Shao**, Institute of Computing Technology, Chinese Academy of Sciences
Chenze Shao is a fifth-year PhD student in Institute of Computing Technology, Chinese Academy of Sciences. His research interests are natural language processing and neural machine translation. His recent research topic is non-autoregressive (NAR) sequence generation. He has published papers on NAR generation in CL, ACL, EMNLP, NAACL, AAAI and NeurIPS.

8

# Main Conference

## Main Conference Program (Overview)

### Main Conference Program (Overview): Day 1

8:00-16:30 *Registration (The Link)*

9:00-9:30 *Plenary: Opening Address (Hall B)*

9:00-10:30 **Session 1** *Plenary: Opening & Keynote 1 - Neil Cohn (Hall B)*

10:30-11:00 Coffee Break

11:00-12:30 **Session 2**

| | |
|---|---|
| **Natural Language Generation 1**<br>*Hall B* | **Resources and Evaluation 1**<br>*Hall A, Room A* |
| **Semantics**<br>*Hall A, Room B* | **Summarization**<br>*Hall A, Room C* |
| **Industry 1**<br>*Hall A, Room D* | **NLP Applications 1**<br>*Collaboratorium* |
| **Poster Sessions 1 & 2**<br>*Atrium* | **Demo Session 1**<br>*Link Admin* |

12:30-14:00 Lunch Break & *D&I Lunch (Capital Suite 7)*

14:00-15:30 **Session 3**

| | |
|---|---|
| **Language Modeling and Analysis of Language Models**<br>*Hall B* | **Sentiment, Stylistic Analysis, Argument Mining & Discourse**<br>*Hall A, Room A* |
| **Speech, Vision, Robotics, Multimodal Grounding 1 & CL**<br>*Hall A, Room B* | **Question Answering 1**<br>*Hall A, Room C* |
| **CL & TACL 1**<br>*Hall A, Room D* | **Ethics & Computational Social Science and Cultural Analytics**<br>*Collaboratorium* |
| **Poster Sessions 3 & 4**<br>*Atrium* | **Demo Session 2**<br>*Link Admin* |

15:30-16:00 Coffee Break

16:00-17:30 **Session 4**

| | |
|---|---|
| **Virtual Portal 1**<br>*Hall A, Room A* | **Virtual Portal 2**<br>*Hall A, Room B* |
| **Virtual Portal 3**<br>*Hall A, Room C* | **Virtual Portal 4**<br>*Hall A, Room D* |
| **Virtual Portal 5**<br>*Hall B* | **Virtual Portal 6**<br>*Collaboratorium* |
| **Poster Sessions 5 & 6**<br>*Atrium* | |

17:30-18:30 **Session 5** *Plenary: Industry Track Keynote - Nazneen Rajani (Hall B)*

# Main Conference Program (Overview): Day 2

8:30-16:30 *Registration (The Link)*

9:00-10:30 **Session 6**

| | |
|---|---|
| **Dialog and Interactive Systems 1**<br>*Hall A, Room A* | **Multilinguality**<br>*Hall A, Room B* |
| **Natural Language Generation 2 & TACL**<br>*Hall A, Room C* | **Efficient Methods for NLP**<br>*Hall A, Room D* |
| **Information Retrieval and Text Mining**<br>*Hall B* | **Industry 2**<br>*Collaboratorium* |
| **Poster Sessions 7 & 8**<br>*Atrium* | **Demo Session 3**<br>*Atrium* |

10:30-11:00 Coffee Break

11:00-12:30 **Session 7**

| | |
|---|---|
| **Interpretability, Interactivity, and Analysis of Models for NLP 1**<br>*Hall A, Room A* | **Machine Learning for NLP**<br>*Hall A, Room B* |
| **Resources and Evaluation 2**<br>*Hall A, Room C* | **Theme Track & CL & Short Papers**<br>*Hall A, Room D* |
| **Information Extraction 1**<br>*Hall B* | **CL & TACL 2**<br>*Collaboratorium* |
| **Poster Sessions 9 & 10**<br>*Atrium* | **Demo Session 4**<br>*Atrium* |
| **Birds of a Feather**<br>*Capital Suite 7, Zoom* | |

12:30-13:15 Lunch Break

13:15-14:00 *Plenary: Business Meeting (Hall B)*

14:00-15:00 **Session 8** *Plenary: Keynote 2 - Gary Marcus (Hall B)*

15:00-15:30 Coffee Break

15:30-17:00 **Session 9**

| | |
|---|---|
| **Virtual Portal 7**<br>*Hall A, Room A* | **Virtual Portal 8**<br>*Hall A, Room B* |
| **Virtual Portal 9**<br>*Hall A, Room C* | **Virtual Portal 10**<br>*Hall A, Room D* |
| **Virtual Portal 11**<br>*Hall B* | **Virtual Portal 12**<br>*Collaboratorium* |
| **Poster Sessions 11 & 12**<br>*Atrium* | **Birds of a Feather**<br>*Capital Suite 7, Zoom* |

17:00-18:00 **Session 10** *Plenary: Industry Track Panel (Hall B)*

18:30 *Buses load for Social Event*

19:00-22:00 *Social Event (Ritz Carlton Abu Dhabi Grand Lawn)*

22:00-22:30 *Buses return from Social Event*

# Main Conference Program (Overview): Day 3

8:30-16:30 *Registration (The Link)*

9:00-10:30 **Session 11**

| | |
|---|---|
| **Machine Translation**<br>*Hall A, Room A* | **Commonsense Reasoning**<br>*Hall A, Room B* |
| **Interpretability, Interactivity, and Analysis of Models for NLP 2**<br>*Hall A, Room C* | **NLP Applications 2 & TACL**<br>*Hall A, Room D* |
| **Unsupervised and Weakly Supervised Methods**<br>*Hall B* | **Industry 3**<br>*Collaboratorium* |
| **Poster Sessions 13 & 14**<br>*Atrium* | **Demo Session 5**<br>*Atrium* |

10:30-11:00 Coffee Break

11:00-12:30 **Session 12**

| | |
|---|---|
| **Question Answering 2**<br>*Hall A, Room A* | **Morphology, Syntax, Linguistics, Psycholinguistics & TACL**<br>*Hall A, Room B* |
| **Dialog and Interactive Systems 2**<br>*Hall A, Room C* | **Speech, Vision, Robotics, Multimodal Grounding 2 & TACL**<br>*Hall A, Room D* |
| **Information Extraction 2**<br>*Hall B* | **CL & TACL 3**<br>*Collaboratorium* |
| **Poster Sessions 15 & 16**<br>*Atrium* | **Demo Session 6**<br>*Atrium* |

12:30-14:00 Lunch Break & *D&I Lunch (Capital Suite 7)*

14:00-15:00 **Session 13** *Plenary: Keynote 3 - Mona Diab (Hall B)*

15:00-15:30 Coffee Break

15:30-17:00 **Session 14**

| | |
|---|---|
| **Virtual Portal 13**<br>*Hall A, Room A* | **Virtual Portal 14**<br>*Hall A, Room B* |
| **Virtual Portal 15**<br>*Hall A, Room C* | **Virtual Portal 16**<br>*Hall A, Room D* |
| **Virtual Portal 17**<br>*Hall B* | **Virtual Portal 18**<br>*Collaboratorium* |
| **Poster Sessions 17 & 18**<br>*Atrium* | |

17:00-18:15 **Session 15** *Plenary: Best Papers & Closing Session (Hall B)*

# Main Conference: Friday, December 9, 2022

## Session 2 - 11:00-12:30

### Natural Language Generation 1

11:00-12:30 (Hall B)

**RankGen: Improving Text Generation with Large Ranking Models**
*Kalpesh Krishna, Yapei Chang, John Wieting and Mohit Iyyer* 11:00-11:15 (Hall B)
Given an input sequence (or prefix), modern language models often assign high probabilities to output sequences that are repetitive, incoherent, or irrelevant to the prefix; as such, model-generated text also contains such artifacts. To address these issues we present RankGen, a 1.2B parameter encoder model for English that scores model generations given a prefix. RankGen can be flexibly incorporated as a scoring function in beam search and used to decode from any pretrained language model. We train RankGen using large-scale contrastive learning to map a prefix close to the ground-truth sequence that follows it and far away from two types of negatives: (1) random sequences from the same document as the prefix, and (2) sequences generated from a large language model conditioned on the prefix. Experiments across four different language models (345M-11B parameters) and two domains show that RankGen significantly outperforms decoding algorithms like nucleus, top-k, and typical sampling on both automatic metrics (85.0 vs 77.3 MAUVE) as well as human evaluations with English writers (74.5% human preference over nucleus sampling). Analysis reveals that RankGen outputs are more relevant to the prefix and improve continuity and coherence compared to baselines. We release our model checkpoints, code, and human preference data with explanations to facilitate future research.

**Linearizing Transformer with Key-Value Memory**
*Yizhe Zhang and Deng Cai* 11:15-11:30 (Hall B)
Efficient transformer variants with linear time complexity have been developed to mitigate the quadratic computational overhead of the vanilla transformer. Among them are low-rank projection methods such as Linformer and kernel-based Transformers. Despite their unique merits, they usually suffer from a performance drop comparing with the vanilla transformer on many sequence generation tasks, and often fail to obtain computation gain when the generation is short. We propose Memsizer, an approach towards closing the performance gap while improving the efficiency even with short generation. It projects the source sequences into lower dimension representations like Linformer, while enjoying efficient recurrent-style incremental computation similar to kernel-based transformers. This yields linear computation time and constant memory complexity at inference time. Memsizer also employs a lightweight multi-head mechanism which renders the computation as light as a single-head model. We demonstrate that Memsizer provides an improved balance between efficiency and accuracy over the vanilla transformer and other efficient transformer variants in three typical sequence generation tasks, including machine translation, abstractive text summarization, and language modeling.

**A Unified Encoder-Decoder Framework with Entity Memory**
*Zhihan Zhang, Wenhao Yu, Chenguang Zhu and Meng Jiang* 11:30-11:45 (Hall B)
Entities, as important carriers of real-world knowledge, play a key role in many NLP tasks. We focus on incorporating entity knowledge into an encoder-decoder framework for informative text generation. Existing approaches tried to index, retrieve, and read external documents as evidence, but they suffered from a large computational overhead. In this work, we propose an encoder-decoder framework with an entity memory, namely EDMem. The entity knowledge is stored in the memory as latent representations, and the memory is pre-trained on Wikipedia along with encoder-decoder parameters. To precisely generate entity names, we design three decoding methods to constrain entity generation by linking entities in the memory. EDMem is a unified framework that can be used on various entity-intensive question answering and generation tasks. Extensive experimental results show that EDMem outperforms both memory-based auto-encoder models and non-memory encoder-decoder models.

**A Distributional Lens for Multi-Aspect Controllable Text Generation**
*Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong and Bing Qin* 11:45-12:00 (Hall B)
Multi-aspect controllable text generation is a more challenging and practical task than single-aspect control. Existing methods achieve complex multi-aspect control by fusing multiple controllers learned from single-aspect, but suffer from attribute degeneration caused by the mutual interference of these controllers. To address this, we provide observations on attribute fusion from a distributional perspective and propose to directly search for the intersection areas of multiple attribute distributions as their combination for generation. Our method first estimates the attribute space with an autoencoder structure. Afterward, we iteratively approach the intersections by jointly minimizing distances to points representing different attributes. Finally, we map them to attribute-relevant sentences with a prefix-tuning-based decoder. Experiments on the three-aspect control task, including sentiment, topic, and detoxification aspects, reveal that our method outperforms several strong baselines on attribute relevance and text quality and achieves the SOTA. Further analysis also supplies some explanatory support for the effectiveness of our approach.

**ELMER: A Non-Autoregressive Pre-trained Language Model for Efficient and Effective Text Generation**
*Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie and Ji-Rong Wen* 12:00-12:15 (Hall B)
We study the text generation task under the approach of pre-trained language models (PLMs). Typically, an auto-regressive (AR) method is adopted for generating texts in a token-by-token manner. Despite many advantages of AR generation, it usually suffers from inefficient inference. Therefore, non-autoregressive (NAR) models are proposed to generate all target tokens simultaneously. However, NAR models usually generate texts of lower quality due to the absence of token dependency in the output text. In this paper, we propose ELMER: an efficient and effective PLM for NAR text generation to explicitly model the token dependency during NAR generation. By leveraging the early exit technique, ELMER enables the token generations at different layers, according to their prediction confidence (a more confident token will exit at a lower layer). Besides, we propose a novel pre-training objective, Layer Permutation Language Modeling, to pre-train ELMER by permuting the exit layer for each token in sequences. Experiments on three text generation tasks show that ELMER significantly outperforms NAR models and further narrows the performance gap with AR PLMs (eg ELMER (29.92) vs BART (30.61) ROUGE-L in XSUM) while achieving over 10 times inference speedup.

**Curriculum Prompt Learning with Self-Training for Abstractive Dialogue Summarization**
*Changqun Li, Linlin Wang, xin Lin, Gerard de Melo and Liang He* 12:15-12:30 (Hall B)
Succinctly summarizing dialogue is a task of growing interest, but inherent challenges, such as insufficient training data and low information density impede our ability to train abstractive models. In this work, we propose a novel curriculum-based prompt learning method with self-training to address these problems. Specifically, prompts are learned using a curriculum learning strategy that gradually increases the degree of prompt perturbation, thereby improving the dialogue understanding and modeling capabilities of our model. Unlabeled dialogue is incorporated by means of self-training so as to reduce the dependency on labeled data. We further investigate topic-aware prompts to better plan for the generation of summaries. Experiments confirm that our model substantially outperforms strong baselines and achieves new state-of-the-art results on the AMI and ICSI datasets. Human evaluations also show the superiority of our model with regard to the summary generation quality.

## Resources and Evaluation 1

11:00-12:30 (Hall A, Room A)

**Multi-VQG: Generating Engaging Questions for Multiple Images**
*Min-Hsuan Yeh, Vincent Chen, Ting-Hao Huang and Lun-Wei Ku* 11:00-11:15 (Hall A, Room A)
Generating engaging content has drawn much recent attention in the NLP community. Asking questions is a natural way to respond to photos and promote awareness. However, most answers to questions in traditional question-answering (QA) datasets are factoids, which reduce individuals' willingness to answer. Furthermore, traditional visual question generation (VQG) confines the source data for question generation to single images, resulting in a limited ability to comprehend time-series information of the underlying event. In this paper, we propose generating engaging questions from multiple images. We present MVQG, a new dataset, and establish a series of baselines, including both end-to-end and dual-stage architectures. Results show that building stories behind the image sequence enables models to generate engaging questions, which confirms our assumption that people typically construct a picture of the event in their minds before asking questions. These results open up an exciting challenge for visual-and-language models to implicitly construct a story behind a series of photos to allow for creativity and experience sharing and hence draw attention to downstream applications.

**Tomayto, Tomahto. Beyond Token-level Answer Equivalence for Question Answering Evaluation**
*Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger and Tal Schuster* 11:15-11:30 (Hall A, Room A)
The predictions of question answering (QA) systems are typically evaluated against manually annotated finite sets of one or more answers. This leads to a coverage limitation that results in underestimating the true performance of systems, and is typically addressed by extending over exact match (EM) with predefined rules or with the token-level F1 measure. In this paper, we present the first systematic conceptual and data-driven analysis to examine the shortcomings of token-level equivalence measures.
To this end, we define the asymmetric notion of answer equivalence (AE), accepting answers that are equivalent to or improve over the reference, and publish over 23k human judgements for candidates produced by multiple QA systems on SQuAD.
Through a careful analysis of this data, we reveal and quantify several concrete limitations of the F1 measure, such as a false impression of graduality, or missing dependence on the question.
Since collecting AE annotations for each evaluated model is expensive, we learn a BERT matching (BEM) measure to approximate this task. Being a simpler task than QA, we find BEM to provide significantly better AE approximations than F1, and to more accurately reflect the performance of systems.
Finally, we demonstrate the practical utility of AE and BEM on the concrete application of minimal accurate prediction sets, reducing the number of required answers by up to X2.6.

**QRelScore: Better Evaluating Generated Questions with Deeper Understanding of Context-aware Relevance**
*Xiaoqiang Wang, Bang Liu, Siliang Tang and Lingfei Wu* 11:30-11:45 (Hall A, Room A)
Existing metrics for assessing question generation not only require costly human reference but also fail to take into account the input context of generation, rendering the lack of deep understanding of the relevance between the generated questions and input contexts. As a result, they may wrongly penalize a legitimate and reasonable candidate question when it (1) involves complicated reasoning with the context or (2) can be grounded by multiple evidences in the context. In this paper, we propose QRelScore, a context-aware Relevance evaluation metric for Question Generation. Based on off-the-shelf language models such as BERT and GPT2, QRelScore employs both word-level hierarchical matching and sentence-level prompt-based generation to cope with the complicated reasoning and diverse generation from multiple evidences, respectively. Compared with existing metrics, our experiments demonstrate that QRelScore is able to achieve a higher correlation with human judgments while being much more robust to adversarial samples.

**Generative Language Models for Paragraph-Level Question Generation**
*Asahi Ushio, Fernando Alva-Manchego and Jose Camacho-Collados* 11:45-12:00 (Hall A, Room A)
Powerful generative models have led to recent progress in question generation (QG). However, it is difficult to measure advances in QG research since there are no standardized resources that allow a uniform comparison among approaches. In this paper, we introduce QG-Bench, a multilingual and multidomain benchmark for QG that unifies existing question answering datasets by converting them to a standard QG setting. It includes general-purpose datasets such as SQuAD for English, datasets from ten domains and two styles, as well as datasets in eight different languages. Using QG-Bench as a reference, we perform an extensive analysis of the capabilities of language models for the task. First, we propose robust QG baselines based on fine-tuning generative language models. Then, we complement automatic evaluation based on standard metrics with an extensive manual evaluation, which in turn sheds light on the difficulty of evaluating QG models. Finally, we analyse both the domain adaptability of these models as well as the effectiveness of multilingual models in languages other than English. QG-Bench is released along with the fine-tuned models presented in the paper (https://github.com/asahi417/lm-question-generation), which are also available as a demo (https://autoqg.net/).

**Cross-document Event Coreference Search: Task, Dataset and Modeling**
*Alon Eirew, Avi Caciularu and Ido Dagan* 12:00-12:15 (Hall A, Room A)
The task of Cross-document Coreference Resolution has been traditionally formulated as requiring to identify all coreference links across a given set of documents. We propose an appealing, and often more applicable, complementary set up for the task – Cross-document Coreference Search, focusing in this paper on event coreference. Concretely, given a mention in context of an event of interest, considered as a query, the task is to find all coreferring mentions for the query event in a large document collection. To support research on this task, we create a corresponding dataset, which is derived from Wikipedia while leveraging annotations in the available Wikipedia Event Coreferecene dataset (WEC-Eng). Observing that the coreference search setup is largely analogous to the setting of Open Domain Question Answering, we adapt the prominent Deep Passage Retrieval (DPR) model to our setting, as an appealing baseline. Finally, we present a novel model that integrates

a powerful coreference scoring scheme into the DPR architecture, yielding improved performance.

**M2D2: A Massively Multi-Domain Language Modeling Dataset**
*Machel Reid, Victor Zhong, Suchin Gururangan and Luke Zettlemoyer* 12:15-12:30 (Hall A, Room A)
We present M2D2, a fine-grained, massively multi-domain corpus for studying domain adaptation in language models (LMs). M2D2 consists of 8.5B tokens and spans 145 domains extracted from Wikipedia and Semantic Scholar. Using ontologies derived from Wikipedia and ArXiv categories, we organize the domains in each data source into 22 groups. This two-level hierarchy enables the study of relationships between domains and their effects on in- and out-of-domain performance after adaptation. We also present a number of insights into the nature of effective domain adaptation in LMs, as examples of the new types of studies M2D2 enables. To improve in-domain performance, we show the benefits of adapting the LM along a domain hierarchy; adapting to smaller amounts of fine-grained domain-specific data can lead to larger in-domain performance gains than larger amounts of weakly relevant data. We further demonstrate a trade-off between in-domain specialization and out-of-domain generalization within and across ontologies, as well as a strong correlation between out-of-domain performance and lexical overlap between domains.

## Semantics

11:00-12:30 (Hall A, Room B)

**UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models**
*Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer and Tao Yu* 11:00-11:15 (Hall A, Room B)
Structured knowledge grounding (SKG) leverages structured knowledge to complete user requests, such as semantic parsing over databases and question answering over knowledge bases. Since the inputs and outputs of SKG tasks are heterogeneous, they have been studied separately by different communities, which limits systematic and compatible research on SKG. In this paper, we overcome this limitation by proposing the UnifiedSKG framework, which unifies 21 SKG tasks into a text-to-text format, aiming to promote systematic SKG research, instead of being exclusive to a single task, domain, or dataset. We use UnifiedSKG to benchmark T5 with different sizes and show that T5, with simple modifications when necessary, achieves state-of-the-art performance on almost all of the 21 tasks. We further demonstrate that multi-task prefix-tuning improves the performance on most tasks, largely improving the overall performance. UnifiedSKG also facilitates the investigation of zero-shot and few-shot learning, and we show that T0, GPT-3, and Codex struggle in zero-shot and few-shot learning for SKG. We also use UnifiedSKG to conduct a series of controlled experiments on structured knowledge encoding variants across SKG tasks. UnifiedSKG is easily extensible to more tasks, and it is open-sourced at https://github.com/hkunlp/unifiedskg.

**Reasoning Like Program Executors**
*Xinyu Pi, Qian Liu, Bei Chen, Morteza Ziyadi, Zeqi Lin, Qiang Fu, Yan Gao, Jian-Guang LOU and Weizhu Chen*11:15-11:30 (Hall A, Room B)
Reasoning over natural language is a long-standing goal for the research community. However, studies have shown that existing language models are inadequate in reasoning. To address the issue, we present POET, a novel reasoning pre-training paradigm. Through pre-training language models with programs and their execution results, POET empowers language models to harvest the reasoning knowledge possessed by program executors via a data-driven approach. POET is conceptually simple and can be instantiated by different kinds of program executors. In this paper, we showcase two simple instances POET-Math and POET-Logic, in addition to a complex instance, POET-SQL. Experimental results on six benchmarks demonstrate that POET can significantly boost model performance in natural language reasoning, such as numerical reasoning, logical reasoning, and multi-hop reasoning. POET opens a new gate on reasoning-enhancement pre-training, and we hope our analysis would shed light on the future research of reasoning like program executors.

**DocInfer: Document-level Natural Language Inference using Optimal Evidence Selection**
*Puneet Mathur, Gautam Kunapuli, Riyaz Bhat, Manish Shrivastava, Dinesh Manocha and Maneesh Singh* 11:30-11:45 (Hall A, Room B)
We present DocInfer - a novel, end-to-end Document-level Natural Language Inference model that builds a hierarchical document graph enriched through inter-sentence relations (topical, entity-based, concept-based), performs paragraph pruning using the novel SubGraph Pooling layer, followed by optimal evidence selection based on REINFORCE algorithm to identify the most important context sentences for a given hypothesis. Our evidence selection mechanism allows it to transcend the input length limitation of modern BERT-like Transformer models while presenting the entire evidence together for inferential reasoning. We show this is an important property needed to reason on large documents where the evidence may be fragmented and located arbitrarily far from each other. Extensive experiments on popular corpora - DocNLI, ContractNLI, and ConTRoL datasets, and our new proposed dataset called CaseHoldNLI on the task of legal judicial reasoning, demonstrate significant performance gains of 8-12% over SOTA methods. Our ablation studies validate the impact of our model. Performance improvement of 3-6% on annotation-scarce downstream tasks of fact verification, multiple-choice QA, and contract clause retrieval demonstrates the usefulness of DocInfer beyond primary NLI tasks.

**Infinite SCAN: An Infinite Model of Diachronic Semantic Change**
*Seiichi Inoue, Mamoru Komachi, Toshinobu Ogiso, Hiroya Takamura and Daichi Mochihashi* 11:45-12:00 (Hall A, Room B)
In this study, we propose a Bayesian model that can jointly estimate the number of senses of words and their changes through time. The model combines a dynamic topic model on Gaussian Markov random fields with a logistic stick-breaking process that realizes Dirichlet process. In the experiments, we evaluated the proposed model in terms of interpretability, accuracy in estimating the number of senses, and tracking their changes using both artificial data and real data. We quantitatively verified that the model behaves as expected through evaluation using artificial data. Using the CCOHA corpus, we showed that our model outperforms the baseline model and investigated the semantic changes of several well-known target words.

**Measuring Context-Word Biases in Lexical Semantic Datasets**
*Qianchu Liu, Diana McCarthy and Anna Korhonen* 12:00-12:15 (Hall A, Room B)
State-of-the-art pretrained contextualized models (PCM) eg. BERT use tasks such as WiC and WSD to evaluate their word-in-context representations. This inherently assumes that performance in these tasks reflect how well a model represents the coupled word and context semantics. We question this assumption by presenting the first quantitative analysis on the context-word interaction being tested in major contextual lexical semantic tasks. To achieve this, we run probing baselines on masked input, and propose measures to calculate and visualize the degree of context or word biases in existing datasets. The analysis was performed on both models and humans. Our findings demonstrate that models are usually not being tested for word-in-context semantics in the same way as humans are in these tasks, which helps us better understand the model-human gap. Specifically, to PCMs, most existing datasets fall into the extreme ends (the retrieval-based tasks exhibit

strong target word bias while WiC-style tasks and WSD show strong context bias); In comparison, humans are less biased and achieve much better performance when both word and context are available than with masked input. We recommend our framework for understanding and controlling these biases for model interpretation and future task design.

**Unobserved Local Structures Make Compositional Generalization Hard**
*Ben Bogin, Shivanshu Gupta and Jonathan Berant* 12:15-12:30 (Hall A, Room B)
While recent work has shown that sequence-to-sequence models struggle to generalize to new compositions (termed compositional generalization), little is known on what makes compositional generalization hard on a particular test instance. In this work, we investigate the factors that make generalization to certain test instances challenging. We first substantiate that some examples are more difficult than others by showing that different models consistently fail or succeed on the same test instances. Then, we propose a criterion for the difficulty of an example: a test instance is hard if it contains a local structure that was not observed at training time. We formulate a simple decision rule based on this criterion and empirically show it predicts instance-level generalization well across 5 different semantic parsing datasets, substantially better than alternative decision rules. Last, we show local structures can be leveraged for creating difficult adversarial compositional splits and also to improve compositional generalization under limited training budgets by strategically selecting examples for the training set.

## Summarization

11:00-12:30 (Hall A, Room C)

**Toward Unifying Text Segmentation and Long Document Summarization**
*Sangwoo Cho, Kaiqiang Song, Xiaoyang Wang, Fei Liu and Dong Yu* 11:00-11:15 (Hall A, Room C)
Text segmentation is important for signaling a document's structure. Without segmenting a long document into topically coherent sections, it is difficult for readers to comprehend the text, let alone find important information. The problem is only exacerbated by a lack of segmentation in transcripts of audio/video recordings. In this paper, we explore the role that section segmentation plays in extractive summarization of written and spoken documents. Our approach learns robust sentence representations by performing summarization and segmentation simultaneously, which is further enhanced by an optimization-based regularizer to promote selection of diverse summary sentences. We conduct experiments on multiple datasets ranging from scientific articles to spoken transcripts to evaluate the model's performance. Our findings suggest that the model can not only achieve state-of-the-art performance on publicly available benchmarks, but demonstrate better cross-genre transferability when equipped with text segmentation. We perform a series of analyses to quantify the impact of section segmentation on summarizing written and spoken documents of substantial length and complexity.

**SNaC: Coherence Error Detection for Narrative Summarization**
*Tanya Goyal, Junyi Jessy Li and Greg Durrett* 11:15-11:30 (Hall A, Room C)
Progress in summarizing long texts is inhibited by the lack of appropriate evaluation frameworks. A long summary that appropriately covers the facets of that text must also present a coherent narrative, but current automatic and human evaluation methods fail to identify gaps in coherence. In this work, we introduce SNaC, a narrative coherence evaluation framework for fine-grained annotations of long summaries. We develop a taxonomy of coherence errors in generated narrative summaries and collect span-level annotations for 6.6k sentences across 150 book and movie summaries. Our work provides the first characterization of coherence errors generated by state-of-the-art summarization models and a protocol for eliciting coherence judgments from crowdworkers. Furthermore, we show that the collected annotations allow us to benchmark past work in coherence modeling and train a strong classifier for automatically localizing coherence errors in generated summaries. Finally, our SNaC framework can support future work in long document summarization and coherence evaluation, including improved summarization modeling and post-hoc summary correction.

**HydraSum: Disentangling Style Features in Text Summarization with Multi-Decoder Models**
*Tanya Goyal, Nazneen Rajani, Wenhao Liu and Wojciech Kryscinski* 11:30-11:45 (Hall A, Room C)
Summarization systems make numerous "decisions" about summary properties during inference, e.g. degree of copying, specificity and length of outputs, etc. However, these are implicitly encoded within model parameters and specific styles cannot be enforced. To address this, we introduce HydraSum, a new summarization architecture that extends the single decoder framework of current models to a mixture-of-experts version with multiple decoders. We show that HydraSum's multiple decoders automatically learn contrasting summary styles when trained under the standard training objective without any extra supervision. Through experiments on three summarization datasets (CNN, Newsroom and XSum), we show that HydraSum provides a simple mechanism to obtain stylistically-diverse summaries by sampling from either individual decoders or their mixtures, outperforming baseline models. Finally, we demonstrate that a small modification to the gating strategy during training can enforce an even stricter style partitioning, e.g. high- vs low-abstractiveness or high- vs low-specificity, allowing users to sample from a larger area in the generation space and vary summary styles along multiple dimensions.

**SEM-F1: an Automatic Way for Semantic Evaluation of Multi-Narrative Overlap Summaries at Scale**
*Naman Bansal, Mousumi Akter and Shubhra Kanti Karmaker Santu* 11:45-12:00 (Hall A, Room C)
Recent work has introduced an important yet relatively under-explored NLP task called Semantic Overlap Summarization (SOS) that entails generating a summary from multiple alternative narratives which conveys the common information provided by those narratives. Previous work also published a benchmark dataset for this task by collecting 2,925 alternative narrative pairs from the web and manually annotating 411 different reference summaries by engaging human annotators. In this paper, we exclusively focus on the automated evaluation of the SOS task using the benchmark dataset. More specifically, we first use the popular ROUGE metric from text-summarization literature and conduct a systematic study to evaluate the SOS task. Our experiments discover that ROUGE is not suitable for this novel task and therefore, we propose a new sentence-level precision-recall style automated evaluation metric, called SEM-F1 (Semantic F1). It is inspired by the benefits of the sentence-wise annotation technique using overlap labels reported by the previous work. Our experiments show that the proposed SEM-F1 metric yields a higher correlation with human judgment and higher inter-rater agreement compared to the ROUGE metric.

**SQuALITY: Building a Long-Document Summarization Dataset the Hard Way**
*Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang and Samuel R. Bowman* 12:00-12:15 (Hall A, Room C)
Summarization datasets are often assembled either by scraping naturally occurring public-domain summaries—which are nearly always in difficult-to-work-with technical domains—or by using approximate heuristics to extract them from everyday text—which frequently yields unfaithful summaries. In this work, we turn to a slower but more straightforward approach to developing summarization benchmark data: We hire highly-qualified contractors to read stories and write original summaries from scratch. To amortize reading time, we collect five summaries per document, with the first giving an overview and the subsequent four addressing specific questions. We use this protocol to collect SQuALITY, a dataset of question-focused summaries built on the same public-domain short stories as the multiple-choice dataset QuALITY (Pang et al., 2021). Experiments with state-of-the-art summarization systems show that our dataset is challenging and that existing automatic

evaluation metrics are weak indicators of quality.

**How Far are We from Robust Long Abstractive Summarization?**
*Huan Yee Koh, Jiaxin Ju, He Zhang, Ming Liu and Shirui Pan* 12:15-12:30 (Hall A, Room C)
Abstractive summarization has made tremendous progress in recent years. In this work, we perform fine-grained human annotations to evaluate long document abstractive summarization systems (i.e., models and metrics) with the aim of implementing them to generate reliable summaries. For long document abstractive models, we show that the constant strive for state-of-the-art ROUGE results can lead us to generate more relevant summaries but not factual ones. For long document evaluation metrics, human evaluation results show that ROUGE remains the best at evaluating the relevancy of a summary. It also reveals important limitations of factuality metrics in detecting different types of factual errors and the reasons behind the effectiveness of BARTScore. We then suggest promising directions in the endeavor of developing factual consistency metrics. Finally, we release our annotated long document dataset with the hope that it can contribute to the development of metrics across a broader range of summarization settings.

## Industry 1

11:00-12:30 (Hall A, Room D)

[INDUSTRY] **Improving Large-Scale Conversational Assistants using Model Interpretation based Training Sample Selection**
*Stefan Schroedl, Manoj Kumar, Kiana Hajebi, Morteza Ziyadi, Sriram Venkatapathy, Anil Ramakrishna, Rahul Gupta and Pradeep Natarajan* 11:00-11:15 (Hall A, Room D)
This paper presents an approach to identify samples from live traffic where the customer implicitly communicated satisfaction with Alexa's responses, by leveraging interpretations of model behavior. Such customer signals are noisy and adding a large number of samples from live traffic to training set makes re-training infeasible. Our work addresses these challenges by identifying a small number of samples that grow training set by 0.05% while producing statistically significant improvements in both offline and online tests.

[INDUSTRY] **CGF: Constrained Generation Framework for Query Rewriting in Conversational AI**
*Jie Hao, Yang Liu, Xing Fan, Saurabh Gupta, Saleh Soltan, Rakesh Chada, Pradeep Natarajan, Chenlei Guo and Gokhan Tur* 11:15-11:30 (Hall A, Room D)
In conversational AI agents, Query Rewriting (QR) plays a crucial role in reducing user frictions and satisfying their daily demands. User frictions are caused by various reasons, such as errors in the conversational AI system, users' accent or their abridged language. In this work, we present a novel Constrained Generation Framework (CGF) for query rewriting at both global and personalized levels. It is based on the encoder-decoder framework, where the encoder takes the query and its previous dialogue turns as the input to form a context-enhanced representation, and the decoder uses constrained decoding to generate the rewrites based on the pre-defined global or personalized constrained decoding space. Extensive offline and online A/B experiments show that the proposed CGF significantly boosts the query rewriting performance.

[INDUSTRY] **SimANS: Simple Ambiguous Negatives Sampling for Dense Text Retrieval**
*Kun Zhou, Yeyun Gong, Xiao Liu, Wayne Xin Zhao, Yelong Shen, Anlei Dong, Jingwen Lu, Rangan Majumder, Ji-Rong Wen, Nan Duan and Weizhu Chen* 11:30-11:45 (Hall A, Room D)
Sampling proper negatives from a large document pool is vital to effectively train a dense retrieval model. However, existing negative sampling strategies suffer from the uninformative or false negative problem. In this work, we empirically show that according to the measured relevance scores, the negatives ranked around the positives are generally more informative and less likely to be false negatives. Intuitively, these negatives are not too hard (*may be false negatives*) or too easy (*uninformative*). They are the ambiguous negatives and need more attention during training. Thus, we propose a simple ambiguous negatives sampling method, SimANS, which incorporates a new sampling probability distribution to sample more ambiguous negatives. Extensive experiments on four public and one industry datasets show the effectiveness of our approach. We made the code and models publicly available in `https://github.com/microsoft/SimXNS`.

[INDUSTRY] **Learning Geolocations for Cold-Start and Hard-to-Resolve Addresses via Deep Metric Learning**
*Govind . and Saurabh Sohoney* 11:45-12:00 (Hall A, Room D)
With evergrowing digital adoption in the society and increasing demand for businesses to deliver to customers doorstep, the last mile hop of transportation planning poses unique challenges in emerging geographies with unstructured addresses. One of the crucial inputs to facilitate effective planning is the task of geolocating customer addresses. Existing systems operate by aggregating historical delivery locations or by resolving/matching addresses to known buildings and campuses to vend a high-precision geolocation. However, by design they fail to cater to a significant fraction of addresses which are new in the system and have inaccurate or missing building level information. We propose a framework to resolve these addresses (referred to as hard-to-resolve henceforth) to a shallower granularity termed as neighbourhood. Specifically, we propose a weakly supervised deep metric learning model to encode the geospatial semantics in address embeddings. We present empirical evaluation on India (IN) and the United Arab Emirates (UAE) hard-to-resolve addresses to show significant improvements in learning geolocations i.e., 22% (IN) & 55% (UAE) reduction in delivery defects (where learnt geocode is >Y meters away from actual location), and 43% (IN) & 90% (UAE) reduction in 50th percentile (p50) distance between learnt and actual delivery locations over the existing production system.

[INDUSTRY] **Large-scale Machine Translation for Indian Languages in E-commerce under Low Resource Constraints**
*Amey Patil and Nikesh Garera* 12:00-12:15 (Hall A, Room D)
The democratization of e-commerce platforms has moved an increasingly diversified Indian user base to shop online. We have deployed reliable and precise large-scale Machine Translation systems for several Indian regional languages in this work. Building such systems is a challenge because of the low-resource nature of the Indian languages. We develop a structured model development pipeline as a closed feedback loop with external manual feedback through an Active Learning component. We show strong synthetic parallel data generation capability and consistent improvements to the model over iterations. Starting with 1.2M parallel pairs for English-Hindi we have compiled a corpus with 400M+ synthetic high quality parallel pairs across different domains. Further, we need colloquial translations to preserve the intent and friendliness of English content in regional languages, and make it easier to understand for our users. We perform robust and effective domain adaptation steps to achieve colloquial such translations. Over iterations, we show 9.02 BLEU points improvement for English to Hindi translation model. Along with Hindi, we show that the overall approach and best practices extends well to other Indian languages, resulting in deployment of our models across 7 Indian Languages.

[INDUSTRY] **Improving Text-to-SQL Semantic Parsing with Fine-grained Query Understanding**
*Jun Wang, Patrick Ng, Alexander Hanbo Li, Jiarong Jiang, Zhiguo Wang, Bing Xiang, Ramesh Nallapati and Sudipta Sengupta* 12:15-12:30 (Hall A, Room D)

Most recent research on Text-to-SQL semantic parsing relies on either parser itself or simple heuristic based approach to understand natural language query (NLQ). When synthesizing a SQL query, there is no explicit semantic information of NLQ available to the parser which leads to undesirable generalization performance. In addition, without lexical-level fine-grained query understanding, linking between query and database can only rely on fuzzy string match which leads to suboptimal performance in real applications. In view of this, in this paper we present a general-purpose, modular neural semantic parsing framework that is based on token-level fine-grained query understanding. Our framework consists of three modules: named entity recognizer (NER), neural entity linker (NEL) and neural semantic parser (NSP). By jointly modeling query and database, NER model analyzes user intents and identifies entities in the query. NEL model links typed entities to schema and cell values in database. Parser model leverages available semantic information and linking results and synthesizes tree-structured SQL queries based on dynamically generated grammar. Experiments on SQUALL, a newly released semantic parsing dataset, show that we can achieve 56.8% execution accuracy on WikiTableQuestions (WTQ) test set, which outperforms the state-of-the-art model by 2.7%.

## NLP Applications 1

11:00-12:30 (Collaboratorium)

**Learning to Generate Question by Asking Question: A Primal-Dual Approach with Uncommon Word Generation**
*Qifan Wang, Li Yang, Xiaojun Quan, Fuli Feng, Dongfang Liu, Zenglin Xu, Sinong Wang and Hao Ma* 11:00-11:15 (Collaboratorium)
Automatic question generation (AQG) is the task of generating a question from a given passage and an answer. Most existing AQG methods aim at encoding the passage and the answer to generate the question. However, limited work has focused on modeling the correlation between the target answer and the generated question. Moreover, unseen or rare word generation has not been studied in previous works. In this paper, we propose a novel approach which incorporates question generation with its dual problem, question answering, into a unified primal-dual framework. Specifically, the question generation component consists of an encoder that jointly encodes the answer with the passage, and a decoder that produces the question. The question answering component then re-asks the generated question on the passage to ensure that the target answer is obtained. We further introduce a knowledge distillation module to improve the model generalization ability. We conduct an extensive set of experiments on SQuAD and HotpotQA benchmarks. Experimental results demonstrate the superior performance of the proposed approach over several state-of-the-art methods.

**PAIR: Prompt-Aware margIn Ranking for Counselor Reflection Scoring in Motivational Interviewing**
*Do June Min, Verónica Pérez-Rosas, Kenneth Resnicow and Rada Mihalcea* 11:15-11:30 (Collaboratorium)
Counselor reflection is a core verbal skill used by mental health counselors to express understanding and affirmation of the client's experience and concerns. In this paper, we propose a system for the analysis of counselor reflections. Specifically, our system takes as input one dialog turn containing a client prompt and a counselor response, and outputs a score indicating the level of reflection in the counselor response. We compile a dataset consisting of different levels of reflective listening skills, and propose the Prompt-Aware margIn Ranking (PAIR) framework that contrasts positive and negative prompt and response pairs using specially designed multi-gap and prompt-aware margin ranking losses. Through empirical evaluations and deployment of our system in a real-life educational environment, we show that our analysis model outperforms several baselines on different metrics, and can be used to provide useful feedback to counseling trainees.

**Robustness of Fusion-based Multimodal Classifiers to Cross-Modal Content Dilutions**
*Gaurav Verma, Vishwa Vinay, Ryan Rossi and Srijan Kumar* 11:30-11:45 (Collaboratorium)
As multimodal learning finds applications in a wide variety of high-stakes societal tasks, investigating their robustness becomes important. Existing work has focused on understanding the robustness of vision-and-language models to imperceptible variations on benchmark tasks. In this work, we investigate the robustness of multimodal classifiers to cross-modal dilutions – a plausible variation. We develop a model that, given a multimodal (image + text) input, generates additional dilution text that (a) maintains relevance and topical coherence with the image and existing text, and (b) when added to the original text, leads to misclassification of the multimodal input. Via experiments on Crisis Humanitarianism and Sentiment Detection tasks, we find that the performance of task-specific fusion-based multimodal classifiers drops by 23.3% and 22.5%, respectively, in the presence of dilutions generated by our model. Metric-based comparisons with several baselines and human evaluations indicate that our dilutions show higher relevance and topical coherence, while simultaneously being more effective at demonstrating the brittleness of the multimodal classifiers. Our work aims to highlight and encourage further research on the robustness of deep multimodal models to realistic variations, especially in human-facing societal applications.

**Translation between Molecules and Natural Language**
*Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho and Heng Ji* 11:45-12:00 (Collaboratorium)
We present MolT5 - a self-supervised learning framework for pretraining models on a vast amount of unlabeled natural language text and molecule strings. MolT5 allows for new, useful, and challenging analogs of traditional vision-language tasks, such as molecule captioning and text-based de novo molecule generation (altogether: translation between molecules and language), which we explore for the first time. Since MolT5 pretrains models on single-modal data, it helps overcome the chemistry domain shortcoming of data scarcity. Furthermore, we consider several metrics, including a new cross-modal embedding-based metric, to evaluate the tasks of molecule captioning and text-based molecule generation. Our results show that MolT5-based models are able to generate outputs, both molecules and captions, which in many cases are high quality.

**Guiding Neural Entity Alignment with Compatibility**
*Bing Liu, Harrisen Scells, Wen Hua, Guido Zuccon, Genghong Zhao and Xia Zhang* 12:00-12:15 (Collaboratorium)
Entity Alignment (EA) aims to find equivalent entities between two Knowledge Graphs (KGs). While numerous neural EA models have been devised, they are mainly learned using labelled data only. In this work, we argue that different entities within one KG should have compatible counterparts in the other KG due to the potential dependencies among the entities. Making compatible predictions thus should be one of the goals of training an EA model along with fitting the labelled data: this aspect however is neglected in current methods. To power neural EA models with compatibility, we devise a training framework by addressing three problems: (1) how to measure the compatibility of an EA model; (2) how to inject the property of being compatible into an EA model; (3) how to optimise parameters of the compatibility model. Extensive experiments on widely-used datasets demonstrate the advantages of integrating compatibility within EA models. In fact, state-of-the-art neural EA models trained within our framework using just 5% of the labelled data can achieve comparable effectiveness with supervised training using 20% of the labelled data.

**How Large Language Models are Transforming Machine-Paraphrase Plagiarism**
*Jan Philip Wahle, Terry Ruas, Frederic Kirstein and Bela Gipp* 12:15-12:30 (Collaboratorium)
The recent success of large language models for text generation poses a severe threat to academic integrity, as plagiarists can generate realistic paraphrases indistinguishable from original work. However, the role of large autoregressive models in generating machine-paraphrased pla-

giarism and their detection is still incipient in the literature. This work explores T5 and GPT3 for machine-paraphrase generation on scientific articles from arXiv, student theses, and Wikipedia. We evaluate the detection performance of six automated solutions and one commercial plagiarism detection software and perform a human study with 105 participants regarding their detection performance and the quality of generated examples. Our results suggest that large language models can rewrite text humans have difficulty identifying as machine-paraphrased (53% mean acc.). Human experts rate the quality of paraphrases generated by GPT-3 as high as original texts (clarity 4.0/5, fluency 4.2/5, coherence 3.8/5). The best-performing detection model (GPT-3) achieves 66% F1-score in detecting paraphrases. We make our code, data, and findings publicly available to facilitate the development of detection solutions.

## Poster Sessions 1 & 2

11:00-12:30 (Atrium)

**TranSHER: Translating Knowledge Graph Embedding with Hyper-Ellipsoidal Restriction**
*Yizhi Li, Wei Fan, Chao Liu, Chenghua Lin and Jiang Qian* 11:00-12:30 (Atrium)
Knowledge graph embedding methods are important for the knowledge graph completion (or link prediction) task. One state-of-the-art method, PairRE, leverages two separate vectors to model complex relations (i.e., 1-to-N, N-to-1, and N-to-N) in knowledge graphs. However, such a method strictly restricts entities on the hyper-ellipsoid surfaces which limits the optimization of entity distribution, leading to suboptimal performance of knowledge graph completion. To address this issue, we propose a novel score function TranSHER, which leverages relation-specific translations between head and tail entities to relax the constraint of hyper-ellipsoid restrictions. By introducing an intuitive and simple relation-specific translation, TranSHER can provide more direct guidance on optimization and capture more semantic characteristics of entities with complex relations. Experimental results show that TranSHER achieves state-of-the-art performance on link prediction and generalizes well to datasets in different domains and scales. Our codes are public available at https://github.com/yizhilll/TranSHER.

**Robots-Dont-Cry: Understanding Falsely Anthropomorphic Utterances in Dialog Systems**
*David Gros, Yu Li and Zhou Yu* 11:00-12:30 (Atrium)
Dialog systems are often designed or trained to output human-like responses. However, some responses may be impossible for a machine to truthfully say (e.g. "that movie made me cry"). Highly anthropomorphic responses might make users uncomfortable or implicitly deceive them into thinking they are interacting with a human. We collect human ratings on the feasibility of approximately 900 two-turn dialogs sampled from 9 diverse data sources. Ratings are for two hypothetical machine embodiments: a futuristic humanoid robot and a digital assistant. We find that for some data-sources commonly used to train dialog systems, 20-30
We explore qualitative and quantitative reasons for these ratings. Finally, we build classifiers and explore how modeling configuration might affect output permissibly, and discuss implications for building less falsely anthropomorphic dialog systems.

**When More Data Hurts: A Troubling Quirk in Developing Broad-Coverage Natural Language Understanding Systems**
*Elias Stengel-Eskin, Emmanouil Antonios Platanios, Adam Pauls, Sam Thomson, Hao Fang, Benjamin Van Durme, Jason Eisner and Yu Su*
11:00-12:30 (Atrium)
In natural language understanding (NLU) production systems, users' evolving needs necessitate the addition of new features over time, indexed by new symbols added to the meaning representation space. This requires additional training data and results in ever-growing datasets. We present the first systematic investigation into this incremental symbol learning scenario. Our analysis reveals a troubling quirk in building broad-coverage NLU systems: as the training dataset grows, performance on a small set of new symbols often decreases. We show that this trend holds for multiple mainstream models on two common NLU tasks: intent recognition and semantic parsing. Rejecting class imbalance as the sole culprit, we reveal that the trend is closely associated with an effect we call source signal dilution, where strong lexical cues for the new symbol become diluted as the training dataset grows. Selectively dropping training examples to prevent dilution often reverses the trend, showing the over-reliance of mainstream neural NLU models on simple lexical cues.

**Less is More: Summary of Long Instructions is Better for Program Synthesis**
*Kirby Kuznia, Swaroop Mishra, Mihir Parmar and Chitta Baral* 11:00-12:30 (Atrium)
Despite the success of large pre-trained language models (LMs) such as Codex, they show below-par performance on the larger and more complicated programming related questions. We show that LMs benefit from the summarized version of complicated questions. Our findings show that superfluous information often present in problem description such as human characters, background stories, and names (which are included to help humans in understanding a task) does not help models in understanding a task. To this extent, we create a meta-dataset from the frequently used APPS dataset and the newly created CodeContests dataset for the program synthesis task. Our meta-dataset consists of human and synthesized summaries of the long and complicated programming questions. Experimental results on Codex show that our proposed approach outperforms baseline by 8.13% on the APPS dataset and 11.88% on the CodeContests dataset on an average in terms of strict accuracy. Our analysis shows that summaries significantly improve performance for introductory (9.86%) and interview (11.48%) related programming questions. However, it shows improvement by a small margin ( 2%) for competitive programming questions, implying the scope for future research direction.

**HashFormers: Towards Vocabulary-independent Pre-trained Transformers**
*Huiyin Xue and Nikolaos Aletras* 11:00-12:30 (Atrium)
Transformer-based pre-trained language models are vocabulary-dependent, mapping by default each token to its corresponding embedding. This one-to-one mapping results into embedding matrices that occupy a lot of memory (i.e. millions of parameters) and grow linearly with the size of the vocabulary. Previous work on on-device transformers dynamically generate token embeddings on-the-fly without embedding matrices using locality-sensitive hashing over morphological information. These embeddings are subsequently fed into transformer layers for text classification. However, these methods are not pre-trained. Inspired by this line of work, we propose HashFormers, a new family of vocabulary-independent pre-trained transformers that support an unlimited vocabulary (i.e. all possible tokens in a corpus) given a substantially smaller fixed-sized embedding matrix. We achieve this by first introducing computationally cheap hashing functions that bucket together individual tokens to embeddings. We also propose three variants that do not require an embedding matrix at all, further reducing the memory requirements. We empirically demonstrate that HashFormers are more memory efficient compared to standard pre-trained transformers while achieving comparable predictive performance when fine-tuned on multiple text classification tasks. For example, our most efficient HashFormer variant has a negligible performance degradation (0.4% on GLUE) using only 99.1K parameters for representing the embeddings compared to 12.3-38M parameters of state-of-the-art models.

**AMAL: Meta Knowledge-Driven Few-Shot Adapter Learning**
*S. K. Hong and Tae Young Jang* 11:00-12:30 (Atrium)
NLP has advanced greatly together with the proliferation of Transformer-based pre-trained language models. To adapt to a downstream task,

the pre-trained language models need to be fine-tuned with a sufficient supply of annotated examples. In recent years, Adapter-based fine-tuning methods have expanded the applicability of pre-trained language models by substantially lowering the required amount of annotated examples. However, existing Adapter-based methods still fail to yield meaningful results in the few-shot regime where only a few annotated examples are provided. In this study, we present a meta-learning-driven low-rank adapter pooling method, called AMAL, for leveraging pre-trained language models even with just a few data points. We evaluate our method on five text classification benchmark datasets. The results show that AMAL significantly outperforms previous few-shot learning methods and achieves a new state-of-the-art.

**Facilitating Contrastive Learning of Discourse Relational Senses by Exploiting the Hierarchy of Sense Relations**
*Wanqiu Long and Bonnie Webber* 11:00-12:30 (Atrium)
Implicit discourse relation recognition is a challenging task that involves identifying the sense or senses that hold between two adjacent spans of text, in the absense of an explicit connective between them. In both PDTB-2 (prasad et al., 2008) and PDTB-3 (Webber et al., 2019), discourse relational senses are organized into a three-level hierarchy ranging from four broad top-level senses, to more specific senses below them. Most previous work on implicitf discourse relation recognition have used the sense hierarchy simply to indicate what sense labels were available. Here we do more — incorporating the sense hierarchy into the recognition process itself and using it to select the negative examples used in contrastive learning. With no additional effort, the approach achieves state-of-the-art performance on the task. Our code is released in https://github.com/wanqiulong 0923/Contrastive_IDRR.

**Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks**
*Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick and Reza Shokri* 11:00-12:30 (Atrium)
The wide adoption and application of Masked language models (MLMs) on sensitive data (from legal to medical) necessitates a thorough quantitative investigation into their privacy vulnerabilities. Prior attempts at measuring leakage of MLMs via membership inference attacks have been inconclusive, implying potential robustness of MLMs to privacy attacks. In this work, we posit that prior attempts were inconclusive because they based their attack solely on the MLM's model score. We devise a stronger membership inference attack based on likelihood ratio hypothesis testing that involves an additional reference MLM to more accurately quantify the privacy risks of memorization in MLMs. We show that masked language models are indeed susceptible to likelihood ratio membership inference attacks: Our empirical results, on models trained on medical notes, show that our attack improves the AUC of prior membership inference attacks from $0.66$ to an alarmingly high $0.90$ level.

**MatchPrompt: Prompt-based Open Relation Extraction with Semantic Consistency Guided Clustering**
*Jiaxin Wang, Lingling Zhang, Jun Liu, Xi Liang, Yujie Zhong and Yaqiang Wu* 11:00-12:30 (Atrium)
Relation clustering is a general approach for open relation extraction (OpenRE). Current methods have two major problems. One is that their good performance relies on large amounts of labeled and pre-defined relational instances for pre-training, which are costly to acquire in reality. The other is that they only focus on learning a high-dimensional metric space to measure the similarity of novel relations and ignore the specific relational representations of clusters. In this work, we propose a new prompt-based framework named MatchPrompt, which can realize OpenRE with efficient knowledge transfer from only a few pre-defined relational instances as well as mine the specific meanings for cluster interpretability. To our best knowledge, we are the first to introduce a prompt-based framework for unlabeled clustering. Experimental results on different datasets show that MatchPrompt achieves the new SOTA results for OpenRE.

**Incorporating Relevance Feedback for Information-Seeking Retrieval using Few-Shot Document Re-Ranking**
*Tim Baumgärtner, Leonardo F. R. Ribeiro, Nils Reimers and Iryna Gurevych* 11:00-12:30 (Atrium)
Pairing a lexical retriever with a neural re-ranking model has set state-of-the-art performance on large-scale information retrieval datasets. This pipeline covers scenarios like question answering or navigational queries, however, for information-seeking scenarios, users often provide information on whether a document is relevant to their query in form of clicks or explicit feedback. Therefore, in this work, we explore how relevance feedback can be directly integrated into neural re-ranking models by adopting few-shot and parameter-efficient learning techniques. Specifically, we introduce a kNN approach that re-ranks documents based on their similarity with the query and the documents the user considers relevant. Further, we explore Cross-Encoder models that we pre-train using meta-learning and subsequently fine-tune for each query, training only on the feedback documents. To evaluate our different integration strategies, we transform four existing information retrieval datasets into the relevance feedback scenario. Extensive experiments demonstrate that integrating relevance feedback directly in neural re-ranking models improves their performance, and fusing lexical ranking with our best performing neural re-ranker outperforms all other methods by 5.2% nDCG@20.

**Extending Logic Explained Networks to Text Classification**
*Rishabh Jain, Gabriele Ciravegna, Pietro Barbiero, Francesco Giannini, Davide Buffelli and Pietro Lio* 11:00-12:30 (Atrium)
Recently, Logic Explained Networks (LENs) have been proposed as explainable-by-design neural models providing logic explanations for their predictions. However, these models have only been applied to vision and tabular data, and they mostly favour the generation of global explanations, while local ones tend to be noisy and verbose. For these reasons, we propose LEN<sup>p</sup>, improving local explanations by perturbing input words, and we test it on text classification. Our results show that (i) LEN<sup>p</sup> provides better local explanations than LIME in terms of sensitivity and faithfulness, and (ii) its logic explanations are more useful and user-friendly than the feature scoring provided by LIME as attested by a human survey.

**Are All Spurious Features in Natural Language Alike? An Analysis through a Causal Lens**
*Nitish Joshi, Xiang Pan and He He* 11:00-12:30 (Atrium)
The term 'spurious correlations' has been used in NLP to informally denote any undesirable feature-label correlations. However, a correlation can be undesirable because (i) the feature is irrelevant to the label (e.g. punctuation in a review), or (ii) the feature's effect on the label depends on the context (e.g. negation words in a review), which is ubiquitous in language tasks. In case (i), we want the model to be invariant to the feature, which is neither necessary nor sufficient for prediction. But in case (ii), even an ideal model (e.g. humans) must rely on the feature, since it is necessary (but not sufficient) for prediction. Therefore, a more fine-grained treatment of spurious features is needed to specify the desired model behavior. We formalize this distinction using a causal model and probabilities of necessity and sufficiency, which delineates the causal relations between a feature and a label. We then show that this distinction helps explain results of existing debiasing methods on different spurious features, and demystifies surprising results such as the encoding of spurious features in model representations after debiasing.

**Human Guided Exploitation of Interpretable Attention Patterns in Summarization and Topic Segmentation**
*Raymond Li, Wen Xiao, Linzi Xing, Lanjun Wang, Gabriel Murray and Giuseppe Carenini* 11:00-12:30 (Atrium)
The multi-head self-attention mechanism of the transformer model has been thoroughly investigated recently. In one vein of study, researchers are interested in understanding why and how transformers work. In another vein, researchers propose new attention augmentation methods to make transformers more accurate, efficient and interpretable. In this paper, we combine these two lines of research in a human-in-the-loop pipeline to first discover important task-specific attention patterns. Then those patterns are injected, not only to smaller models, but also to the original model. The benefits of our pipeline and discovered patterns are demonstrated in two case studies with extractive summarization

and topic segmentation. After discovering interpretable patterns in BERT-based models fine-tuned for the two downstream tasks, experiments indicate that when we inject the patterns into attention heads, the models show considerable improvements in accuracy and efficiency.

**InforMask: Unsupervised Informative Masking for Language Model Pretraining**
*Nafis Sadeq, Canwen Xu and Julian McAuley* 11:00-12:30 (Atrium)
Masked language modeling is widely used for pretraining large language models for natural language understanding (NLU). However, random masking is suboptimal, allocating an equal masking rate for all tokens. In this paper, we propose InforMask, a new unsupervised masking strategy for training masked language models. InforMask exploits Pointwise Mutual Information (PMI) to select the most informative tokens to mask. We further propose two optimizations for InforMask to improve its efficiency. With a one-off preprocessing step, InforMask outperforms random masking and previously proposed masking strategies on the factual recall benchmark LAMA and the question answering benchmark SQuAD v1 and v2.

**Subword Evenness (SuE) as a Predictor of Cross-lingual Transfer to Low-resource Languages**
*Olga Pelloni, Anastassia Shaitarova and Tanja Samardzic* 11:00-12:30 (Atrium)
Pre-trained multilingual models, such as mBERT, XLM-R and mT5, are used to improve the performance on various tasks in low-resource languages via cross-lingual transfer. In this framework, English is usually seen as the most natural choice for a transfer language (for fine-tuning or continued training of a multilingual pre-trained model), but it has been revealed recently that this is often not the best choice. The success of cross-lingual transfer seems to depend on some properties of languages, which are currently hard to explain. Successful transfer often happens between unrelated languages and it often cannot be explained by data-dependent factors.

In this study, we show that languages written in non-Latin and non-alphabetic scripts (mostly Asian languages) are the best choices for improving performance on the task of Masked Language Modelling (MLM) in a diverse set of 30 low-resource languages and that the success of the transfer is well predicted by our novel measure of Subword Evenness (SuE). Transferring language models over the languages that score low on our measure results in the lowest average perplexity over target low-resource languages. Our correlation coefficients obtained with three different pre-trained multilingual models are consistently higher than all the other predictors, including text-based measures (type-token ratio, entropy) and linguistically motivated choice (genealogical and typological proximity).

**Don't Prompt, Search! Mining-based Zero-Shot Learning with Language Models**
*Mozes van de Kar, Mengzhou Xia, Danqi Chen and Mikel Artetxe* 11:00-12:30 (Atrium)
Masked language models like BERT can perform text classification in a zero-shot fashion by reformulating downstream tasks as text infilling. However, this approach is highly sensitive to the template used to prompt the model, yet practitioners are blind when designing them in strict zero-shot settings. In this paper, we propose an alternative mining-based approach for zero-shot learning. Instead of prompting language models, we use regular expressions to mine labeled examples from unlabeled corpora, which can optionally be filtered through prompting, and used to finetune a pretrained model. Our method is more flexible and interpretable than prompting, and outperforms it on a wide range of tasks when using comparable templates. Our results suggest that the success of prompting can partly be explained by the model being exposed to similar examples during pretraining, which can be directly retrieved through regular expressions.

**Fine-Tuning Pre-trained Transformers into Decaying Fast Weights**
*Huanru Henry Mao* 11:00-12:30 (Atrium)
Autoregressive Transformers are strong language models but incur O(T) complexity during per-token generation due to the self-attention mechanism. Recent work proposes kernel-based methods to approximate causal self-attention by replacing it with recurrent formulations with various update rules and feature maps to achieve O(1) time and memory complexity. We explore these approaches and find that they are unnecessarily complex, and propose a simple alternative - decaying fast weights - that runs fast on GPU, outperforms prior methods, and retains 99% of attention's performance for GPT-2. We also show competitive performance on WikiText-103 against more complex attention substitutes.

**Efficient Large Scale Language Modeling with Mixtures of Experts**
*Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, Giridharan Anantharaman, Xian Li, Shuohui Chen, Halil Akin, Mandeep Baines, Louis Martin, Xing Zhou, Punit Singh Koura, Brian O'Horo, Jeffrey Wang, Luke Zettlemoyer, Mona Diab, Zornitsa Kozareva and Veselin Stoyanov* 11:00-12:30 (Atrium)
Mixture of Experts layers (MoEs) enable efficient scaling of language models through conditional computation. This paper presents a detailed empirical study of how autoregressive MoE language models scale in comparison with dense models in a wide range of settings: in- and out-of-domain language modeling, zero- and few-shot priming, and full-shot fine-tuning. With the exception of fine-tuning, we find MoEs to be substantially more compute efficient. At more modest training budgets, MoEs can match the performance of dense models using 4 times less compute. This gap narrows at scale, but our largest MoE model (1.1T parameters) consistently outperforms a compute-equivalent dense model (6.7B parameters). Overall, this performance gap varies greatly across tasks and domains, suggesting that MoE and dense models generalize differently in ways that are worthy of future study. We make our code and models publicly available for research use.

**The Curious Case of Control**
*Elias Stengel-Eskin and Benjamin Van Durme* 11:00-12:30 (Atrium)
Children acquiring English make systematic errors on subject control sentences even after they have reached near-adult competence (Chomsky, 1969), possibly due to heuristics based on semantic roles (Maratsos, 1974). Given the advanced fluency of large generative language models, we ask whether model outputs are consistent with these heuristics, and to what degree different models are consistent with each other. We find that models can be categorized by behavior into three separate groups, with broad differences between the groups. The outputs of models in the largest group are consistent with positional heuristics that succeed on subject control but fail on object control. This result is surprising, given that object control is orders of magnitude more frequent in the text data used to train such models. We examine to what degree the models are sensitive to prompting with agent-patient information, finding that raising the salience of agent and patient relations results in significant changes in the outputs of most models. Based on this observation, we leverage an existing dataset of semantic proto-role annotations (White et al. 2020) to explore the connections between control and labeling event participants with properties typically associated with agents and patients.

**RLPrompt: Optimizing Discrete Text Prompts with Reinforcement Learning**
*Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing and Zhiting Hu* 11:00-12:30 (Atrium)
Prompting has shown impressive success in enabling large pre-trained language models (LMs) to perform diverse NLP tasks, especially with only few downstream data. Automatically finding the optimal prompt for each task, however, is challenging. Most existing work resorts to tuning *soft* prompts (e.g., embeddings) which fall short of interpretability, reusability across LMs, and applicability when gradients are not accessible. *Discrete* prompts, on the other hand, are difficult to optimize, and are often created by "enumeration (e.g., paraphrasing)-then-selection" heuristics that do not explore the prompt space systematically. This paper proposes RLPrompt, an efficient discrete prompt optimization approach with reinforcement learning (RL). RLPrompt formulates a parameter-efficient policy network that generates the opti-

mized discrete prompt after training with reward. To harness the complex and stochastic reward signals from the large LM environment, we incorporate effective reward stabilization that substantially enhances training efficiency. RLPrompt is flexibly applicable to different types of LMs, such as masked (e.g., BERT) and left-to-right models (e.g., GPTs), for both classification and generation tasks. Experiments on few-shot classification and unsupervised text style transfer show superior performance over a wide range of existing fine-tuning or prompting methods. Interestingly, the resulting optimized prompts are often ungrammatical gibberish text; and surprisingly, those gibberish prompts are transferrable between different LMs to retain significant performance, indicating that LM prompting may not follow human language patterns.

**Natural Language to Code Translation with Execution**
*Freda Shi, Daniel Fried, Marjan Ghazvininejad, Luke Zettlemoyer and Sida I. Wang* 11:00-12:30 (Atrium)
Generative models of code, pretrained on large corpora of programs, have shown great success in translating natural language to code (Chen et al., 2021; Austin et al., 2021; Li et al., 2022, inter alia). While these models do not explicitly incorporate program semantics (i.e., execution results) during training, they are able to generate correct solutions for many problems. However, choosing a single correct program from a generated set for each problem remains challenging. In this work, we introduce execution result–based minimum Bayes risk decoding (MBR-EXEC) for program selection and show that it improves the few-shot performance of pretrained code models on natural-language-to-code tasks. We select output programs from a generated candidate set by marginalizing over program implementations that share the same semantics. Because exact equivalence is intractable, we execute each program on a small number of test inputs to approximate semantic equivalence. Across datasets, execution or simulated execution significantly outperforms the methods that do not involve program semantics. We find that MBR-EXEC consistently improves over all execution-unaware selection methods, suggesting it as an effective approach for natural language to code translation.

[CL] **Neural Embedding Allocation: Distributed Representations of Topic Models**
*Kamrun Naher Keya, Yannis Papanikolaou and James R. Foulds* 11:00-12:30 (Atrium)
We propose a method which uses neural embeddings to improve the performance of any given LDA-style topic model. Our method, called neural embedding allocation (NEA), deconstructs topic models (LDA or otherwise) into interpretable vector-space embeddings of words, topics, documents, authors, and so on, by learning neural embeddings to mimic the topic model. We demonstrate that NEA improves coherence scores of the original topic model by smoothing out the noisy topics when the number of topics is large. Furthermore, we show NEA's effectiveness and generality in deconstructing and smoothing LDA, author-topic models, and the recent mixed membership skip-gram topic model and achieve better performance with the embeddings compared to several state-of-the-art models.

**Chunk-based Nearest Neighbor Machine Translation**
*Pedro Henrique Martins, Zita Marinho and André F. T. Martins* 11:00-12:30 (Atrium)
Semi-parametric models, which augment generation with retrieval, have led to impressive results in language modeling and machine translation, due to their ability to retrieve fine-grained information from a datastore of examples. One of the most prominent approaches, kNN-MT, exhibits strong domain adaptation capabilities by retrieving tokens from domain-specific datastores (Khandelwal et al., 2021). However, kNN-MT requires an expensive retrieval operation for every single generated token, leading to a very low decoding speed (around 8 times slower than a parametric model). In this paper, we introduce a chunk-based kNN-MT model which retrieves chunks of tokens from the datastore, instead of a single token. We propose several strategies for incorporating the retrieved chunks into the generation process, and for selecting the steps at which the model needs to search for neighbors in the datastore. Experiments on machine translation in two settings, static and "on-the-fly" domain adaptation, show that the chunk-based kNN-MT model leads to significant speed-ups (up to 4 times) with only a small drop in translation quality.

**ConNER: Consistency Training for Cross-lingual Named Entity Recognition**
*Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, Luo Si and Chunyan Miao* 11:00-12:30 (Atrium)
Cross-lingual named entity recognition (NER) suffers from data scarcity in the target languages, especially under zero-shot settings. Existing translate-train or knowledge distillation methods attempt to bridge the language gap, but often introduce a high level of noise. To solve this problem, consistency training methods regularize the model to be robust towards perturbations on data or hidden states. However, such methods are likely to violate the consistency hypothesis, or mainly focus on coarse-grain consistency. We propose ConNER as a novel consistency training framework for cross-lingual NER, which comprises of: (1) translation-based consistency training on unlabeled target-language data, and (2) dropout-based consistency training on labeled source-language data. ConNER effectively leverages unlabeled target-language data and alleviates overfitting on the source language to enhance the cross-lingual adaptability. Experimental results show our ConNER achieves consistent improvement over various baseline methods.

**Transforming Sequence Tagging Into A Seq2Seq Task**
*Karthik Raman, Iftekhar Naim, Jiecao Chen, Kazuma Hashimoto, Kiran Yalasangi and Krishna Srinivasan* 11:00-12:30 (Atrium)
Pretrained, large, generative language models (LMs) have had great success in a wide range of sequence tagging and structured prediction tasks. Casting a sequence tagging task as a Seq2Seq one requires deciding the formats of the input and output sequences. However, we lack a principled understanding of the trade-offs associated with these formats (such as the effect on model accuracy, sequence length, multilingual generalization, hallucination). In this paper, we rigorously study different formats one could use for casting input text sentences and their output labels into the input and target (i.e., output) of a Seq2Seq model. Along the way, we introduce a new format, which we show to to be both simpler and more effective. Additionally the new format demonstrates significant gains in the multilingual settings – both zero-shot transfer learning and joint training. Lastly, we find that the new format is more robust and almost completely devoid of hallucination – an issue we find common in existing formats. With well over a 1000 experiments studying 14 different formats, over 7 diverse public benchmarks – including 3 multilingual datasets spanning 7 languages – we believe our findings provide a strong empirical basis in understanding how we should tackle sequence tagging tasks.

[INDUSTRY] **Accelerating the Discovery of Semantic Associations from Medical Literature: Mining Relations Between Diseases and Symptoms**
*Alberto Purpura, Francesca Bonin and Joao H. Bettencourt-Silva* 11:00-12:30 (Atrium)
Medical literature is a vast and constantly expanding source of information about diseases, their diagnoses and treatments. One of the ways to extract insights from this type of data is through mining association rules between such entities. However, existing solutions do not take into account the semantics of sentences from which entity co-occurrences are extracted. We propose a scalable solution for the automated discovery of semantic associations between different entities such as diseases and their symptoms. Our approach employs the UMLS semantic network and a binary relation classification model trained with distant supervision to validate and help ranking the most likely entity associations pairs extracted with frequency-based association rule mining algorithms. We evaluate the proposed system on the task of extracting disease-symptom associations from a collection of over 14M PubMed abstracts and validate our results against a publicly available known list of disease-symptom pairs.

[INDUSTRY] **Machine translation impact in E-commerce multilingual search**
*Hang Zhang and amita misra* 11:00-12:30 (Atrium)

Previous work suggests that performance of cross-lingual information retrieval correlates highly with the quality of Machine Translation. However, there may be a threshold beyond which improving query translation quality yields little or no benefit to further improve the retrieval performance. This threshold may depend upon multiple factors including the source and target languages, the existing MT system quality and the search pipeline. In order to identify the benefit of improving an MT system for a given search pipeline, we investigate the sensitivity of retrieval quality to the presence of different levels of MT quality using experimental datasets collected from actual traffic. We systematically improve the performance of our MT systems quality on language pairs as measured by MT evaluation metrics including Bleu and Chrf to determine their impact on search precision metrics and extract signals that help to guide the improvement strategies. Using this information we develop techniques to compare query translations for multiple language pairs and identify the most promising language pairs to invest and improve.

[INDUSTRY] **Exploiting In-Domain Bilingual Corpora for Zero-Shot Transfer Learning in NLU of Intra-Sentential Code-Switching Chatbot Interactions**
*Maia Aguirre, Manex Serras, Laura García-Sardiña, Jacobo López-Fernández, Ariane Méndez and Arantza del Pozo* 11:00-12:30 (Atrium)
Code-switching (CS) is a very common phenomenon in regions with various co-existing languages. Since CS is such a frequent habit in informal communications, both spoken and written, it also arises naturally in Human-Machine Interactions. Therefore, in order for natural language understanding (NLU) not to be degraded, CS must be taken into account when developing chatbots. The co-existence of multiple languages in a single NLU model has become feasible with multilingual language representation models such as mBERT. In this paper, the efficacy of zero-shot cross-lingual transfer learning with mBERT for NLU is evaluated on a Basque-Spanish CS chatbot corpus, comparing the performance of NLU models trained using in-domain chatbot utterances in Basque and/or Spanish without CS. The results obtained indicate that training joint multi-intent classification and entity recognition models on both languages simultaneously achieves best performance, better capturing the CS patterns.

[INDUSTRY] **Calibrating Imbalanced Classifiers with Focal Loss: An Empirical Study**
*Cheng Wang, Jorge Balazs, György Szarvas, Patrick Ernst, Lahari Poddar and Pavel Danchenko* 11:00-12:30 (Atrium)
Imbalanced data distribution is a practical and common challenge in building production-level machine learning (ML) models in industry, where data usually exhibits long-tail distributions. For instance, in virtual AI Assistants, such as Google Assistant, Amazon Alexa and Apple Siri, the "play music" or "set timer" utterance is exposed to an order of magnitude more traffic than other skills. This can easily cause trained models to overfit to the majority classes, categories or intents, lead to model miscalibration. The uncalibrated models output unreliable (mostly overconfident) predictions, which are at high risk of affecting downstream decision-making systems. In this work, we study the calibration of production models in the industry use-case of predicting product return reason codes in customer service conversations of an online retail store; The returns reasons also exhibit class imbalance. To alleviate the resulting miscalibration in the production ML model, we streamline the model development and deployment using focal loss [**lin2017focal**]. We empirically show the effectiveness of model training with focal loss in learning better calibrated models, as compared to standard cross-entropy loss. Better calibration, in turn, enables better control of the precision-recall trade-off for the models deployed in production.

[INDUSTRY] **Unsupervised training data re-weighting for natural language understanding with local distribution approximation**
*Jose Garrido Ramas, Dieu-Thu Le, Bei Chen, Manoj Kumar and Kay Rottmann* 11:00-12:30 (Atrium)
One of the major challenges of training Natural Language Understanding (NLU) production models lies in the discrepancy between the distributions of the offline training data and of the online live data, due to, e.g., biased sampling scheme, cyclic seasonality shifts, annotated training data coming from a variety of different sources, and a changing pool of users. Consequently, the model trained by the offline data is biased. We often observe this problem especially in task-oriented conversational systems, where topics of interest and the characteristics of users using the system change over time. In this paper we propose an unsupervised approach to mitigate the offline training data sampling bias in multiple NLU tasks. We show that a local distribution approximation in the pre-trained embedding space enables the estimation of importance weights for training samples guiding re-sampling for an effective bias mitigation. We illustrate our novel approach using multiple NLU datasets and show improvements obtained without additional annotation, making this a general approach for mitigating effects of sampling bias.

[INDUSTRY] **Cross-Encoder Data Annotation for Bi-Encoder Based Product Matching**
*Justin Chiu and Keiji Shinzato* 11:00-12:30 (Atrium)
Matching a seller listed item to an appropriate product is an important step for an e-commerce platform. With the recent advancement in deep learning, there are different encoder based approaches being proposed as solution. When textual data for two products are available, cross-encoder approaches encode them jointly while bi-encoder approaches encode them separately. Since cross-encoders are computationally heavy, approaches based on bi-encoders are a common practice for this challenge. In this paper, we propose cross-encoder data annotation; a technique to annotate or refine human annotated training data for bi-encoder models using a cross-encoder model. This technique enables us to build a robust model without annotation on newly collected training data or further improve model performance on annotated training data. We evaluate the cross-encoder data annotation on the product matching task using a real-world e-commerce dataset containing 104 million products. Experimental results show that the cross-encoder data annotation improves 4% absolute accuracy when no annotation for training data is available, and 2% absolute accuracy when annotation for training data is available.

[INDUSTRY] **Multi-Tenant Optimization For Few-Shot Task-Oriented FAQ Retrieval**
*Asha Vishwanathan, Rajeev Unnikrishnan Warrier, Gautham Vadakkekara Suresh and Chandra Shekhar Kandpal* 11:00-12:30 (Atrium)
Business-specific Frequently Asked Questions (FAQ) retrieval in task-oriented dialog systems poses unique challenges vis à vis community based FAQs. Each FAQ question represents an intent which is usually an umbrella term for many related user queries. We evaluate performance for such Business FAQs both with standard FAQ retrieval techniques using query-Question (q-Q) similarity and few-shot intent detection techniques. Implementing a real-world solution for FAQ retrieval in order to support multiple tenants (FAQ sets) entails optimizing speed, accuracy and cost. We propose a novel approach to scale multi-tenant FAQ applications in real-world context by contrastive fine-tuning of the last layer in sentence Bi-Encoders along with tenant-specific weight switching.

**Life is a Circus and We are the Clowns: Automatically Finding Analogies between Situations and Processes**
*Oren Sultan and Dafna Shahaf* 11:00-12:30 (Atrium)
Analogy-making gives rise to reasoning, abstraction, flexible categorization and counterfactual inference – abilities lacking in even the best AI systems today. Much research has suggested that analogies are key to non-brittle systems that can adapt to new domains. Despite their importance, analogies received little attention in the NLP community, with most research focusing on simple word analogies. Work that tackled more complex analogies relied heavily on manually constructed, hard-to-scale input representations. In this work, we explore a more realistic, challenging setup: our input is a pair of natural language procedural texts, describing a situation or a process (e.g., how the heart works/how a pump works). Our goal is to automatically extract entities and their relations from the text and find a mapping between the different domains based on relational similarity (e.g., blood is mapped to water).

We develop an interpretable, scalable algorithm and demonstrate that it identifies the correct mappings 87% of the time for procedural texts and 94% for stories from cognitive-psychology literature. We show it can extract analogies from a large dataset of procedural texts, achieving

79% precision (analogy prevalence in data: 3%). Lastly, we demonstrate that our algorithm is robust to paraphrasing the input texts

**Automatic Generation of Socratic Subquestions for Teaching Math Word Problems**
*Kumar Shridhar, Jakub Macina, Mennatallah El-Assady, Tanmay Sinha, Manu Kapur and Mrinmaya Sachan* 11:00-12:30 (Atrium)
Socratic questioning is an educational method that allows students to discover answers to complex problems by asking them a series of thoughtful questions. Generation of didactically sound questions is challenging, requiring understanding of the reasoning process involved in the problem. We hypothesize that such questioning strategy can not only enhance the human performance, but also assist the math word problem (MWP) solvers. In this work, we explore the ability of large language models (LMs) in generating sequential questions for guiding math word problem-solving. We propose various guided question generation schemes based on input conditioning and reinforcement learning. On both automatic and human quality evaluations, we find that LMs constrained with desirable question properties generate superior questions and improve the overall performance of a math word problem solver. We conduct a preliminary user study to examine the potential value of such question generation models in the education domain. Results suggest that the difficulty level of problems plays an important role in determining whether questioning improves or hinders human performance. We discuss the future of using such questioning strategies in education.

**Factual Accuracy is not Enough: Planning Consistent Description Order for Radiology Report Generation**
*Toru Nishino, Yasuhide Miura, Tomoki Taniguchi, Tomoko Ohkuma, Yuki Suzuki, Shoji Kido and Noriyuki Tomiyama* 11:00-12:30 (Atrium)
Radiology report generation systems have the potential to reduce the workload of radiologists by automatically describing the findings in medical images. To broaden the application of the report generation system, the system should generate reports that are not only factually accurate but also chronologically consistent, describing images that are presented in time order, that is, the correct order. We employ a planning-based radiology report generation system that generates the overall structure of reports as "plans'" prior to generating reports that are accurate and consistent in order. Additionally, we propose a novel reinforcement learning and inference method, Coordinated Planning (CoPlan), that includes a content planner and a text generator to train and infer in a coordinated manner to alleviate the cascading of errors that are often inherent in planning-based models. We conducted experiments with single-phase diagnostic reports in which the factual accuracy is critical and multi-phase diagnostic reports in which the description order is critical. Our proposed CoPlan improves the content order score by 5.1 pt in time series critical scenarios and the clinical factual accuracy F-score by 9.1 pt in time series irrelevant scenarios, compared those of the baseline models without CoPlan.

**Differentially Private Language Models for Secure Data Sharing**
*Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schoelkopf and Mrinmaya Sachan* 11:00-12:30 (Atrium)
To protect the privacy of individuals whose data is being shared, it is of high importance to develop methods allowing researchers and companies to release textual data while providing formal privacy guarantees to its originators. In the field of NLP, substantial efforts have been directed at building mechanisms following the framework of local differential privacy, thereby anonymizing individual text samples before releasing them. In practice, these approaches are often dissatisfying in terms of the quality of their output language due to the strong noise required for local differential privacy. In this paper, we approach the problem at hand using global differential privacy, particularly by training a generative language model in a differentially private manner and consequently sampling data from it. Using natural language prompts and a new prompt-mismatch loss, we are able to create highly accurate and fluent textual datasets taking on specific desired attributes such as sentiment or topic and resembling statistical properties of the training data. We perform thorough experiments indicating that our synthetic datasets do not leak information from our original data and are of high language quality and highly suitable for training models for further analysis on real-world data. Notably, we also demonstrate that training classifiers on private synthetic data outperforms directly training classifiers with DP-SGD.

**Hard Gate Knowledge Distillation - Leverage Calibration for Robust and Reliable Language Model**
*Dongkyu Lee, Zhiliang Tian, Yingxiu Zhao, Ka Chun Cheung and Nevin Zhang* 11:00-12:30 (Atrium)
In knowledge distillation, a student model is trained with supervisions from both knowledge from a teacher and observations drawn from a training data distribution. Knowledge of a teacher is considered a subject that holds inter-class relations which send a meaningful supervision to a student; hence, much effort has been put to find such knowledge to be distilled. In this paper, we explore a question that has been given little attention: "when to distill such knowledge." The question is answered in our work with the concept of model calibration; we view a teacher model not only as a source of knowledge but also as a gauge to detect miscalibration of a student. This simple and yet novel view leads to a hard gate knowledge distillation scheme that switches between learning from a teacher model and training data. We verify the gating mechanism in the context of natural language generation at both the token-level and the sentence-level. Empirical comparisons with strong baselines show that hard gate knowledge distillation not only improves model generalization, but also significantly lowers model calibration error.

**Improving Iterative Text Revision by Learning Where to Edit from Other Revision Tasks**
*Zae Myung Kim, Wanyu Du, Vipul Raheja, Dhruv Kumar and Dongyeop Kang* 11:00-12:30 (Atrium)
Iterative text revision improves text quality by fixing grammatical errors, rephrasing for better readability or contextual appropriateness, or reorganizing sentence structures throughout a document. Most recent research has focused on understanding and classifying different types of edits in the iterative revision process from human-written text instead of building accurate and robust systems for iterative text revision. In this work, we aim to build an end-to-end text revision system that can iteratively generate helpful edits by explicitly detecting editable spans (where-to-edit) with their corresponding edit intents and then instructing a revision model to revise the detected edit spans. Leveraging datasets from other related text editing NLP tasks, combined with the specification of editable spans, leads our system to more accurately model the process of iterative text refinement, as evidenced by empirical results and human evaluations. Our system significantly outperforms previous baselines on our text revision tasks and other standard text revision tasks, including grammatical error correction, text simplification, sentence fusion, and style transfer. Through extensive qualitative and quantitative analysis, we make vital connections between edit intentions and writing quality, and better computational modeling of iterative text revisions.

**FiE: Building a Global Probability Space by Leveraging Early Fusion in Encoder for Open-Domain Question Answering**
*Akhil Kedia, Mohd Abbas Zaidi and Haejun Lee* 11:00-12:30 (Atrium)
Generative models have recently started to outperform extractive models in Open Domain Question Answering, largely by leveraging their decoder to attend over multiple encoded passages and combining their information. However, generative models tend to be larger than extractive models due to the need for a decoder, run slower during inference due to auto-regressive decoder beam search, and their generated output often suffers from hallucinations. We propose to extend transformer encoders with the ability to fuse information from multiple passages, using global representation to provide cross-sample attention over all tokens across samples. Furthermore, we propose an alternative answer span probability calculation to better aggregate answer scores in the global space of all samples. Using our proposed method, we outperform the current state-of-the-art method by 2.5 Exact Match score on the Natural Question dataset while using only 25% of parameters and 35% of the latency during inference, and 4.4 Exact Match on WebQuestions dataset. When coupled with synthetic data augmentation, we outperform larger models on the TriviaQA dataset as well. The latency and parameter savings of our method make it particularly attractive for open-domain question answering, as these models are often compute-intensive.

**monoQA: Multi-Task Learning of Reranking and Answer Extraction for Open-Retrieval Conversational Question Answering**
*Sarawoot Kongyoung, Craig Macdonald and Iadh Ounis* 11:00-12:30 (Atrium)
To address the Conversational Question Answering (ORConvQA) task, previous work has considered an effective three-stage architecture, consisting of a retriever, a reranker, and a reader to extract the answers. In order to effectively answer the users' questions, a number of existing approaches have applied multi-task learning, such that the same model is shared between the reranker and the reader. Such approaches also typically tackle reranking and reading as classification tasks. On the other hand, recent text generation models, such as monoT5 and UnifiedQA, have been shown to respectively yield impressive performances in passage reranking and reading. However, no prior work has combined monoT5 and UnifiedQA to share a single text generation model that directly extracts the answers for the users instead of predicting the start/end positions in a retrieved passage. In this paper, we investigate the use of Multi-Task Learning (MTL) to improve performance on the ORConvQA task by sharing the reranker and reader's learned structure in a generative model. In particular, we propose monoQA, which uses a text generation model with multi-task learning for both the reranker and reader. Our model, which is based on the T5 text generation model, is fine-tuned simultaneously for both reranking (in order to improve the precision of the top retrieved passages) and extracting the answer. Our results on the OR-QuAC and OR-CoQA datasets demonstrate the effectiveness of our proposed model, which significantly outperforms existing strong baselines with improvements ranging from +12.31% to +19.51% in MAP and from +5.70% to +23.34% in F1 on all used test sets.

**Empowering Language Models with Knowledge Graph Reasoning for Open-Domain Question Answering**
*Ziniu Hu, Yichong Xu, Wenhao Yu, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Kai-Wei Chang and Yizhou Sun* 11:00-12:30 (Atrium)
Answering open-domain questions requires world knowledge about in-context entities. As pre-trained Language Models (LMs) lack the power to store all required knowledge, external knowledge sources, such as knowledge graphs, are often used to augment LMs. In this work, we propose knOwledge REasOning empowered Language Model (OREO-LM), which consists of a novel Knowledge Interaction Layer that can be flexibly plugged into existing Transformer-based LMs to interact with a differentiable Knowledge Graph Reasoning module collaboratively. In this way, LM guides KG to walk towards the desired answer, while the retrieved knowledge improves LM. By adopting OREO-LM to RoBERTa and T5, we show significant performance gain, achieving state-of-art results in the Closed-Book setting. The performance enhancement is mainly from the KG reasoning's capacity to infer missing relational facts. In addition, OREO-LM provides reasoning paths as rationales to interpret the model's decision.

**FigMemes: A Dataset for Figurative Language Identification in Politically-Opinionated Memes**
*Chen Liu, Gregor Geigle, Robin Krebs and Iryna Gurevych* 11:00-12:30 (Atrium)
Real-world politically-opinionated memes often rely on figurative language to cloak propaganda and radical ideas to help them spread. It is not only a scientific challenge to develop machine learning models to recognize them in memes, but also sociologically beneficial to understand hidden meanings at scale and raise awareness. These memes are fast-evolving (in both topics and visuals) and it remains unclear whether current multimodal machine learning models are robust to such distribution shifts. To enable future research into this area, we first present FigMemes, a dataset for figurative language classification in politically-opinionated memes. We evaluate the performance of state-of-the-art unimodal and multimodal models and provide comprehensive benchmark results. The key contributions of this proposed dataset include annotations of six commonly used types of figurative language in politically-opinionated memes, and a wide range of topics and visual styles. We also provide analyses on the ability of multimodal models to generalize across distribution shifts in memes. Our dataset poses unique machine learning challenges and our results show that current models have significant room for improvement in both performance and robustness to distribution shifts.

**Detecting Label Errors by Using Pre-Trained Language Models**
*Derek Chong, Jenny Hong and Christopher Manning* 11:00-12:30 (Atrium)
We show that large pre-trained language models are inherently highly capable of identifying label errors in natural language datasets: simply examining out-of-sample data points in descending order of fine-tuned task loss significantly outperforms more complex error-detection mechanisms proposed in previous work. To this end, we contribute a novel method for introducing realistic, human-originated label noise into existing crowdsourced datasets such as SNLI and TweetNLP. We show that this noise has similar properties to real, hand-verified label errors, and is harder to detect than existing synthetic noise, creating challenges for model robustness. We argue that human-originated noise is a better standard for evaluation than synthetic noise. Finally, we use crowdsourced verification to evaluate the detection of real errors on IMDB, Amazon Reviews, and Recon, and confirm that pre-trained models perform at a 9–36% higher absolute Area Under the Precision-Recall Curve than existing models.

**Evaluating the Knowledge Dependency of Questions**
*Hyeongdon Moon, Yoonseok Yang, Hangyeol Yu, Seunghyun Lee, Myeongho Jeong, juneyoung park, Jamin Shin, Minsam Kim and Seungtaek Choi* 11:00-12:30 (Atrium)
The automatic generation of Multiple Choice Questions (MCQ) has the potential to reduce the time educators spend on student assessment significantly. However, existing evaluation metrics for MCQ generation, such as BLEU, ROUGE, and METEOR, focus on the n-gram based similarity of the generated MCQ to the gold sample in the dataset and disregard their educational value. They fail to evaluate the MCQ's ability to assess the student's knowledge of the corresponding target fact. To tackle this issue, we propose a novel automatic evaluation metric, coined Knowledge Dependent Answerability (KDA), which measures the MCQ's answerability given knowledge of the target fact. Specifically, we first show how to measure KDA based on student responses from a human survey. Then, we propose two automatic evaluation metrics, KDA_disc and KDA_cont, that approximate KDA by leveraging pre-trained language models to imitate students' problem-solving behavior. Through our human studies, we show that KDA_disc and KDA_soft have strong correlations with both (1) KDA and (2) usability in an actual classroom setting, labeled by experts. Furthermore, when combined with n-gram based similarity metrics, KDA_disc and KDA_cont are shown to have a strong predictive power for various expert-labeled MCQ quality measures.

**On the Limitations of Reference-Free Evaluations of Generated Text**
*Daniel Deutsch, Rotem Dror and Dan Roth* 11:00-12:30 (Atrium)
There is significant interest in developing evaluation metrics which accurately estimate the quality of generated text without the aid of a human-written reference text, which can be time consuming and expensive to collect or entirely unavailable in online applications. However, in this work, we demonstrate that these reference-free metrics are inherently biased and limited in their ability to evaluate generated text, and we argue that they should not be used to measure progress on tasks like machine translation or summarization. We show how reference-free metrics are equivalent to using one generation model to evaluate another, which has several limitations: (1) the metrics can be optimized at test time to find the approximate best-possible output, (2) they are inherently biased toward models which are more similar to their own, and (3) they can be biased against higher-quality outputs, including those written by humans. Therefore, we recommend that reference-free metrics should be used as diagnostic tools for analyzing and understanding model behavior instead of measures of how well models perform a task, in which the goal is to achieve as high of a score as possible.

**Three Real-World Datasets and Neural Computational Models for Classification Tasks in Patent Landscaping**

*Subhash Pujari, Jannik Strötgen, Mark Giereth, Michael Gertz and Annemarie Friedrich* 11:00-12:30 (Atrium)
Patent Landscaping, one of the central tasks of intellectual property management, includes selecting and grouping patents according to user-defined technical or application-oriented criteria. While recent transformer-based models have been shown to be effective for classifying patents into taxonomies such as CPC or IPC, there is yet little research on how to support real-world Patent Landscape Studies (PLSs) using natural language processing methods. With this paper, we release three labeled datasets for PLS-oriented classification tasks covering two diverse domains. We provide a qualitative analysis and report detailed corpus statistics.

Most research on neural models for patents has been restricted to leveraging titles and abstracts. We compare strong neural and non-neural baselines, proposing a novel model that takes into account textual information from the patents' full texts as well as embeddings created based on the patents' CPC labels. We find that for PLS-oriented classification tasks, going beyond title and abstract is crucial, CPC labels are an effective source of information, and combining all features yields the best results.

**Natural Language Deduction with Incomplete Information**
*Zayne Sprague, Kaj Bostrom, Swarat Chaudhuri and Greg Durrett* 11:00-12:30 (Atrium)
A growing body of work studies how to answer a question or verify a claim by generating a natural language "proof:" a chain of deductive inferences yielding the answer based on a set of premises. However, these methods can only make sound deductions when they follow from evidence that is given. We propose a new system that can handle the underspecified setting where not all premises are stated at the outset; that is, additional assumptions need to be materialized to prove a claim. By using a natural language generation model to abductively infer a premise given another premise and a conclusion, we can impute missing pieces of evidence needed for the conclusion to be true. Our system searches over two fringes in a bidirectional fashion, interleaving deductive (forward-chaining) and abductive (backward-chaining) generation steps. We sample multiple possible outputs for each step to achieve coverage of the search space, at the same time ensuring correctness by filtering low-quality generations with a round-trip validation procedure. Results on a modified version of the EntailmentBank dataset and a new dataset called Everyday Norms: Why Not? Show that abductive generation with validation can recover premises across in- and out-of-domain settings.

**Evaluating the Impact of Model Scale for Compositional Generalization in Semantic Parsing**
*Linlu Qiu, Peter Shaw, Panupong Pasupat, Tianze Shi, Jonathan Herzig, Emily Pitler, Fei Sha and Kristina Toutanova* 11:00-12:30 (Atrium)
Despite their strong performance on many tasks, pre-trained language models have been shown to struggle on out-of-distribution compositional generalization. Meanwhile, recent work has shown considerable improvements on many NLP tasks from model scaling. Can scaling up model size also improve compositional generalization in semantic parsing? We evaluate encoder-decoder models up to 11B parameters and decoder-only models up to 540B parameters, and compare model scaling curves for three different methods for applying a pre-trained language model to a new task: fine-tuning all parameters, prompt tuning, and in-context learning. We observe that fine-tuning generally has flat or negative scaling curves on out-of-distribution compositional generalization in semantic parsing evaluations. In-context learning has positive scaling curves, but is generally outperformed by much smaller fine-tuned models. Prompt-tuning can outperform fine-tuning, suggesting further potential improvements from scaling as it exhibits a more positive scaling curve. Additionally, we identify several error trends that vary with model scale. For example, larger models are generally better at modeling the syntax of the output space, but are also more prone to certain types of overfitting. Overall, our study highlights limitations of current techniques for effectively leveraging model scale for compositional generalization, while our analysis also suggests promising directions for future work.

**Mitigating Spurious Correlation in Natural Language Understanding with Counterfactual Inference**
*Can Udomcharoenchaikit, Wuttikorn Ponwitayarat, Patomporn Payoungkhamdee, Kanruethai Masuk, Weerayut Buaphet, Ekapol Chuangsuwanich and Sarana Nutanong* 11:00-12:30 (Atrium)
Despite their promising results on standard benchmarks, NLU models are still prone to make predictions based on shortcuts caused by unintended bias in the dataset. For example, an NLI model may use lexical overlap as a shortcut to make entailment predictions due to repetitive data generation patterns from annotators, also called annotation artifacts. In this paper, we propose a causal analysis framework to help debias NLU models. We show that (1) by defining causal relationships, we can introspect how much annotation artifacts affect the outcomes. (2) We can utilize counterfactual inference to mitigate bias with this knowledge. We found that viewing a model as a treatment can mitigate bias more effectively than viewing annotation artifacts as treatment. (3) In addition to bias mitigation, we can interpret how much each debiasing strategy is affected by annotation artifacts. Our experimental results show that using counterfactual inference can improve out-of-distribution performance in all settings while maintaining high in-distribution performance.

**TRIPS: Efficient Vision-and-Language Pre-training with Text-Relevant Image Patch Selection**
*Chaoya Jiang, Haiyang Xu, Chenliang Li, Ming Yan, Wei Ye, Shikun Zhang, Bin Bi and Songfang Huang* 11:00-12:30 (Atrium)
Vision Transformers (ViTs) have been widely used in large-scale Vision and Language Pre-training (VLP) models. Though previous VLP works have proved the effectiveness of ViTs, they still suffer from computational efficiency brought by the long visual sequence. To tackle this problem, in this paper, we propose an efficient vision-and-language pre-training model with Text-Relevant Image Patch Selection, namely TRIPS, which reduces the visual sequence progressively with a text-guided patch-selection layer in the visual backbone for efficient training and inference. The patch-selection layer can dynamically compute text-dependent visual attention to identify the attentive image tokens with text guidance and fuse inattentive ones in an end-to-end manner. Meanwhile, TRIPS does not introduce extra parameters to ViTs. Experimental results on a variety of popular benchmark datasets demonstrate that TRIPS gain a speedup of 40% over previous similar VLP models, yet with competitive or better downstream task performance.

**Adaptive Contrastive Learning on Multimodal Transformer for Review Helpfulness Prediction**
*Thong Nguyen, Xiaobao Wu, Anh Tuan Luu, Zhen Hai and Lidong Bing* 11:00-12:30 (Atrium)
Modern Review Helpfulness Prediction systems are dependent upon multiple modalities, typically texts and images. Unfortunately, those contemporary approaches pay scarce attention to polish representations of cross-modal relations and tend to suffer from inferior optimization. This might cause harm to model's predictions in numerous cases. To overcome the aforementioned issues, we propose Multi-modal Contrastive Learning for Multimodal Review Helpfulness Prediction (MRHP) problem, concentrating on mutual information between input modalities to explicitly elaborate cross-modal relations. In addition, we introduce Adaptive Weighting scheme for our contrastive learning approach in order to increase flexibility in optimization. Lastly, we propose Multimodal Interaction module to address the unalignment nature of multimodal data, thereby assisting the model in producing more reasonable multimodal representations. Experimental results show that our method outperforms prior baselines and achieves state-of-the-art results on two publicly available benchmark datasets for MRHP problem.

**Mutual Information Alleviates Hallucinations in Abstractive Summarization**
*Liam van der Poel, Clara Meister and Ryan Cotterell* 11:00-12:30 (Atrium)
Despite significant progress in the quality of language generated from abstractive summarization models, these models still exhibit the tendency to hallucinate, i.e., output content not supported by the source document. A number of works have tried to fix—or at least uncover the source of—the problem with limited success. In this paper, we identify a simple criterion under which models are significantly more likely to assign more probability to hallucinated content during generation: high model uncertainty. This finding offers a potential explanation for hallucinations: models default to favoring text with high marginal probability, i.e., high-frequency occurrences in the training set, when

uncertain about a continuation. It also motivates possible routes for real-time intervention during decoding to prevent such hallucinations. We propose a decoding strategy that switches to optimizing for pointwise mutual information of the source and target token—rather than purely the probability of the target token—when the model exhibits uncertainty. Experiments on the xsum dataset show that our method decreases the probability of hallucinated tokens while maintaining the Rouge and BERT-S scores of top-performing decoding strategies.

**Salience Allocation as Guidance for Abstractive Summarization**
*Fei Wang, Kaiqiang Song, Hongming Zhang, Lifeng Jin, Sangwoo Cho, Wenlin Yao, Xiaoyang Wang, Muhao Chen and Dong Yu* 11:00-12:30 (Atrium)
Abstractive summarization models typically learn to capture the salient information from scratch implicitly. Recent literature adds extractive summaries as guidance for abstractive summarization models to provide hints of salient content and achieves better performance. However, extractive summaries as guidance could be over strict, leading to information loss or noisy signals. Furthermore, it cannot easily adapt to documents with various abstractiveness. As the number and allocation of salience content pieces varies, it is hard to find a fixed threshold deciding which content should be included in the guidance. In this paper, we propose a novel summarization approach with a flexible and reliable salience guidance, namely SEASON (SaliencE Allocation as Guidance for Abstractive SummarizatiON). SEASON utilizes the allocation of salience expectation to guide abstractive summarization and adapts well to articles in different abstractiveness. Automatic and human evaluations on two benchmark datasets show that the proposed method is effective and reliable. Empirical results on more than one million news articles demonstrate a natural fifteen-fifty salience split for news article sentences, providing a useful insight for composing news articles.

**Improving Factual Consistency in Summarization with Compression-Based Post-Editing**
*Alex Fabbri, Prafulla Kumar Choubey, Jesse Vig, Chien-Sheng Wu and caiming xiong* 11:00-12:30 (Atrium)
State-of-the-art summarization models still struggle to be factually consistent with the input text. A model-agnostic way to address this problem is post-editing the generated summaries. However, existing approaches typically fail to remove entity errors if a suitable input entity replacement is not available or may insert erroneous content. In our work, we focus on removing extrinsic entity errors, or entities not in the source, to improve consistency while retaining the summary's essential information and form. We propose to use sentence-compression data to train the post-editing model to take a summary with extrinsic entity errors marked with special tokens and output a compressed, well-formed summary with those errors removed. We show that this model improves factual consistency while maintaining ROUGE, improving entity precision by up to 30% on XSum, and that this model can be applied on top of another post-editor, improving entity precision by up to a total of 38%. We perform an extensive comparison of post-editing approaches that demonstrate trade-offs between factual consistency, informativeness, and grammaticality, and we analyze settings where post-editors show the largest improvements.

**Learning with Rejection for Abstractive Text Summarization**
*Meng Cao, Yue Dong, Jingyi He and Jackie Chi Kit Cheung* 11:00-12:30 (Atrium)
State-of-the-art abstractive summarization systems frequently hallucinate content that is not supported by the source document, mainly due to noise in the training dataset. Existing methods opt to drop the noisy samples or tokens from the training set entirely, reducing the effective training set size and creating an artificial propensity to copy words from the source. In this work, we propose a training objective for abstractive summarization based on rejection learning, in which the model learns whether or not to reject potentially noisy tokens. We further propose a regularized decoding objective that penalizes non-factual candidate summaries during inference by using the rejection probability learned during training. We show that our method considerably improves the factuality of generated summaries in automatic and human evaluations when compared to five baseline models, and that it does so while increasing the abstractiveness of the generated summaries.

**Fast-R2D2: A Pretrained Recursive Neural Network based on Pruned CKY for Grammar Induction and Text Representation**
*Xiang Hu, Haitao Mi, Liang Li and Gerard de Melo* 11:00-12:30 (Atrium)
Chart-based models have shown great potential in unsupervised grammar induction, running recursively and hierarchically, but requiring $O(n^3)$ time-complexity. The Recursive Transformer based on Differentiable Trees (R2D2) makes it possible to scale to large language model pretraining even with a complex tree encoder, by introducing a heuristic pruning method. However, its rule-based pruning process suffers from local optima and slow inference. In this paper, we propose a unified R2D2 method that overcomes these issues. We use a top-down unsupervised parser as a model-guided pruning method, which also enables parallel encoding during inference. Our parser casts parsing as a split point scoring task by first scoring all split points for a given sentence and then using the highest-scoring one to recursively split a span into two parts. The reverse order of the splits is considered as the order of pruning in the encoder. We optimize the unsupervised parser by minimizing the Kullback–Leibler distance between tree probabilities from the parser and the R2D2 model. Our experiments show that our Fast-R2D2 significantly improves the grammar induction quality and achieves competitive results in downstream tasks.

## Demo Session 1

11:00-12:30 (Link Admin)

[DEMO] **LM-Debugger: An Interactive Tool for Inspection and Intervention in Transformer-Based Language Models**
*Mor Geva, Avi Caciularu, Guy Dar, Paul Roit, Shoval Sadde, Micah Shlain, Bar Tamir and Yoav Goldberg* 11:00-12:30 (Link Admin)
The opaque nature and unexplained behavior of transformer-based language models (LMs) have spurred a wide interest in interpreting their predictions. However, current interpretation methods mostly focus on probing models from outside, executing behavioral tests, and analyzing salience input features, while the internal prediction construction process is largely not understood. In this work, we introduce LM-Debugger, an interactive debugger tool for transformer-based LMs, which provides a fine-grained interpretation of the model's internal prediction process, as well as a powerful framework for intervening in LM behavior. For its backbone, LM-Debugger relies on a recent method that interprets the inner token representations and their updates by the feed-forward layers in the vocabulary space. We demonstrate the utility of LM-Debugger for single-prediction debugging, by inspecting the internal disambiguation process done by GPT2. Moreover, we show how easily LM-Debugger allows to shift model behavior in a direction of the user's choice, by identifying a few vectors in the network and inducing effective interventions to the prediction process. We release LM-Debugger as an open-source tool and a demo over GPT2 models.

[DEMO] **FairLib: A Unified Framework for Assessing and Improving Fairness**
*Xudong Han, Aili Shen, Yitong Li, Lea Frermann, Timothy Baldwin and Trevor Cohn* 11:00-12:30 (Link Admin)
This paper presents FairLib, an open-source python library for assessing and improving model fairness. It provides a systematic framework for quickly accessing benchmark datasets, reproducing existing debiasing baseline models, developing new methods, evaluating models with different metrics, and visualizing their results. Its modularity and extensibility enable the framework to be used for diverse types of inputs, including natural language, images, and audio. We implement 14 debiasing methods, including pre-processing, at-training-time, and post-processing approaches. The built-in metrics cover the most commonly acknowledged fairness criteria and can be further generalized and customized for fairness evaluation.

[DEMO] **Snoopy: An Online Interface for Exploring the Effect of Pretraining Term Frequencies on Few-Shot LM Performance**
*Yasaman Razeghi, Raja sekhar Reddy Mekala, Robert L Logan IV, Matt Gardner and Sameer Singh* 11:00-12:30 (Link Admin)
Current evaluation schemes for large language models often fail to consider the impact of the overlap between pretraining corpus and test data on model performance statistics. Snoopy is an online interface that allows researchers to study this impact in few-shot learning settings. Our demo provides term frequency statistics for the Pile, which is an 800 GB corpus, accompanied by the precomputed performance of EleutherAI/GPT models on more than 20 NLP benchmarks, including numerical, commonsense reasoning, natural language understanding, and question-answering tasks. Snoopy allows a user to interactively align specific terms in test instances with their frequency in the Pile, enabling exploratory analysis of how term frequency is related to the accuracy of the models, which are hard to discover through automated means. A user can look at correlations over various model sizes and numbers of in-context examples and visualize the result across multiple (potentially aggregated) datasets. Using Snoopy, we show that a researcher can quickly replicate prior analyses for numerical tasks, while simultaneously allowing for much more expansive exploration that was previously challenging. Snoopy is available at https://nlp.ics.uci.edu/snoopy.

[DEMO] **Azimuth: Systematic Error Analysis for Text Classification**
*Gabrielle Gauthier-Melancon, Orlando Marquez Ayala, Lindsay Brin, Chris Tyler, Frederic Branchaud-Charron, Joseph Marinier, Karine Grande and Di Le* 11:00-12:30 (Link Admin)
We present Azimuth, an open-source and easy-to-use tool to perform error analysis for text classification. Compared to other stages of the ML development cycle, such as model training and hyper-parameter tuning, the process and tooling for the error analysis stage are less mature. However, this stage is critical for the development of reliable and trustworthy AI systems. To make error analysis more systematic, we propose an approach comprising dataset analysis and model quality assessment, which Azimuth facilitates. We aim to help AI practitioners discover and address areas where the model does not generalize by leveraging and integrating a range of ML techniques, such as saliency maps, similarity, uncertainty, and behavioral analyses, all in one tool. Our code and documentation are available at github.com/servicenow/azimuth.

# Session 3 - 14:00-15:30

## Language Modeling and Analysis of Language Models

14:00-15:30 (Hall B)

**The Geometry of Multilingual Language Model Representations**
*Tyler Chang, Zhuowen Tu and Benjamin Bergen* 14:00-14:15 (Hall B)
We assess how multilingual language models maintain a shared multilingual representation space while still encoding language-sensitive information in each language. Using XLM-R as a case study, we show that languages occupy similar linear subspaces after mean-centering, evaluated based on causal effects on language modeling performance and direct comparisons between subspaces for 88 languages. The subspace means differ along language-sensitive axes that are relatively stable throughout middle layers, and these axes encode information such as token vocabularies. Shifting representations by language means is sufficient to induce token predictions in different languages. However, we also identify stable language-neutral axes that encode information such as token positions and part-of-speech. We visualize representations projected onto language-sensitive and language-neutral axes, identifying language family and part-of-speech clusters, along with spirals, toruses, and curves representing token position information. These results demonstrate that multilingual language models encode information along orthogonal language-sensitive and language-neutral axes, allowing the models to extract a variety of features for downstream tasks and cross-lingual transfer learning.

**What Makes Instruction Learning Hard? An Investigation and a New Challenge in a Synthetic Environment**
*Matthew Finlayson, Kyle Richardson, Ashish Sabharwal and Peter Clark* 14:15-14:30 (Hall B)
The instruction learning paradigm—where a model learns to perform new tasks from task descriptions alone—has become popular in research on general-purpose models. The capabilities of large transformer models as instruction learners, however, remain poorly understood. We use a controlled synthetic environment to characterize such capabilities. Specifically, we use the task of deciding whether a given string matches a regular expression (viewed as an instruction) to identify properties of tasks, instructions, and instances that make instruction learning challenging. For instance, we find that our model, a fine-tuned T5-based text2text transformer, struggles with large regular languages, suggesting that less precise instructions are challenging for models. Instruction executions that require tracking longer contexts of prior steps are also difficult. We use our findings to systematically construct a challenging instruction learning dataset, which we call Hard RegSet. Fine-tuning on Hard RegSet, our large transformer learns to correctly interpret (with at least 90

**Language Model Pre-Training with Sparse Latent Typing**
*Liliang Ren, Zixuan Zhang, Han Wang, Clare Voss, ChengXiang Zhai and Heng Ji* 14:30-14:45 (Hall B)
Modern large-scale Pre-trained Language Models (PLMs) have achieved tremendous success on a wide range of downstream tasks. However, most of the LM pre-training objectives only focus on text reconstruction, but have not sought to learn latent-level interpretable representations of sentences. In this paper, we manage to push the language models to obtain a deeper understanding of sentences by proposing a new pre-training objective, Sparse Latent Typing, which enables the model to sparsely extract sentence-level keywords with diverse latent types. Experimental results show that our model is able to learn interpretable latent type categories in a self-supervised manner without using any external knowledge. Besides, the language model pre-trained with such an objective also significantly improves Information Extraction related downstream tasks in both supervised and few-shot settings. Our code is publicly available at https://github.com/renll/SparseLT.

**Ground-Truth Labels Matter: A Deeper Look into Input-Label Demonstrations**
*Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee and Taeuk Kim* 14:45-15:00 (Hall B)
Despite recent explosion of interests in in-context learning, the underlying mechanism and the precise impact of the quality of demonstrations remain elusive. Intuitively, ground-truth labels should have as much impact in in-context learning (ICL) as supervised learning, but recent work reported that the input-label correspondence is significantly less important than previously thought. Intrigued by this counter-intuitive observation, we re-examine the importance of ground-truth labels in in-context learning. With the introduction of two novel metrics, namely Label-Correctness Sensitivity and Ground-truth Label Effect Ratio (GLER), we were able to conduct quantifiable analysis on the impact of ground-truth label demonstrations. Through extensive analyses, we find that the correct input-label mappings can have varying impacts on the downstream in-context learning performances, depending on the experimental configuration. Through additional studies, we identify key components, such as the verbosity of prompt templates and the language model size, as the controlling factor to achieve more noise-resilient

ICL.

**Language Model Decomposition: Quantifying the Dependency and Correlation of Language Models**
*Hao Zhang* 15:00-15:15 (Hall B)
Pre-trained language models (LMs), such as BERT (Devlin et al., 2018) and its variants, have led to significant improvements on various NLP tasks in past years. However, a theoretical framework for studying their relationships is still missing. In this paper, we fill this gap by investigating the linear dependency between pre-trained LMs. The linear dependency of LMs is defined analogously to the linear dependency of vectors. We propose Language Model Decomposition (LMD) to represent a LM using a linear combination of other LMs as basis, and derive the closed-form solution. A goodness-of-fit metric for LMD similar to the coefficient of determination is defined and used to measure the linear dependency of a set of LMs. In experiments, we find that BERT and eleven (11) BERT-like LMs are 91% linearly dependent. This observation suggests that current state-of-the-art (SOTA) LMs are highly "correlated". To further advance SOTA we need more diverse and novel LMs that are less dependent on existing LMs.

**Iteratively Prompt Pre-trained Language Models for Chain of Thought**
*Boshi Wang, Xiang Deng and Huan Sun* 15:15-15:30 (Hall B)
While Pre-trained Language Models (PLMs) internalize a great amount of world knowledge, they have been shown incapable of recalling these knowledge to solve tasks requiring complex & multi-step reasoning. Similar to how humans develop a "chain of thought" for these tasks, how can we equip PLMs with such abilities? In this work, we explore an iterative prompting framework, a new prompting paradigm which progressively elicits relevant knowledge from PLMs for multi-step inference. We identify key limitations of existing prompting methods, namely they are either restricted to queries with a single identifiable relation/predicate, or being agnostic to input contexts, which makes it difficult to capture variabilities across different inference steps. We propose an iterative context-aware prompter, which addresses these limitations by learning to dynamically synthesize prompts conditioned on the current step's contexts. Experiments on three datasets involving multi-step reasoning show the effectiveness of the iterative scheme and the context-aware prompter design.

## Sentiment, Stylistic Analysis, Argument Mining & Discourse

14:00-15:30 (Hall A, Room A)

**Curriculum Knowledge Distillation for Emoji-supervised Cross-lingual Sentiment Analysis**
*Jianyang Zhang, Tao Liang, Mingyang Wan, Guowu Yang and Fengmao Lv* 14:00-14:15 (Hall A, Room A)
Existing sentiment analysis models have achieved great advances with the help of sufficient sentiment annotations. Unfortunately, many languages do not have sufficient sentiment corpus. To this end, recent studies have proposed cross-lingual sentiment analysis to transfer sentiment analysis models from resource-rich languages to low-resource languages. However, these studies either rely on external cross-lingual supervision (e.g., parallel corpora and translation model), or are limited by the cross-lingual gaps. In this work, based on the intuitive assumption that the relationships between emojis and sentiments are consistent across different languages, we investigate transferring sentiment knowledge across languages with the help of emojis. To this end, we propose a novel cross-lingual sentiment analysis approach dubbed Curriculum Knowledge Distiller (CKD). The core idea of CKD is to use emojis to bridge the source and target languages. Note that, compared with texts, emojis are more transferable, but cannot reveal the precise sentiment. Thus, we distill multiple Intermediate Sentiment Classifiers (ISC) on source language corpus with emojis to get ISCs with different attention weights of texts. To transfer them into the target language, we distill ISCs into the Target Language Sentiment Classifier (TSC) following the curriculum learning mechanism. In this way, TSC can learn delicate sentiment knowledge, meanwhile, avoid being affected by cross-lingual gaps. Experimental results on five cross-lingual benchmarks clearly verify the effectiveness of our approach.

**Sentence-Incremental Neural Coreference Resolution**
*Matt Grenander, Shay B. Cohen and Mark Steedman* 14:15-14:30 (Hall A, Room A)
We propose a sentence-incremental neural coreference resolution system which incrementally builds clusters after marking mention boundaries in a shift-reduce method. The system is aimed at bridging two recent approaches at coreference resolution: (1) state-of-the-art non-incremental models that incur quadratic complexity in document length with high computational cost, and (2) memory network-based models which operate incrementally but do not generalize beyond pronouns. For comparison, we simulate an incremental setting by constraining non-incremental systems to form partial coreference chains before observing new sentences. In this setting, our system outperforms comparable state-of-the-art methods by 2 F1 on OntoNotes and 6.8 F1 on the CODI-CRAC 2021 corpus. In a conventional coreference setup, our system achieves 76.3 F1 on OntoNotes and 45.5 F1 on CODI-CRAC 2021, which is comparable to state-of-the-art baselines. We also analyze variations of our system and show that the degree of incrementality in the encoder has a surprisingly large effect on the resulting performance.

**A Multifaceted Framework to Evaluate Evasion, Content Preservation, and Misattribution in Authorship Obfuscation Techniques**
*Malik Altakrori, Thomas Scialom, Benjamin C. M. Fung and Jackie Chi Kit Cheung* 14:30-14:45 (Hall A, Room A)
Authorship obfuscation techniques have commonly been evaluated based on their ability to hide the author's identity (evasion) while preserving the content of the original text. However, to avoid overstating the systems' effectiveness, evasion detection must be evaluated using competitive identification techniques in settings that mimic real-life scenarios, and the outcomes of the content-preservation evaluation have to be interpretable by potential users of these obfuscation tools. Motivated by recent work on cross-topic authorship identification and content preservation in summarization, we re-evaluate different authorship obfuscation techniques on detection evasion and content preservation. Furthermore, we propose a new information-theoretic measure to characterize the misattribution harm that can be caused by detection evasion. Our results reveal key weaknesses in state-of-the-art obfuscation techniques and a surprisingly competitive effectiveness from a back-translation baseline in all evaluation aspects.

**Affective Idiosyncratic Responses to Music**
*Sky CH-Wang, Evan Li, Oliver Li, Smaranda Muresan and Zhou Yu* 14:45-15:00 (Hall A, Room A)
Affective responses to music are highly personal. Despite consensus that idiosyncratic factors play a key role in regulating how listeners emotionally respond to music, precisely measuring the marginal effects of these variables has proved challenging. To address this gap, we develop computational methods to measure affective responses to music from over 403M listener comments on a Chinese social music platform. Building on studies from music psychology in systematic and quasi-causal analyses, we test for musical, lyrical, contextual, demographic, and mental health effects that drive listener affective responses. Finally, motivated by the social phenomenon known as wǎng-yì-yún, we identify influencing factors of platform user self-disclosures, the social support they receive, and notable differences in discloser user activity.

**Varifocal Question Generation for Fact-checking**
*Nedjma Ousidhoum, Zhangdie Yuan and Andreas Vlachos* 15:00-15:15 (Hall A, Room A)

Fact-checking requires retrieving evidence related to a claim under investigation. The task can be formulated as question generation based on a claim, followed by question answering. However, recent question generation approaches assume that the answer is known and typically contained in a passage given as input, whereas such passages are what is being sought when verifying a claim. In this paper, we present *Varifocal*, a method that generates questions based on different focal points within a given claim, i.e. different spans of the claim and its metadata, such as its source and date. Our method outperforms previous work on a fact-checking question generation dataset on a wide range of automatic evaluation metrics. These results are corroborated by our manual evaluation, which indicates that our method generates more relevant and informative questions. We further demonstrate the potential of focal points in generating sets of clarification questions for product descriptions.

**Topic-Regularized Authorship Representation Learning**
*Jitkapat Sawatphol, Nonthakit Chaiwong, Can Udomcharoenchaikit and Sarana Nutanong* 15:15-15:30 (Hall A, Room A)
Authorship attribution is a task that aims to identify the author of a given piece of writing. We aim to develop a generalized solution that can handle a large number of texts from authors and topics unavailable in training data. Previous studies have proposed strategies to address only either unseen authors or unseen topics. Authorship representation learning has been shown to work in open-set environments with a large number of unseen authors but has not been explicitly designed for cross-topic environments at the same time. To handle a large number of unseen authors and topics, we propose Authorship Representation Regularization (ARR), a distillation framework that creates authorship representation with reduced reliance on topic-specific information. To assess the performance of our framework, we also propose a cross-topic-open-set evaluation method. Our proposed method has improved performances in the cross-topic-open set setup over baselines in 4 out of 6 cases.

## Speech, Vision, Robotics, Multimodal Grounding 1 & CL

14:00-15:30 (Hall A, Room B)

**Normalized Contrastive Learning for Text-Video Retrieval**
*Yookoon Park, Mahmoud Azab, Seungwhan Moon, Bo Xiong, Florian Metze, Gourab Kundu and Kirmani Ahmed*14:00-14:15 (Hall A, Room B)
Cross-modal contrastive learning has led the recent advances in multimodal retrieval with its simplicity and effectiveness. In this work, however, we reveal that cross-modal contrastive learning suffers from incorrect normalization of the sum retrieval probabilities of each text or video instance. Specifically, we show that many test instances are either over- or under-represented during retrieval, significantly hurting the retrieval performance. To address this problem, we propose Normalized Contrastive Learning (NCL) which utilizes the Sinkhorn-Knopp algorithm to compute the instance-wise biases that properly normalize the sum retrieval probabilities of each instance so that every text and video instance is fairly represented during cross-modal retrieval. Empirical study shows that NCL brings consistent and significant gains in text-video retrieval on different model architectures, with new state-of-the-art multimodal retrieval metrics on the ActivityNet, MSVD, and MSR-VTT datasets without any architecture engineering.

**Abstract Visual Reasoning with Tangram Shapes**
*Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert Hawkins and Yoav Artzi* 14:15-14:30 (Hall A, Room B)
We introduce KiloGram, a resource for studying abstract visual reasoning in humans and machines. Drawing on the history of tangram puzzles as stimuli in cognitive science, we build a richly annotated dataset that, with >1k distinct stimuli, is orders of magnitude larger and more diverse than prior resources. It is both visually and linguistically richer, moving beyond whole shape descriptions to include segmentation maps and part labels. We use this resource to evaluate the abstract visual reasoning capacities of recent multi-modal models. We observe that pre-trained weights demonstrate limited abstract reasoning, which dramatically improves with fine-tuning. We also observe that explicitly describing parts aids abstract reasoning for both humans and models, especially when jointly encoding the linguistic and visual inputs.

**Z-LaVI: Zero-Shot Language Solver Fueled by Visual Imagination**
*Yue Yang, Wenlin Yao, Hongming Zhang, Xiaoyang Wang, Dong Yu and Jianshu Chen* 14:30-14:45 (Hall A, Room B)
Large-scale pretrained language models have made significant advances in solving downstream language understanding tasks. However, they generally suffer from reporting bias, the phenomenon describing the lack of explicit commonsense knowledge in written text, e.g., "an orange is orange". To overcome this limitation, we develop a novel approach, Z-LaVI, to endow language models with visual imagination capabilities. Specifically, we leverage two complementary types of "imaginations": (i) recalling existing images through retrieval and (ii) synthesizing nonexistent images via text-to-image generation. Jointly exploiting the language inputs and the imagination, a pretrained vision-language model (e.g., CLIP) eventually composes a zero-shot solution to the original language tasks. Notably, fueling language models with imagination can effectively leverage visual knowledge to solve plain language tasks. In consequence, Z-LaVI consistently improves the zero-shot performance of existing language models across a diverse set of language tasks.

**DANLI: Deliberative Agent for Following Natural Language Instructions**
*Yichi Zhang, Jianing Yang, Jiayi Pan, Shane Storks, Nikhil Devraj, Ziqiao Ma, Keunwoo Yu, Yuwei Bao and Joyce Chai* 14:45-15:00 (Hall A, Room B)
Recent years have seen an increasing amount of work on embodied AI agents that can perform tasks by following human language instructions. However, most of these agents are reactive, meaning that they simply learn and imitate behaviors encountered in the training data. These reactive agents are insufficient for long-horizon complex tasks. To address this limitation, we propose a neuro-symbolic deliberative agent that, while following language instructions, proactively applies reasoning and planning based on its neural and symbolic representations acquired from past experience (e.g., natural language and egocentric vision). We show that our deliberative agent achieves greater than 70% improvement over reactive baselines on the challenging TEACh benchmark. Moreover, the underlying reasoning and planning processes, together with our modular framework, offer impressive transparency and explainability to the behaviors of the agent. This enables an in-depth understanding of the agent's capabilities, which shed light on challenges and opportunities for future embodied agents for instruction following. The code is available at https://github.com/sled-group/DANLI.

**Learning a Grammar Inducer from Massive Uncurated Instructional Videos**
*Songyang Zhang, Linfeng Song, Lifeng Jin, Haitao Mi, Kun Xu, Dong Yu and Jiebo Luo* 15:00-15:15 (Hall A, Room B)
Video-aided grammar induction aims to leverage video information for finding more accurate syntactic grammars for accompanying text. While previous work focuses on building systems for inducing grammars on text that are well-aligned with video content, we investigate the scenario, in which text and video are only in loose correspondence. Such data can be found in abundance online, and the weak correspondence is similar to the indeterminacy problem studied in language acquisition. Furthermore, we build a new model that can better learn video-span correlation without manually designed features adopted by previous work. Experiments show that our model trained only on large-scale YouTube data with no text-video alignment reports strong and robust performances across three unseen datasets, despite domain shift and

noisy label issues. Furthermore our model yields higher F1 scores than the previous state-of-the-art systems trained on in-domain data.

**How well can Text-to-Image Generative Models understand Ethical Natural Language Interventions?**
*Hritik Bansal, Da Yin, Masoud Monajatipoor and Kai-Wei Chang* 15:15-15:30 (Hall A, Room B)
Text-to-image generative models have achieved unprecedented success in generating high-quality images based on natural language descriptions. However, it is shown that these models tend to favor specific social groups when prompted with neutral text descriptions (e.g., 'a photo of a lawyer'). Following Zhao et al. (2021), we study the effect on the diversity of the generated images when adding *ethical intervention* that supports equitable judgment (e.g., 'if all individuals can be a lawyer irrespective of their gender') in the input prompts. To this end, we introduce an Ethical NaTural Language Interventions in Text-to-Image GENeration (ENTIGEN) benchmark dataset to evaluate the change in image generations conditional on ethical interventions across three social axes – gender, skin color, and culture. Through CLIP-based and human evaluation on minDALL.E, DALL.E-mini and Stable Diffusion, we find that the model generations cover diverse social groups while preserving the image quality. In some cases, the generations would be anti-stereotypical (e.g., models tend to create images with individuals that are perceived as man when fed with prompts about makeup) in the presence of ethical intervention. Preliminary studies indicate that a large change in the model predictions is triggered by certain phrases such as 'irrespective of gender' in the context of gender bias in the ethical interventions. We release code and annotated data at https://github.com/Hritikbansal/entigen_emnlp.

## Question Answering 1

14:00-15:30 (Hall A, Room C)

**Generating Natural Language Proofs with Verifier-Guided Search**
*Kaiyu Yang, Jia Deng and Danqi Chen* 14:00-14:15 (Hall A, Room C)
Reasoning over natural language is a challenging problem in NLP. In this work, we focus on proof generation: Given a hypothesis and a set of supporting facts, the model generates a proof tree indicating how to derive the hypothesis from supporting facts. Compared to generating the entire proof in one shot, stepwise generation can better exploit the compositionality and generalize to longer proofs but has achieved limited success on real-world data. Existing stepwise methods struggle to generate proof steps that are both logically valid and relevant to the hypothesis. Instead, they tend to hallucinate invalid steps given the hypothesis. In this paper, we present a novel stepwise method, NLProofS (Natural Language Proof Search), which learns to generate relevant steps conditioning on the hypothesis. At the core of our approach, we train an independent verifier to check the validity of the proof steps to prevent hallucination. Instead of generating steps greedily, we search for proofs maximizing a global proof score judged by the verifier. NLProofS achieves state-of-the-art performance on EntailmentBank and RuleTaker. Specifically, it improves the correctness of predicted proofs from 27.7% to 33.3% in the distractor setting of EntailmentBank, demonstrating the effectiveness of NLProofS in generating challenging human-authored proofs.

**Improving Complex Knowledge Base Question Answering via Question-to-Action and Question-to-Question Alignment**
*Yechun Tang, Xiaoxia Cheng and Weiming Lu* 14:15-14:30 (Hall A, Room C)
Complex knowledge base question answering can be achieved by converting questions into sequences of predefined actions. However, there is a significant semantic and structural gap between natural language and action sequences, which makes this conversion difficult. In this paper, we introduce an alignment-enhanced complex question answering framework, called ALCQA, which mitigates this gap through question-to-action alignment and question-to-question alignment. We train a question rewriting model to align the question and each action, and utilize a pretrained language model to implicitly align the question and KG artifacts. Moreover, considering that similar questions correspond to similar action sequences, we retrieve top-k similar question-answer pairs at the inference stage through question-to-question alignment and propose a novel reward-guided action sequence selection strategy to select from candidate action sequences. We conduct experiments on CQA and WQSP datasets, and the results show that our approach outperforms state-of-the-art methods and obtains a 9.88
https://github.com/TTTTTTTTy/ALCQA.

**Successive Prompting for Decomposing Complex Questions**
*Dheeru Dua, Shivanshu Gupta, Sameer Singh and Matt Gardner* 14:30-14:45 (Hall A, Room C)
Answering complex questions that require making latent decisions is a challenging task, especially when limited supervision is available. Recent works leverage the capabilities of large language models (LMs) to perform complex question answering in a few-shot setting by demonstrating how to output intermediate rationalizations while solving the complex question in a single pass. We introduce "Successive Prompting" where, we iteratively break down a complex task into a simple task, solve it, and then repeat the process until we get the final solution. Successive prompting decouples the supervision for decomposing complex questions from the supervision for answering simple questions, allowing us to (1) have multiple opportunities to query in-context examples at each reasoning step (2) learn question decomposition separately from question answering, including using synthetic data, and (3) use bespoke (fine-tuned) components for reasoning steps where a large LM does not perform well. The intermediate supervision is typically manually written, which can be expensive to collect. We introduce a way to generate synthetic dataset which can be used to bootstrap model's ability to decompose and answer intermediate questions. Our best model (with successive prompting) achieves an improvement in F1 of 5% when compared with a state-of-the-art model with synthetic augmentations and few-shot version of the DROP dataset.

**M3: A Multi-View Fusion and Multi-Decoding Network for Multi-Document Reading Comprehension**
*Liang Wen, Houfeng Wang, Yingwei Luo and Xiaolin Wang* 14:45-15:00 (Hall A, Room C)
Multi-document reading comprehension task requires collecting evidences from different documents for answering questions. Previous research works either use the extractive modeling method to naively integrate the scores from different documents on the encoder side or use the generative modeling method to collect the clues from different documents on the decoder side individually. However, any single modeling method cannot make full of the advantages of both. In this work, we propose a novel method that tries to employ a multi-view fusion and multi-decoding mechanism to achieve it. For one thing, our approach leverages question-centered fusion mechanism and cross-attention mechanism to gather fine-grained fusion of evidence clues from different documents in the encoder and decoder concurrently. For another, our method simultaneously employs both the extractive decoding approach and the generative decoding method to effectively guide the training process. Compared with existing methods, our method can perform both extractive decoding and generative decoding independently and optionally. Our experiments on two mainstream multi-document reading comprehension datasets (Natural Questions and TriviaQA) demonstrate that our method can provide consistent improvements over previous state-of-the-art methods.

**Semantic Framework based Query Generation for Temporal Question Answering over Knowledge Graphs**
*Wentao Ding, Hao Chen, Huayu Li and Yuzhong Qu* 15:00-15:15 (Hall A, Room C)
Answering factual questions with temporal intent over knowledge graphs (temporal KGQA) attracts rising attention in recent years. In the generation of temporal queries, existing KGQA methods ignore the fact that some intrinsic connections between events can make them tem-

porally related, which may limit their capability. We systematically analyze the possible interpretation of temporal constraints and conclude the interpretation structures as the Semantic Framework of Temporal Constraints, SF-TCons. Based on the semantic framework, we propose a temporal question answering method, SF-TQA, which generates query graphs by exploring the relevant facts of mentioned entities, where the exploring process is restricted by SF-TCons. Our evaluations show that SF-TQA significantly outperforms existing methods on two benchmarks over different knowledge graphs.

**Improving compositional generalization for multi-step quantitative reasoning in question answering**
*Armineh Nourbakhsh, Cathy Jiao, Sameena Shah and Carolyn Rosé* 15:15-15:30 (Hall A, Room C)
Quantitative reasoning is an important aspect of question answering, especially when numeric and verbal cues interact to indicate sophisticated, multi-step programs. In this paper, we demonstrate how modeling the compositional nature of quantitative text can enhance the performance and robustness of QA models, allowing them to capture arithmetic logic that is expressed verbally. Borrowing from the literature on semantic parsing, we propose a method that encourages the QA models to adjust their attention patterns and capture input/output alignments that are meaningful to the reasoning task. We show how this strategy improves program accuracy and renders the models more robust against overfitting as the number of reasoning steps grows. Our approach is designed as a standalone module which can be prepended to many existing models and trained in an end-to-end fashion without the need for additional supervisory signal. As part of this exercise, we also create a unified dataset building on four previously released numerical QA datasets over tabular data.

## CL & TACL 1

14:00-15:30 (Hall A, Room D)

[TACL] **OPAL: Ontology-Aware Pretrained Language Model for End-to-End Task-Oriented Dialogue**
*Zhi Chen, Yuncong Liu, Lu Chen, Su Zhu, Mengyue Wu and Kai Yu* 14:00-14:15 (Hall A, Room D)
This paper presents an ontology-aware pretrained language model (OPAL) for end-to-end task-oriented dialogue (TOD). Unlike chit-chat dialogue models, task-oriented dialogue models fulfill at least two task-specific modules: dialogue state tracker (DST) and response generator (RG). The dialogue state consists of the domain-slot-value triples, which are regarded as the user's constraints to search the domain-related databases. The large-scale task-oriented dialogue data with the annotated structured dialogue state usually are inaccessible. It prevents the development of the pretrained language model for the task-oriented dialogue. We propose a simple yet effective pretraining method to alleviate this problem, which consists of two pretraining phases. The first phase is to pretrain on large-scale contextual text data, where the structured information of the text is extracted by the information extracting tool. To bridge the gap between the pretraining method and downstream tasks, we design two pretraining tasks: ontology recovery and next-text generation, which simulates the DST and RG, respectively. The second phase is to fine-tune the pretrained model on the TOD data. The experimental results show that our proposed method achieves an exciting boost and get competitive performance even without any TOD data on CamRest676 and MultiWOZ benchmarks.

[TACL] **True Few-Shot Learning With Prompts - A Real-World Perspective**
*Timo Schick and Hinrich Schütze* 14:15-14:30 (Hall A, Room D)
Prompt-based approaches excel at few-shot learning. However, Perez et al. (2021) recently cast doubt on their performance as they had difficulty getting good results in a "true" few-shot setting in which prompts and hyperparameters cannot be tuned on a dev set. In view of this, we conduct an extensive study of PET, a method that combines textual instructions with example-based finetuning. We show that, if correctly configured, PET performs strongly in true few-shot settings without a dev set. Crucial for this strong performance is a number of design choices, including PET's ability to intelligently handle multiple prompts. We put our findings to a real-world test by running PET on RAFT, a benchmark of tasks taken from realistic NLP applications for which no labeled dev or test sets are available. PET achieves a new state of the art on RAFT and performs close to non-expert humans for 7 out of 11 tasks. These results demonstrate that prompt-based learners can successfully be applied in true few-shot settings and underpin our belief that learning from instructions will play an important role on the path towards human-like few-shot learning capabilities.

[TACL] **Generate, Annotate, and Learn: NLP with Synthetic Text**
*Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari and Mohammad Norouzi* 14:30-14:45 (Hall A, Room D)
This paper studies the use of language models as a source of synthetic unlabeled text for NLP. We formulate a general framework called "generate, annotate, and learn (GAL)" to take advantage of synthetic text within knowledge distillation, self-training, and few-shot learning applications. To generate high-quality task-specific text, we either fine-tune LMs on inputs from the task of interest, or prompt large LMs with few examples. We use the best available classifier to annotate synthetic text with soft pseudo labels for knowledge distillation and self-training, and use LMs to obtain hard labels for few-shot learning. We train new supervised models on the combination of labeled and pseudo-labeled data, which results in significant gains across several applications. We investigate key components of GAL and present theoretical and empirical arguments against the use of class-conditional LMs to generate synthetic labeled text instead of unlabeled text. GAL achieves new state-of-the-art knowledge distillation results for 6-layer transformers on the GLUE leaderboard.

[TACL] **Structural Persistence in Language Models: Priming as a Window into Abstract Language Representations**
*Arabella Sinclair, Jaap Jumelet, Willem Zuidema and Raquel Fernández* 14:45-15:00 (Hall A, Room D)
We investigate the extent to which modern, neural language models are susceptible to structural priming, the phenomenon whereby the structure of a sentence makes the same structure more probable in a follow-up sentence. We explore how priming can be used to study the potential of these models to learn abstract structural information, which is a prerequisite for good performance on tasks that require natural language understanding skills. We introduce a novel metric and release Prime-LM, a large corpus where we control for various linguistic factors which interact with priming strength. We find that Transformer models indeed show evidence of structural priming, but also that the generalisations they learned are to some extent modulated by semantic information. Our experiments also show that the representations acquired by the models may not only encode abstract sequential structure but involve certain level of hierarchical syntactic information. More generally, our study shows that the priming paradigm is a useful, additional tool for gaining insights into the capacities of language models and opens the door to future priming-based investigations that probe the model's internal states.

[TACL] **ProoFVer: Natural Logic Theorem Proving for Fact Verification**
*Amrith Krishna, Sebastian Riedel and Andreas Vlachos* 15:00-15:15 (Hall A, Room D)
Fact verification systems typically rely on neural network classifiers for veracity prediction which lack explainability. This paper proposes ProoFVer, which uses a seq2seq model to generate natural logic-based inferences as proofs. These proofs consist of lexical mutations between spans in the claim and the evidence retrieved, each marked with a natural logic operator. Claim veracity is determined solely based on the sequence of these operators. Hence, these proofs are faithful explanations, and this makes ProoFVer faithful by construction. Currently, ProoFVer has the highest label accuracy and the second-best Score on the FEVER leaderboard. Furthermore, it improves by 13.21% points

over the next best model on a dataset with counterfactual instances, demonstrating its robustness. As explanations, the proofs show better overlap with human rationales than attention-based highlights and the proofs help humans predict model decisions correctly more often than using the evidence directly.

[TACL] **Investigating Reasons for Disagreement in Natural Language Inference**
*Nan-Jiang Jiang and Marie-Catherine de Marneffe* 15:15-15:30 (Hall A, Room D)
We investigate how disagreement in natural language inference (NLI) annotation arises. We developed a taxonomy of disagreement sources with 10 categories spanning 3 high-level classes. We found that some disagreements are due to uncertainty in the sentence meaning, others to annotator biases and task artifacts, leading to different interpretations of the label distribution. We explore two modeling approaches for detecting items with potential disagreement: a 4-way classification with a "Complicated" label in addition to the three standard NLI labels, and a multilabel classification approach. We found that the multilabel classification is more expressive and gives better recall of the possible interpretations in the data.

## Ethics & Computational Social Science and Cultural Analytics

14:00-15:30 (Collaboratorium)

**Gendered Mental Health Stigma in Masked Language Models**
*Inna Lin, Lucille Njoo, Anjalie Field, Ashish Sharma, Katharina Reinecke, Tim Althoff and Yulia Tsvetkov* 14:00-14:15 (Collaboratorium)
Mental health stigma prevents many individuals from receiving the appropriate care, and social psychology studies have shown that mental health tends to be overlooked in men. In this work, we investigate gendered mental health stigma in masked language models. In doing so, we operationalize mental health stigma by developing a framework grounded in psychology research: we use clinical psychology literature to curate prompts, then evaluate the models' propensity to generate gendered words. We find that masked language models capture societal stigma about gender in mental health: models are consistently more likely to predict female subjects than male in sentences about having a mental health condition (32% vs. 19%), and this disparity is exacerbated for sentences that indicate treatment-seeking behavior. Furthermore, we find that different models capture dimensions of stigma differently for men and women, associating stereotypes like anger, blame, and pity more with women with mental health conditions than with men. In showing the complex nuances of models' gendered mental health stigma, we demonstrate that context and overlapping dimensions of identity are important considerations when assessing computational models' social biases.

**SafeText: A Benchmark for Exploring Physical Safety in Language Models**
*Sharon Levy, Emily Allaway, Melanie Subbiah, Lydia Chilton, Desmond Patton, Kathleen McKeown and William Yang Wang* 14:15-14:30 (Collaboratorium)
Understanding what constitutes safe text is an important issue in natural language processing and can often prevent the deployment of models deemed harmful and unsafe. One such type of safety that has been scarcely studied is commonsense physical safety, i.e. text that is not explicitly violent and requires additional commonsense knowledge to comprehend that it leads to physical harm. We create the first benchmark dataset, SafeText, comprising real-life scenarios with paired safe and physically unsafe pieces of advice. We utilize SafeText to empirically study commonsense physical safety across various models designed for text generation and commonsense reasoning tasks. We find that state-of-the-art large language models are susceptible to the generation of unsafe text and have difficulty rejecting unsafe advice. As a result, we argue for further studies of safety and the assessment of commonsense physical safety in models before release.

**Prompting for Multimodal Hateful Meme Classification**
*Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong and Jing Jiang* 14:30-14:45 (Collaboratorium)
Hateful meme classification is a challenging multimodal task that requires complex reasoning and contextual background knowledge. Ideally, we could leverage an explicit external knowledge base to supplement contextual and cultural information in hateful memes. However, there is no known explicit external knowledge base that could provide such hate speech contextual information. To address this gap, we propose PromptHate, a simple yet effective prompt-based model that prompts pre-trained language models (PLMs) for hateful meme classification. Specifically, we construct simple prompts and provide a few in-context examples to exploit the implicit knowledge in the pre-trained RoBERTa language model for hateful meme classification. We conduct extensive experiments on two publicly available hateful and offensive meme datasets. Our experiment results show that PromptHate is able to achieve a high AUC of 90.96, outperforming state-of-the-art baselines on the hateful meme classification task. We also perform fine-grain analyses and case studies on various prompt settings and demonstrate the effectiveness of the prompts on hateful meme classification.

**Modeling Information Change in Science Communication with Semantically Matched Paraphrases**
*Dustin Wright, Jiaxin Pei, David Jurgens and Isabelle Augenstein* 14:45-15:00 (Collaboratorium)
Whether the media faithfully communicate scientific information has long been a core issue to the science community. Automatically identifying paraphrased scientific findings could enable large-scale tracking and analysis of information changes in the science communication process, but this requires systems to understand the similarity between scientific information across multiple domains. To this end, we present the SCIENTIFIC PARAPHRASE AND INFORMATION CHANGE DATASET (SPICED), the first paraphrase dataset of scientific findings annotated for degree of information change. SPICED contains 6,000 scientific finding pairs extracted from news stories, social media discussions, and full texts of original papers. We demonstrate that SPICED poses a challenging task and that models trained on SPICED improve downstream performance on evidence retrieval for fact checking of real-world scientific claims. Finally, we show that models trained on SPICED can reveal large-scale trends in the degrees to which people and organizations faithfully communicate new scientific findings. Data, code, and pre-trained models are available at http://www.copenlu.com/publication/2022_emnlp_wright/.

**Tracing Semantic Variation in Slang**
*Zhewei Sun and Yang Xu* 15:00-15:15 (Collaboratorium)
The meaning of a slang term can vary in different communities. However, slang semantic variation is not well understood and under-explored in the natural language processing of slang. One existing view argues that slang semantic variation is driven by culture-dependent communicative needs. An alternative view focuses on slang's social functions suggesting that the desire to foster semantic distinction may have led to the historical emergence of community-specific slang senses. We explore these theories using computational models and test them against historical slang dictionary entries, with a focus on characterizing regularity in the geographical variation of slang usages attested in the US and the UK over the past two centuries. We show that our models are able to predict the regional identity of emerging slang word meanings from historical slang records. We offer empirical evidence that both communicative need and semantic distinction play a role in the variation of slang meaning yet their relative importance fluctuates over the course of history. Our work offers an opportunity for incorporating historical cultural elements into the natural language processing of slang.

**An Empirical Analysis of Memorization in Fine-tuned Autoregressive Language Models**
*Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David Evans and Taylor Berg-Kirkpatrick* 15:15-15:30 (Collaboratorium)
Large language models are shown to present privacy risks through memorization of training data, and several recent works have studied such risks for the pre-training phase. Little attention, however, has been given to the fine-tuning phase and it is not well understood how different fine-tuning methods (such as fine-tuning the full model, the model head, and adapter) compare in terms of memorization risk. This presents increasing concern as the "pre-train and fine-tune" paradigm proliferates. In this paper, we empirically study memorization of fine-tuning methods using membership inference and extraction attacks, and show that their susceptibility to attacks is very different. We observe that fine-tuning the head of the model has the highest susceptibility to attacks, whereas fine-tuning smaller adapters appears to be less vulnerable to known extraction attacks.

## Poster Sessions 3 & 4

14:00-15:30 (Atrium)

**A Systematic Investigation of Commonsense Knowledge in Large Language Models**
*Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d'Autume, Phil Blunsom and Aida Nematzadeh* 14:00-15:30 (Atrium)
Language models (LMs) trained on large amounts of data have shown impressive performance on many NLP tasks under the zero-shot and few-shot setup. Here we aim to better understand the extent to which such models learn commonsense knowledge — a critical component of many NLP applications. We conduct a systematic and rigorous zero-shot and few-shot commonsense evaluation of large pre-trained LMs, where we: (i) carefully control for the LMs' ability to exploit potential surface cues and annotation artefacts, and (ii) account for variations in performance that arise from factors that are not related to commonsense knowledge. Our findings highlight the limitations of pre-trained LMs in acquiring commonsense knowledge without task-specific supervision; furthermore, using larger models or few-shot evaluation is insufficient to achieve human-level commonsense performance.

**How to disagree well: Investigating the dispute tactics used on Wikipedia**
*Christine De Kock and Andreas Vlachos* 14:00-15:30 (Atrium)
Disagreements are frequently studied from the perspective of either detecting toxicity or analysing argument structure. We propose a framework of dispute tactics which unifies these two perspectives, as well as other dialogue acts which play a role in resolving disputes, such as asking questions and providing clarification. This framework includes a preferential ordering among rebuttal-type tactics, ranging from ad hominem attacks to refuting the central argument. Using this framework, we annotate 213 disagreements (3,865 utterances) from Wikipedia Talk pages. This allows us to investigate research questions around the tactics used in disagreements; for instance, we provide empirical validation of the approach to disagreement recommended by Wikipedia. We develop models for multilabel prediction of dispute tactics in an utterance, achieving the best performance with a transformer-based label powerset model. Adding an auxiliary task to incorporate the ordering of rebuttal tactics further yields a statistically significant increase. Finally, we show that these annotations can be used to provide useful additional signals to improve performance on the task of predicting escalation.

**Dictionary-Assisted Supervised Contrastive Learning**
*Patrick Wu, Richard Bonneau, Joshua Tucker and Jonathan Nagler* 14:00-15:30 (Atrium)
Text analysis in the social sciences often involves using specialized dictionaries to reason with abstract concepts, such as perceptions about the economy or abuse on social media. These dictionaries allow researchers to impart domain knowledge and note subtle usages of words relating to a concept(s) of interest. We introduce the dictionary-assisted supervised contrastive learning (DASCL) objective, allowing researchers to leverage specialized dictionaries when fine-tuning pretrained language models. The text is first keyword simplified: a common, fixed token replaces any word in the corpus that appears in the dictionary(ies) relevant to the concept of interest. During fine-tuning, a supervised contrastive objective draws closer the embeddings of the original and keyword-simplified texts of the same class while pushing further apart the embeddings of different classes. The keyword-simplified texts of the same class are more textually similar than their original text counterparts, which additionally draws the embeddings of the same class closer together. Combining DASCL and cross-entropy improves classification performance metrics in few-shot learning settings and social science applications compared to using cross-entropy alone and alternative contrastive and data augmentation methods.

**FineD-Eval: Fine-grained Automatic Dialogue-Level Evaluation**
*Chen Zhang, Luis Fernando D'Haro, Qiquan Zhang, Thomas Friedrichs and Haizhou Li* 14:00-15:30 (Atrium)
Recent model-based reference-free metrics for open-domain dialogue evaluation exhibit promising correlations with human judgment. However, they either perform turn-level evaluation or look at a single dialogue quality dimension. One would expect a good evaluation metric to assess multiple quality dimensions at the dialogue level. To this end, we are motivated to propose a multi-dimensional dialogue-level metric, which consists of three sub-metrics with each targeting a specific dimension. The sub-metrics are trained with novel self-supervised objectives and exhibit strong correlations with human judgment for their respective dimensions. Moreover, we explore two approaches to combine the sub-metrics: metric ensemble and multitask learning. Both approaches yield a holistic metric that significantly outperforms individual sub-metrics. Compared to the existing state-of-the-art metric, the combined metrics achieve around 16% relative improvement on average across three high-quality dialogue-level evaluation benchmarks.

**FlowEval: A Consensus-Based Dialogue Evaluation Framework Using Segment Act Flows**
*Jianqiao Zhao, Yanyang Li, Wanyu Du, Yangfeng Ji, Dong Yu, Michael Lyu and Liwei Wang* 14:00-15:30 (Atrium)
Despite recent progress in open-domain dialogue evaluation, how to develop automatic metrics remains an open problem. We explore the potential of dialogue evaluation featuring dialog act information, which was hardly explicitly modeled in previous methods. However, defined at the utterance level in general, dialog act is of coarse granularity, as an utterance can contain multiple segments possessing different functions. Hence, we propose segment act, an extension of dialog act from utterance level to segment level, and crowdsource a large-scale dataset for it. To utilize segment act flows, sequences of segment acts, for evaluation, we develop the first consensus-based dialogue evaluation framework, FlowEval. This framework provides a reference-free approach for dialog evaluation by finding pseudo-references. Extensive experiments against strong baselines on three benchmark datasets demonstrate the effectiveness and other desirable characteristics of our FlowEval, pointing out a potential path for better dialogue evaluation.

**Injecting Domain Knowledge in Language Models for Task-oriented Dialogue Systems**
*Denis Emelin, Daniele Bonadiman, Sawsan Alqahtani, Yi Zhang and Saab Mansour* 14:00-15:30 (Atrium)
Pre-trained language models (PLM) have advanced the state-of-the-art across NLP applications, but lack domain-specific knowledge that does

not naturally occur in pre-training data. Previous studies augmented PLMs with symbolic knowledge for different downstream NLP tasks. However, knowledge bases (KBs) utilized in these studies are usually large-scale and static, in contrast to small, domain-specific, and modifiable knowledge bases that are prominent in real-world task-oriented dialogue (TOD) systems. In this paper, we showcase the advantages of injecting domain-specific knowledge prior to fine-tuning on TOD tasks. To this end, we utilize light-weight adapters that can be easily integrated with PLMs and serve as a repository for facts learned from different KBs. To measure the efficacy of proposed knowledge injection methods, we introduce Knowledge Probing using Response Selection (KPRS) – a probe designed specifically for TOD models. Experiments on KPRS and the response generation task show improvements of knowledge injection with adapters over strong baselines.

**An Unsupervised, Geometric and Syntax-aware Quantification of Polysemy**
*Anmol Goel, Charu Sharma and Ponnurangam Kumaraguru* 14:00-15:30 (Atrium)
Polysemy is the phenomenon where a single word form possesses two or more related senses. It is an extremely ubiquitous part of natural language and analyzing it has sparked rich discussions in the linguistics, psychology and philosophy communities alike. With scarce attention paid to polysemy in computational linguistics, and even scarcer attention toward quantifying polysemy, in this paper, we propose a novel, unsupervised framework to compute and estimate polysemy scores for words in multiple languages. We infuse our proposed quantification with syntactic knowledge in the form of dependency structures. This informs the final polysemy scores of the lexicon motivated by recent linguistic findings that suggest there is an implicit relation between syntax and ambiguity/polysemy. We adopt a graph based approach by computing the discrete Ollivier Ricci curvature on a graph of the contextual nearest neighbors. We test our framework on curated datasets controlling for different sense distributions of words in 3 typologically diverse languages - English, French and Spanish. The effectiveness of our framework is demonstrated by significant correlations of our quantification with expert human annotated language resources like WordNet. We observe a 0.3 point increase in the correlation coefficient as compared to previous quantification studies in English. Our research leverages contextual language models and syntactic structures to empirically support the widely held theoretical linguistic notion that syntax is intricately linked to ambiguity/polysemy.

**Model Cascading: Towards Jointly Improving Efficiency and Accuracy of NLP Systems**
*Neeraj Varshney and Chitta Baral* 14:00-15:30 (Atrium)
Do all instances need inference through the big models for a correct prediction? Perhaps not; some instances are easy and can be answered correctly by even small capacity models. This provides opportunities for improving the computational efficiency of systems. In this work, we present an explorative study on 'model cascading', a simple technique that utilizes a collection of models of varying capacities to accurately yet efficiently output predictions. Through comprehensive experiments in multiple task settings that differ in the number of models available for cascading (K value), we show that cascading improves both the computational efficiency and the prediction accuracy. For instance, in K=3 setting, cascading saves up to 88.93% computation cost and consistently achieves superior prediction accuracy with an improvement of up to 2.18%. We also study the impact of introducing additional models in the cascade and show that it further increases the efficiency improvements. Finally, we hope that our work will facilitate development of efficient NLP systems making their widespread adoption in real-world applications possible.

**Better Few-Shot Relation Extraction with Label Prompt Dropout**
*Peiyuan Zhang and Wei Lu* 14:00-15:30 (Atrium)
Few-shot relation extraction aims to learn to identify the relation between two entities based on very limited training examples. Recent efforts found that textual labels (i.e., relation names and relation descriptions) could be extremely useful for learning class representations, which will benefit the few-shot learning task. However, what is the best way to leverage such label information in the learning process is an important research question. Existing works largely assume such textual labels are always present during both learning and prediction. In this work, we argue that such approaches may not always lead to optimal results. Instead, we present a novel approach called label prompt dropout, which randomly removes label descriptions in the learning process. Our experiments show that our approach is able to lead to improved class representations, yielding significantly better results on the few-shot relation extraction task.

**EDIN: An End-to-end Benchmark and Pipeline for Unknown Entity Discovery and Indexing**
*Nora Kassner, Fabio Petroni, Mikhail Plekhanov, Sebastian Riedel and Nicola Cancedda* 14:00-15:30 (Atrium)
Existing work on Entity Linking mostly assumes that the reference knowledge base is complete, and therefore all mentions can be linked. In practice this is hardly ever the case, as knowledge bases are incomplete and because novel concepts arise constantly. We introduce the temporally segmented Unknown Entity Discovery and Indexing (EDIN)-benchmark where unknown entities, that is entities not part of the knowledge base and without descriptions and labeled mentions, have to be integrated into an existing entity linking system. By contrasting EDIN with zero-shot entity linking, we provide insight on the additional challenges it poses. Building on dense-retrieval based entity linking, we introduce the end-to-end EDIN-pipeline that detects, clusters, and indexes mentions of unknown entities in context. Experiments show that indexing a single embedding per entity unifying the information of multiple mentions works better than indexing mentions independently.

**Towards Reinterpreting Neural Topic Models via Composite Activations**
*Jia Peng Lim and Hady Lauw* 14:00-15:30 (Atrium)
Most Neural Topic Models (NTM) use a variational auto-encoder framework producing K topics limited to the size of the encoder's output. These topics are interpreted through the selection of the top activated words via the weights or reconstructed vector of the decoder that are directly connected to each neuron. In this paper, we present a model-free two-stage process to reinterpret NTM and derive further insights on the state of the trained model. Firstly, building on the original information from a trained NTM, we generate a pool of potential candidate "composite topics" by exploiting possible co-occurrences within the original set of topics, which decouples the strict interpretation of topics from the original NTM. This is followed by a combinatorial formulation to select a final set of composite topics, which we evaluate for coherence and diversity on a large external corpus. Lastly, we employ a user study to derive further insights on the reinterpretation process.

**Saving Dense Retriever from Shortcut Dependency in Conversational Search**
*Sungdong Kim and Gangwoo Kim* 14:00-15:30 (Atrium)
Conversational search (CS) needs a holistic understanding of conversational inputs to retrieve relevant passages. In this paper, we demonstrate the existence of a *retrieval shortcut* in CS, which causes models to retrieve passages solely relying on partial history while disregarding the latest question. With in-depth analysis, we first show that naively trained dense retrievers heavily exploit the shortcut and hence perform poorly when asked to answer history-independent questions. To build more robust models against shortcut dependency, we explore various hard negative mining strategies. Experimental results show that training with the model-based hard negatives effectively mitigates the dependency on the shortcut, significantly improving dense retrievers on recent CS benchmarks. In particular, our retriever outperforms the previous state-of-the-art model by 11.0 in Recall@10 on QReCC.

**Communication breakdown: On the low mutual intelligibility between human and neural captioning**
*Roberto Dessì, Eleonora Gualdoni, Francesca Franzon, Gemma Boleda and Marco Baroni* 14:00-15:30 (Atrium)
We compare the 0-shot performance of a neural caption-based image retriever when given as input either human-produced captions or captions generated by a neural captioner. We conduct this comparison on the recently introduced ImageCoDe data-set (Krojer et al. 2022), which

contains hard distractors nearly identical to the images to be retrieved. We find that the neural retriever has much higher performance when fed neural rather than human captions, despite the fact that the former, unlike the latter, were generated without awareness of the distractors that make the task hard. Even more remarkably, when the same neural captions are given to human subjects, their retrieval performance is almost at chance level. Our results thus add to the growing body of evidence that, even when the "language" of neural models resembles English, this superficial resemblance might be deeply misleading.

**Why Should Adversarial Perturbations be Imperceptible? Rethink the Research Paradigm in Adversarial NLP**
*Yangyi Chen, Hongcheng Gao, Ganqu CUI, Fanchao Qi, Longtao Huang, Zhiyuan Liu and Maosong Sun* 14:00-15:30 (Atrium)
Textual adversarial samples play important roles in multiple subfields of NLP research, including security, evaluation, explainability, and data augmentation. However, most work mixes all these roles, obscuring the problem definitions and research goals of the security role that aims to reveal the practical concerns of NLP models. In this paper, we rethink the research paradigm of textual adversarial samples in security scenarios. We discuss the deficiencies in previous work and propose our suggestions that the research on the Security-oriented adversarial NLP (SoadNLP) should: (1) evaluate their methods on security tasks to demonstrate the real-world concerns; (2) consider real-world attackers' goals, instead of developing impractical methods. To this end, we first collect, process, and release a security datasets collection Advbench. Then, we reformalize the task and adjust the emphasis on different goals in SoadNLP. Next, we propose a simple method based on heuristic rules that can easily fulfill the actual adversarial goals to simulate real-world attack methods. We conduct experiments on both the attack and the defense sides on Advbench. Experimental results show that our method has higher practical value, indicating that the research paradigm in SoadNLP may start from our new benchmark. All the code and data of Advbench can be obtained at `https://github.com/thunlp/Advbench`.

[CL] **Hierarchical Interpretation of Neural Text Classification**
*Hanqi Yan, Lin Gui and Yulan He* 14:00-15:30 (Atrium)
Recent years have witnessed increasing interest in developing interpretable models in Natural Language Processing (NLP). Most existing models aim at identifying input features such as words or phrases important for model predictions. Neural models developed in NLP, however, often compose word semantics in a hierarchical manner. As such, interpretation by words or phrases only cannot faithfully explain model decisions in text classification. This article proposes a novel Hierarchical Interpretable Neural Text classifier, called HINT, which can automatically generate explanations of model predictions in the form of label-associated topics in a hierarchical manner. Model interpretation is no longer at the word level, but built on topics as the basic semantic unit. Experimental results on both review datasets and news datasets show that our proposed approach achieves text classification results on par with existing state-of-the-art text classifiers, and generates interpretations more faithful to model predictions and better understood by humans than other interpretable neural text classifiers.

**Evidence > Intuition: Transferability Estimation for Encoder Selection**
*Elisa Bassignana, Max Müller-Eberstein, Mike Zhang and Barbara Plank* 14:00-15:30 (Atrium)
With the increase in availability of large pre-trained language models (LMs) in Natural Language Processing (NLP), it becomes critical to assess their fit for a specific target task a priori—as fine-tuning the entire space of available LMs is computationally prohibitive and unsustainable. However, encoder transferability estimation has received little to no attention in NLP. In this paper, we propose to generate quantitative evidence to predict which LM, out of a pool of models, will perform best on a target task without having to fine-tune all candidates. We provide a comprehensive study on LM ranking for 10 NLP tasks spanning the two fundamental problem types of classification and structured prediction. We adopt the state-of-the-art Logarithm of Maximum Evidence (LogME) measure from Computer Vision (CV) and find that it positively correlates with final LM performance in 94% of the setups. In the first study of its kind, we further compare transferability measures with the de facto standard of human practitioner ranking, finding that evidence from quantitative metrics is more robust than pure intuition and can help identify unexpected LM candidates.

**Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?**
*Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi and Luke Zettlemoyer* 14:00-15:30 (Atrium)
Large language models (LMs) are able to in-context learn—perform a new task via inference alone by conditioning on a few input-label pairs (demonstrations) and making predictions for new inputs. However, there has been little understanding of how the model learns and which aspects of the demonstrations contribute to end task performance. In this paper, we show that ground truth demonstrations are in fact not required—randomly replacing labels in the demonstrations barely hurts performance on a range of classification and multi-choce tasks, consistently over 12 different models including GPT-3. Instead, we find that other aspects of the demonstrations are the key drivers of end task performance, including the fact that they provide a few examples of (1) the label space, (2) the distribution of the input text, and (3) the overall format of the sequence. Together, our analysis provides a new way of understanding how and why in-context learning works, while opening up new questions about how much can be learned from large language models through inference alone.

**Efficiently Tuned Parameters Are Task Embeddings**
*Wangchunshu Zhou, Canwen Xu and Julian McAuley* 14:00-15:30 (Atrium)
Intermediate-task transfer can benefit a wide range of NLP tasks with properly selected source datasets. However, it is computationally infeasible to experiment with all intermediate transfer combinations, making choosing a useful source task a challenging problem. In this paper, we anticipate that task-specific parameters updated in parameter-efficient tuning methods are likely to encode task-specific information. Therefore, such parameters can be predictive for inter-task transferability. Thus, we propose to exploit these efficiently tuned parameters as off-the-shelf task embeddings for the efficient selection of source datasets for intermediate-task transfer. We experiment with 11 text classification tasks and 11 question answering tasks. Experimental results show that our approach consistently outperforms existing inter-task transferability prediction methods while being conceptually simple and computationally efficient. Our analysis also reveals that the ability of efficiently tuned parameters on transferability prediction is disentangled with their in-task performance. This allows us to use parameters from early checkpoints as task embeddings to further improve efficiency.

**Machine Translation Robustness to Natural Asemantic Variation**
*Jacob Bremerman, Xiang Ren and Jonathan May* 14:00-15:30 (Atrium)
Current Machine Translation (MT) models still struggle with more challenging input, such as noisy data and tail-end words and phrases. Several works have addressed this robustness issue by identifying specific categories of noise and variation then tuning models to perform better on them. An important yet under-studied category involves minor variations in nuance (non-typos) that preserve meaning w.r.t. the target language. We introduce and formalize this category as Natural Asemantic Variation (NAV) and investigate it in the context of MT robustness. We find that existing MT models fail when presented with NAV data, but we demonstrate strategies to improve performance on NAV by fine-tuning them with human-generated variations. We also show that NAV robustness can be transferred across languages and find that synthetic perturbations can achieve some but not all of the benefits of organic NAV data.

**SimQA: Detecting Simultaneous MT Errors through Word-by-Word Question Answering**
*HyoJung Han, Marine Carpuat and Jordan Boyd-Graber* 14:00-15:30 (Atrium)
Detractors of neural machine translation admit that while its translations are fluent, it sometimes gets key facts wrong. This is particularly

important in simultaneous interpretation where translations have to be provided as fast as possible: before a sentence is complete. Yet, evaluations of simultaneous machine translation (SimulMT) fail to capture if systems correctly translate the most salient elements of a question: people, places, and dates. To address this problem, we introduce a downstream word-by-word question answering evaluation task (SimQA): given a source language question, translate the question word by word into the target language, and answer as soon as possible. SimQA jointly measures whether the SimulMT models translate the question quickly and accurately, and can reveal shortcomings in existing neural systems—hallucinating or omitting facts.

**Sampling-Based Approximations to Minimum Bayes Risk Decoding for Neural Machine Translation**
*Bryan Eikema and Wilker Aziz* 14:00-15:30 (Atrium)
In NMT we search for the mode of the model distribution to form predictions. The mode and other high-probability translations found by beam search have been shown to often be inadequate in a number of ways. This prevents improving translation quality through better search, as these idiosyncratic translations end up selected by the decoding algorithm, a problem known as the beam search curse. Recently, an approximation to minimum Bayes risk (MBR) decoding has been proposed as an alternative decision rule that would likely not suffer from the same problems. We analyse this approximation and establish that it has no equivalent to the beam search curse. We then design approximations that decouple the cost of exploration from the cost of robust estimation of expected utility. This allows for much larger hypothesis spaces, which we show to be beneficial. We also show that mode-seeking strategies can aid in constructing compact sets of promising hypotheses and that MBR is effective in identifying good translations in them. We conduct experiments on three language pairs varying in amounts of resources available: English into and from German, Romanian, and Nepali.

[INDUSTRY] **Tractable & Coherent Multi-Document Summarization: Discrete Optimization of Multiple Neural Modeling Streams via Integer Linear Programming**
*Litton J Kurisinkel and Nancy Chen* 14:00-15:30 (Atrium)
One key challenge in multi-document summarization is the generated summary is often less coherent compared to single document summarization due to the larger heterogeneity of the input source content. In this work, we propose a generic framework to jointly consider coherence and informativeness in multi-document summarization and offers provisions to replace individual components based on the domain of source text. In particular, the framework characterizes coherence through verb transitions and entity mentions and takes advantage of syntactic parse trees and neural modeling for intra-sentential noise pruning. The framework cast the entire problem as an integer linear programming optimization problem with neural and non-neural models as linear components. We evaluate our method in the news and legal domains. The proposed approach consistently performs better than competitive baselines for both objective metrics and human evaluation.

[INDUSTRY] **Semi-supervised Adversarial Text Generation based on Seq2Seq models**
*Hieu Le, Dieu-Thu Le, Verena Weber, Chris Church, Kay Rottmann, Melanie Bradford and Peter Chin* 14:00-15:30 (Atrium)
To improve deep learning models' robustness, adversarial training has been frequently used in computer vision with satisfying results. However, adversarial perturbation on text have turned out to be more challenging due to the discrete nature of text. The generated adversarial text might not sound natural or does not preserve semantics, which is the key for real world applications where text classification is based on semantic meaning. In this paper, we describe a new way for generating adversarial samples by using pseudo-labeled in-domain text data to train a seq2seq model for adversarial generation and combine it with paraphrase detection. We showcase the benefit of our approach for a real-world Natural Language Understanding (NLU) task, which maps a user's request to an intent. Furthermore, we experiment with gradient-based training for the NLU task and try using token importance scores to guide the adversarial text generation. We show that our approach can generate realistic and relevant adversarial samples compared to other state-of-the-art adversarial training methods. Applying adversarial training using these generated samples helps the NLU model to recover up to 70% of these types of errors and makes the model more robust, especially in the tail distribution in a large scale real world application.

[INDUSTRY] **Is it out yet? Automatic Future Product Releases Extraction from Web Data**
*Gilad Fuchs, Ido Ben-Shaul and Matan Mandelbrod* 14:00-15:30 (Atrium)
Identifying the release of new products and their predicted demand in advance is highly valuable for E-Commerce marketplaces and retailers. The information of an upcoming product release is used for inventory management, marketing campaigns and pre-order suggestions. Often, the announcement of an upcoming product release is widely available in multiple web pages such as blogs, chats or news articles. However, to the best of our knowledge, an automatic system to extract future product releases from web data has not been presented. In this work we describe an ML-powered multi-stage pipeline to automatically identify future product releases and rank their predicted demand from unstructured pages across the whole web. Our pipeline includes a novel Longformer-based model which uses a global attention mechanism guided by pre-calculated Named Entity Recognition predictions related to product releases. The model training data is based on a new corpus of 30K web pages manually annotated to identify future product releases. We made the dataset openly available at https://doi.org/10.5281/zenodo.6894770.

[INDUSTRY] **SpeechNet: Weakly Supervised, End-to-End Speech Recognition at Industrial Scale**
*Raphael Tang, Karun Kumar, Gefei Yang, Akshat Pandey, Yajie Mao, Vladislav Belyaev, Madhuri Emmadi, Craig Murray, Ferhan Ture and Jimmy Lin* 14:00-15:30 (Atrium)
End-to-end automatic speech recognition systems represent the state of the art, but they rely on thousands of hours of manually annotated speech for training, as well as heavyweight computation for inference. Of course, this impedes commercialization since most companies lack vast human and computational resources. In this paper, we explore training and deploying an ASR system in the label-scarce, compute-limited setting. To reduce human labor, we use a third-party ASR system as a weak supervision source, supplemented with labeling functions derived from implicit user feedback. To accelerate inference, we propose to route production-time queries across a pool of CUDA graphs of varying input lengths, the distribution of which best matches the traffic's. Compared to our third-party ASR, we achieve a relative improvement in word-error rate of 8% and a speedup of 600%. Our system, called SpeechNet, currently serves 12 million queries per day on our voice-enabled smart television. To our knowledge, this is the first time a large-scale, Wav2vec-based deployment has been described in the academic literature.

[INDUSTRY] **Reinforced Question Rewriting for Conversational Question Answering**
*Zhiyu Chen, Jie Zhao, Anjie Fang, Besnik Fetahu, Oleg Rokhlenko and Shervin Malmasi* 14:00-15:30 (Atrium)
Conversational Question Answering (CQA) aims to answer questions contained within dialogues, which are not easily interpretable without context. Developing a model to rewrite conversational questions into self-contained ones is an emerging solution in industry settings as it allows using existing single-turn QA systems to avoid training a CQA model from scratch. Previous work trains rewriting models using human rewrites as supervision. However, such objectives are disconnected with QA models and therefore more human-like rewrites do not guarantee better QA performance. In this paper we propose using QA feedback to supervise the rewriting model with reinforcement learning. Experiments show that our approach can effectively improve QA performance over baselines for both extractive and retrieval QA. Furthermore, human evaluation shows that our method can generate more accurate and detailed rewrites when compared to human annotations.

[INDUSTRY] **PAIGE: Personalized Adaptive Interactions Graph Encoder for Query Rewriting in Dialogue Systems**

*Daniel Biś, Saurabh Gupta, Jie Hao, Xing Fan and Chenlei Guo* 14:00-15:30 (Atrium)
Unexpected responses or repeated clarification questions from conversational agents detract from the users' experience with technology meant to streamline their daily tasks. To reduce these frictions, Query Rewriting (QR) techniques replace transcripts of faulty queries with alternatives that lead to responses that satisfy the users' needs. Despite their successes, existing QR approaches are limited in their ability to fix queries that require considering users' personal preferences. We improve QR by proposing Personalized Adaptive Interactions Graph Encoder (PAIGE). PAIGE is the first QR architecture that jointly models user's affinities and query semantics end-to-end. The core idea is to represent previous user-agent interactions and world knowledge in a structured form — a heterogeneous graph — and apply message passing to propagate latent representations of users' affinities to refine utterance embeddings. Using these embeddings, PAIGE can potentially provide different rewrites given the same query for users with different preferences. Our model, trained without any human-annotated data, improves the rewrite retrieval precision of state-of-the-art baselines by 12.5–17.5% while having nearly ten times fewer parameters.

[INDUSTRY] **Fast Vocabulary Transfer for Language Model Compression**
*Leonidas Gee, Andrea Zugarini, Leonardo Rigutini and Paolo Torroni* 14:00-15:30 (Atrium)
Real-world business applications require a trade-off between language model performance and size. We propose a new method for model compression that relies on vocabulary transfer. We evaluate the method on various vertical domains and downstream tasks. Our results indicate that vocabulary transfer can be effectively used in combination with other compression techniques, yielding a significant reduction in model size and inference time while marginally compromising on performance.

[INDUSTRY] **Multimodal Context Carryover**
*Prashan Wanigasekara, Nalin Gupta, Fan Yang, Emre Barut, Zeynab Raeesy, Kechen Qin, Stephen Rawls, Xinyue Liu, Chengwei Su and Spurthi Sandiri* 14:00-15:30 (Atrium)
Multi-modality support has become an integral part of creating a seamless user experience with modern voice assistants with smart displays. Users refer to images, video thumbnails, or the accompanying text descriptions on the screen through voice communication with AI powered devices. This raises the need to either augment existing commercial voice only dialogue systems with state-of-the-art multimodal components, or to introduce entirely new architectures; where the latter can lead to costly system revamps. To support the emerging visual navigation and visual product selection use cases, we propose to augment commercially deployed voice-only dialogue systems with additional multi-modal components. In this work, we present a novel yet pragmatic approach to expand an existing dialogue-based context carryover system (Chen et al., 2019a) in a voice assistant with state-of-the-art multimodal components to facilitate quick delivery of visual modality support with minimum changes. We demonstrate a 35% accuracy improvement over the existing system on an in-house multi-modal visual navigation data set.

**Data-Efficient Strategies for Expanding Hate Speech Detection into Under-Resourced Languages**
*Paul Röttger, Debora Nozza, Federico Bianchi and Dirk Hovy* 14:00-15:30 (Atrium)
Hate speech is a global phenomenon, but most hate speech datasets so far focus on English-language content. This hinders the development of more effective hate speech detection models in hundreds of languages spoken by billions across the world. More data is needed, but annotating hateful content is expensive, time-consuming and potentially harmful to annotators. To mitigate these issues, we explore data-efficient strategies for expanding hate speech detection into under-resourced languages. In a series of experiments with mono- and multilingual models across five non-English languages, we find that 1) a small amount of target-language fine-tuning data is needed to achieve strong performance, 2) the benefits of using more such data decrease exponentially, and 3) initial fine-tuning on readily-available English data can partially substitute target-language data and improve model generalisability. Based on these findings, we formulate actionable recommendations for hate speech detection in low-resource language settings.

**Does Corpus Quality Really Matter for Low-Resource Languages?**
*Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de-Viñaspre and Aitor Soroa* 14:00-15:30 (Atrium)
The vast majority of non-English corpora are derived from automatically filtered versions of CommonCrawl. While prior work has identified major issues on the quality of these datasets (Kreutzer et al., 2021), it is not clear how this impacts downstream performance. Taking representation learning in Basque as a case study, we explore tailored crawling (manually identifying and scraping websites with high-quality content) as an alternative to filtering CommonCrawl. Our new corpus, called EusCrawl, is similar in size to the Basque portion of popular multilingual corpora like CC100 and mC4, yet it has a much higher quality according to native annotators. For instance, 66% of documents are rated as high-quality for EusCrawl, in contrast with <33% for both mC4 and CC100. Nevertheless, we obtain similar results on downstream NLU tasks regardless of the corpus used for pre-training. Our work suggests that NLU performance in low-resource languages is not primarily constrained by the quality of the data, and other factors like corpus size and domain coverage can play a more important role.

**Few-shot Learning with Multilingual Generative Language Models**
*Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov and Xian Li* 14:00-15:30 (Atrium)
Large-scale generative language models such as GPT-3 are competitive few-shot learners. While these models are known to be able to jointly represent many different languages, their training data is dominated by English, potentially limiting their cross-lingual generalization. In this work, we train multilingual generative language models on a corpus covering a diverse set of languages, and study their few- and zero-shot learning capabilities in a wide range of tasks. Our largest model with 7.5 billion parameters sets new state of the art in few-shot learning in more than 20 representative languages, outperforming GPT-3 of comparable size in multilingual commonsense reasoning (with +7.4% absolute accuracy improvement in 0-shot settings and +9.4% in 4-shot settings) and natural language inference (+5.4% in each of 0-shot and 4-shot settings). On the FLORES-101 machine translation benchmark, our model outperforms GPT-3 on 171 out of 182 directions with 32 training examples, while surpassing the official supervised baseline in 45 directions. We conduct an in-depth analysis of different multilingual prompting approaches, showing in particular that strong few-shot learning performance across languages can be achieved via cross-lingual transfer through both templates and demonstration examples.

**ScienceWorld: Is your Agent Smarter than a 5th Grader?**
*Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté and Prithviraj Ammanabrolu* 14:00-15:30 (Atrium)
We present ScienceWorld, a benchmark to test agents' scientific reasoning abilities in a new interactive text environment at the level of a standard elementary school science curriculum. Despite the transformer-based progress seen in question-answering and scientific text processing, we find that current models cannot reason about or explain learned science concepts in novel contexts. For instance, models can easily answer what the conductivity of a known material is but struggle when asked how they would conduct an experiment in a grounded environment to find the conductivity of an unknown material. This begs the question of whether current models are simply retrieving answers by way of seeing a large number of similar examples or if they have learned to reason about concepts in a reusable manner. We hypothesize that agents need to be grounded in interactive environments to achieve such reasoning capabilities. Our experiments provide empirical evidence supporting this hypothesis – showing that a 1.5 million parameter agent trained interactively for 100k steps outperforms a 11 billion parameter model statically trained for scientific question-answering and reasoning from millions of expert demonstrations.

**MedJEx: A Medical Jargon Extraction Model with Wiki's Hyperlink Span and Contextualized Masked Language Model Score**
*Sunjae Kwon, Zonghai Yao, Harmon Jordan, David Levy, Brian Corner and hong yu* 14:00-15:30 (Atrium)
This paper proposes a new natural language processing (NLP) application for identifying medical jargon terms potentially difficult for patients to comprehend from electronic health record (EHR) notes. We first present a novel and publicly available dataset with expert-annotated medical jargon terms from 18K+ EHR note sentences (MedJ). Then, we introduce a novel medical jargon extraction (MedJEx) model which has been shown to outperform existing state-of-the-art NLP models. First, MedJEx improved the overall performance when it was trained on an auxiliary Wikipedia hyperlink span dataset, where hyperlink spans provide additional Wikipedia articles to explain the spans (or terms), and then fine-tuned on the annotated MedJ data. Secondly, we found that a contextualized masked language model score was beneficial for detecting domain-specific unfamiliar jargon terms. Moreover, our results show that training on the auxiliary Wikipedia hyperlink span datasets improved six out of eight biomedical named entity recognition benchmark datasets. MedJEx is publicly available.

**TABS: Efficient Textual Adversarial Attack for Pre-trained NL Code Model Using Semantic Beam Search**
*YunSeok Choi, Hyojun Kim and Jee-Hyong Lee* 14:00-15:30 (Atrium)
As pre-trained models have shown successful performance in program language processing as well as natural language processing, adversarial attacks on these models also attract attention. However, previous works on black-box adversarial attacks generated adversarial examples in a very inefficient way with simple greedy search. They also failed to find out better adversarial examples because it was hard to reduce the search space without performance loss. In this paper, we propose TABS, an efficient beam search black-box adversarial attack method. We adopt beam search to find out better adversarial examples, and contextual semantic filtering to effectively reduce the search space. Contextual semantic filtering reduces the number of candidate adversarial words considering the surrounding context and the semantic similarity. Our proposed method shows good performance in terms of attack success rate, the number of queries, and semantic similarity in attacking models for two tasks: NL code search classification and retrieval tasks.

**VisToT: Vision-Augmented Table-to-Text Generation**
*Prajwal Gatti, Anand Mishra, Manish Gupta and Mithun Das Gupta* 14:00-15:30 (Atrium)
Table-to-text generation has been widely studied in the Natural Language Processing community in the recent years. We give a new perspective to this problem by incorporating signals from both tables as well as associated images to generate relevant text. While tables contain a structured list of facts, images are a rich source of unstructured visual information. For example, in the tourism domain, images can be used to infer knowledge such as the type of landmark (e.g., church), its architecture (e.g., Ancient Roman), and composition (e.g., white marble). Therefore, in this paper, we introduce the novel task of Vision-augmented Table-To-Text Generation (VisToT, defined as follows: given a table and an associated image, produce a descriptive sentence conditioned on the multimodal input. For the task, we present a novel multimodal table-to-text dataset, WikiLandmarks, covering 73,084 unique world landmarks. Further, we also present a competitive architecture, namely, VT3 that generates accurate sentences conditioned on the image and table pairs. Through extensive analyses and experiments, we show that visual cues from images are helpful in (i) inferring missing information from incomplete or sparse tables, and (ii) strengthening the importance of useful information from noisy tables for natural language generation. We make the code and data publicly available.

**PACIFIC: Towards Proactive Conversational Question Answering over Tabular and Textual Data in Finance**
*Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam and Tat-Seng Chua* 14:00-15:30 (Atrium)
To facilitate conversational question answering (CQA) over hybrid contexts in finance, we present a new dataset, named PACIFIC. Compared with existing CQA datasets, PACIFIC exhibits three key features: (i) proactivity, (ii) numerical reasoning, and (iii) hybrid context of tables and text. A new task is defined accordingly to study Proactive Conversational Question Answering (PCQA), which combines clarification question generation and CQA. In addition, we propose a novel method, namely UniPCQA, to adapt a hybrid format of input and output content in PCQA into the Seq2Seq problem, including the reformulation of the numerical reasoning process as code generation. UniPCQA performs multi-task learning over all sub-tasks in PCQA and incorporates a simple ensemble strategy to alleviate the error propagation issue in the multi-task learning by cross-validating top-$k$ sampled Seq2Seq outputs. We benchmark the PACIFIC dataset with extensive baselines and provide comprehensive evaluations on each sub-task of PCQA.

**MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition**
*David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson KALIPE, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, MBONING TCHIAZE Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia and Joyce Nakatumba-Nabende* 14:00-15:30 (Atrium)
African languages are spoken by over a billion people, but they are under-represented in NLP research and development. Multiple challenges exist, including the limited availability of annotated training and evaluation datasets as well as the lack of understanding of which settings, languages, and recently proposed methods like cross-lingual transfer will be effective. In this paper, we aim to move towards solutions for these challenges, focusing on the task of named entity recognition (NER). We present the creation of the largest to-date human-annotated NER dataset for 20 African languages. We study the behaviour of state-of-the-art cross-lingual transfer methods in an Africa-centric setting, empirically demonstrating that the choice of source transfer language significantly affects performance. While much previous work defaults to using English as the source language, our results show that choosing the best transfer language improves zero-shot F1 scores by an average of 14

**RuCoLA: Russian Corpus of Linguistic Acceptability**
*Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov and Ekaterina Artemova* 14:00-15:30 (Atrium)
Linguistic acceptability (LA) attracts the attention of the research community due to its many uses, such as testing the grammatical knowledge of language models and filtering implausible texts with acceptability classifiers. However, the application scope of LA in languages other than English is limited due to the lack of high-quality resources. To this end, we introduce the Russian Corpus of Linguistic Acceptability (RuCoLA), built from the ground up under the well-established binary LA approach. RuCoLA consists of 9.8k in-domain sentences from linguistic publications and 3.6k out-of-domain sentences produced by generative models. The out-of-domain set is created to facilitate the practical use of acceptability for improving language generation. Our paper describes the data collection protocol and presents a fine-grained analysis of acceptability classification experiments with a range of baseline approaches. In particular, we demonstrate that the most widely used language models still fall behind humans by a large margin, especially when detecting morphological and semantic errors. We release RuCoLA, the code of experiments, and a public leaderboard to assess the linguistic competence of language models for Russian.

**PHEE: A Dataset for Pharmacovigilance Event Extraction from Text**
*Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, Byron Wallace, Bino John, Nigel Greene, Joseph Kim and Yulan He* 14:00-15:30 (Atrium)
The primary goal of drug safety researchers and regulators is to promptly identify adverse drug reactions. Doing so may in turn prevent or reduce the harm to patients and ultimately improve public health. Evaluating and monitoring drug safety (i.e., pharmacovigilance) involves

analyzing an ever growing collection of spontaneous reports from health professionals, physicians, and pharmacists, and information voluntarily submitted by patients. In this scenario, facilitating analysis of such reports via automation has the potential to rapidly identify safety signals. Unfortunately, public resources for developing natural language models for this task are scant. We present PHEE, a novel dataset for pharmacovigilance comprising over 5000 annotated events from medical case reports and biomedical literature, making it the largest such public dataset to date. We describe the hierarchical event schema designed to provide coarse and fine-grained information about patients' demographics, treatments and (side) effects. Along with the discussion of the dataset, we present a thorough experimental evaluation of current state-of-the-art approaches for biomedical event extraction, point out their limitations, and highlight open challenges to foster future research in this area.

**DEMETR: Diagnosing Evaluation Metrics for Translation**
*Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta and Mohit Iyyer* 14:00-15:30 (Atrium)
While machine translation evaluation metrics based on string overlap (e.g., BLEU) have their limitations, their computations are transparent: the BLEU score assigned to a particular candidate translation can be traced back to the presence or absence of certain words. The operations of newer learned metrics (e.g., BLEURT, COMET), which leverage pretrained language models to achieve higher correlations with human quality judgments than BLEU, are opaque in comparison. In this paper, we shed light on the behavior of these learned metrics by creating DEMETR, a diagnostic dataset with 31K English examples (translated from 10 source languages) for evaluating the sensitivity of MT evaluation metrics to 35 different linguistic perturbations spanning semantic, syntactic, and morphological error categories. All perturbations were carefully designed to form minimal pairs with the actual translation (i.e., differ in only one aspect). We find that learned metrics perform substantially better than string-based metrics on DEMETR. Additionally, learned metrics differ in their sensitivity to various phenomena (e.g., BERTScore is sensitive to untranslated words but relatively insensitive to gender manipulation, while COMET is much more sensitive to word repetition than to aspectual changes). We publicly release DEMETR to spur more informed future development of machine translation evaluation metrics

**Agent-Specific Deontic Modality Detection in Legal Language**
*Abhilasha Sancheti, Aparna Garimella, Balaji Vasan Srinivasan and Rachel Rudinger* 14:00-15:30 (Atrium)
Legal documents are typically long and written in legalese, which makes it particularly difficult for laypeople to understand their rights and duties. While natural language understanding technologies can be valuable in supporting such understanding in the legal domain, the limited availability of datasets annotated for deontic modalities in the legal domain, due to the cost of hiring experts and privacy issues, is a bottleneck. To this end, we introduce, LEXDEMOD, a corpus of English contracts annotated with deontic modality expressed with respect to a contracting party or agent along with the modal triggers. We benchmark this dataset on two tasks: (i) agent-specific multi-label deontic modality classification, and (ii) agent-specific deontic modality and trigger span detection using Transformer-based (Vaswani et al., 2017) language models. Transfer learning experiments show that the linguistic diversity of modal expressions in LEXDEMOD generalizes reasonably from lease to employment and rental agreements. A small case study indicates that a model trained on LEXDEMOD can detect red flags with high recall. We believe our work offers a new research direction for deontic modality detection in the legal domain.

**AX-MABSA: A Framework for Extremely Weakly Supervised Multi-label Aspect Based Sentiment Analysis**
*Sabyasachi Kamila, Walid Magdy, Sourav Dutta and MingXue Wang* 14:00-15:30 (Atrium)
Aspect Based Sentiment Analysis is a dominant research area with potential applications in social media analytics, business, finance, and health. Prior works in this area are primarily based on supervised methods, with a few techniques using weak supervision limited to predicting a single aspect category per review sentence. In this paper, we present an extremely weakly supervised multi-label Aspect Category Sentiment Analysis framework which does not use any labelled data. We only rely on a single word per class as an initial indicative information. We further propose an automatic word selection technique to choose these seed categories and sentiment words. We explore unsupervised language model post-training to improve the overall performance, and propose a multi-label generator model to generate multiple aspect category-sentiment pairs per review sentence. Experiments conducted on four benchmark datasets showcase our method to outperform other weakly supervised baselines by a significant margin.

**McQueen: a Benchmark for Multimodal Conversational Query Rewrite**
*Yifei Yuan, Chen Shi, Runze Wang, Liyi Chen, Feijun Jiang, Yuan You and Wai Lam* 14:00-15:30 (Atrium)
The task of query rewrite aims to convert an in-context query to its fully-specified version where ellipsis and coreference are completed and referred-back according to the history context. Although much progress has been made, less efforts have been paid to real scenario conversations that involve drawing information from more than one modalities. In this paper, we propose the task of multimodal conversational query rewrite (McQR), which performs query rewrite under the multimodal visual conversation setting. We collect a large-scale dataset named McQueen based on manual annotation, which contains 15k visual conversations and over 80k queries where each one is associated with a fully-specified rewrite version. In addition, for entities appearing in the rewrite, we provide the corresponding image box annotation. We then use the McQueen dataset to benchmark a state-of-the-art method for effectively tackling the McQR task, which is based on a multimodal pre-trained model with pointer generator. Extensive experiments are performed to demonstrate the effectiveness of our model on this task.

**mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections**
*Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou and Luo Si* 14:00-15:30 (Atrium)
Large-scale pre-trained foundation models have been an emerging paradigm for building artificial intelligence (AI) systems, which can be quickly adapted to a wide range of downstream tasks. This paper presents mPLUG, a new vision-language foundation model for both cross-modal understanding and generation. Most existing pre-trained models suffer from inefficiency and linguistic signal overwhelmed by long visual sequences in cross-modal alignment. To address both problems, mPLUG introduces an effective and efficient vision-language architecture with novel cross-modal skip-connections.

mPLUG is pre-trained end-to-end on large-scale image-text pairs with both discriminative and generative objectives. It achieves state-of-the-art results on a wide range of vision-language downstream tasks, including image captioning, image-text retrieval, visual grounding and visual question answering. mPLUG also demonstrates strong zero-shot transferability on vision-language and video-language tasks. The code and pre-trained models are available at https://github.com/alibaba/AliceMind

**Why Do You Feel This Way? Summarizing Triggers of Emotions in Social Media Posts**
*Hongli Zhan, Tiberiu Sosea, Cornelia Caragea and Junyi Jessy Li* 14:00-15:30 (Atrium)
Crises such as the COVID-19 pandemic continuously threaten our world and emotionally affect billions of people worldwide in distinct ways. Understanding the triggers leading to people's emotions is of crucial importance. Social media posts can be a good source of such analysis, yet these texts tend to be charged with multiple emotions, with triggers scattering across multiple sentences. This paper takes a novel angle, namely, emotion detection and trigger summarization, aiming to both detect perceived emotions in text, and summarize events and their appraisals that trigger each emotion. To support this goal, we introduce CovidET (Emotions and their Triggers during Covid-19), a dataset of 1,900 English Reddit posts related to COVID-19, which contains manual annotations of perceived emotions and abstractive summaries of their triggers described in the post. We develop strong baselines to jointly detect emotions and summarize emotion triggers. Our analyses

show that CovidET presents new challenges in emotion-specific summarization, as well as multi-emotion detection in long social media posts.

**Bridging Fairness and Environmental Sustainability in Natural Language Processing**
*Marius Hessenthaler, Emma Strubell, Dirk Hovy and Anne Lauscher* 14:00-15:30 (Atrium)
Fairness and environmental impact are important research directions for the sustainable development of artificial intelligence. However, while each topic is an active research area in natural language processing (NLP), there is a surprising lack of research on the interplay between the two fields. This lacuna is highly problematic, since there is increasing evidence that an exclusive focus on fairness can actually hinder environmental sustainability, and vice versa. In this work, we shed light on this crucial intersection in NLP by (1) investigating the efficiency of current fairness approaches through surveying example methods for reducing unfair stereotypical bias from the literature, and (2) evaluating a common technique to reduce energy consumption (and thus environmental impact) of English NLP models, knowledge distillation (KD), for its impact on fairness. In this case study, we evaluate the effect of important KD factors, including layer and dimensionality reduction, with respect to: (a) performance on the distillation task (natural language inference and semantic similarity prediction), and (b) multiple measures and dimensions of stereotypical bias (e.g., gender bias measured via the Word Embedding Association Test). Our results lead us to clarify current assumptions regarding the effect of KD on unfair bias: contrary to other findings, we show that KD can actually decrease model fairness.

**Modal-specific Pseudo Query Generation for Video Corpus Moment Retrieval**
*Minjoon Jung, SeongHo Choi, JooChan Kim, Jin-Hwa Kim and Byoung-Tak Zhang* 14:00-15:30 (Atrium)
Video corpus moment retrieval (VCMR) is the task to retrieve the most relevant video moment from a large video corpus using a natural language query. For narrative videos, e.g., drama or movies, the holistic understanding of temporal dynamics and multimodal reasoning are crucial. Previous works have shown promising results; however, they relied on the expensive query annotations for the VCMR, i.e., the corresponding moment intervals. To overcome this problem, we propose a self-supervised learning framework: Modal-specific Pseudo Query Generation Network (MPGN). First, MPGN selects candidate temporal moments via subtitle-based moment sampling. Then, it generates pseudo queries exploiting both visual and textual information from the selected temporal moments. Through the multimodal information in the pseudo queries, we show that MPGN successfully learns to localize the video corpus moment without any explicit annotation. We validate the effectiveness of MPGN on TVR dataset, showing the competitive results compared with both supervised models and unsupervised setting models.

**Robustifying Sentiment Classification by Maximally Exploiting Few Counterfactuals**
*Maarten De Raedt, Fréderic Godin, Chris Develder and Thomas Demeester* 14:00-15:30 (Atrium)
For text classification tasks, finetuned language models perform remarkably well. Yet, they tend to rely on spurious patterns in training data, thus limiting their performance on out-of-distribution (OOD) test data. Among recent models aiming to avoid this spurious pattern problem, adding extra counterfactual samples to the training data has proven to be very effective. Yet, counterfactual data generation is costly since it relies on human annotation. Thus, we propose a novel solution that only requires annotation of a small fraction (e.g., 1%) of the original training data, and uses automatic generation of extra counterfactuals in an encoding vector space. We demonstrate the effectiveness of our approach in sentiment classification, using IMDb data for training and other sets for OOD tests (i.e., Amazon, SemEval and Yelp). We achieve noticeable accuracy improvements by adding only 1% manual counterfactuals: +3% compared to adding +100% in-distribution training samples, +1.3% compared to alternate counterfactual approaches.

## Demo Session 2

14:00-15:30 (Link Admin)

[DEMO] **JoeyS2T: Minimalistic Speech-to-Text Modeling with JoeyNMT**
*Mayumi Ohta, Julia Kreutzer and Stefan Riezler* 14:00-15:30 (Link Admin)
JoeyS2T is a JoeyNMT extension for speech-to-text tasks such as automatic speech recognition and end-to-end speech translation. It inherits the core philosophy of JoeyNMT, a minimalist NMT toolkit built on PyTorch, seeking simplicity and accessibility. JoeyS2T's workflow is self-contained, starting from data pre-processing, over model training and prediction to evaluation, and is seamlessly integrated into JoeyNMT's compact and simple code base. On top of JoeyNMT's state-of-the-art Transformer-based Encoder-Decoder architecture, JoeyS2T provides speech-oriented components such as convolutional layers, SpecAugment, CTC-loss, and WER evaluation. Despite its simplicity compared to prior implementations, JoeyS2T performs competitively on English speech recognition and English-to-German speech translation benchmarks. The implementation is accompanied by a walk-through tutorial and available on `https://github.com/may-/joeys2t`.

[DEMO] **Camelira: An Arabic Multi-Dialect Morphological Disambiguator**
*Ossama Obeid, Go Inoue and Nizar Habash* 14:00-15:30 (Link Admin)
We present Camelira, a web-based Arabic multi-dialect morphological disambiguation tool that covers four major variants of Arabic: Modern Standard Arabic, Egyptian, Gulf, and Levantine. Camelira offers a user-friendly web interface that allows researchers and language learners to explore various linguistic information, such as part-of-speech, morphological features, and lemmas. Our system also provides an option to automatically choose an appropriate dialect-specific disambiguator based on the prediction of a dialect identification component. Camelira is publicly accessible at http://camelira.camel-lab.com.

[DEMO] **TweetNLP: Cutting-Edge Natural Language Processing for Social Media**
*Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, Eugenio Martínez-Cámara, Gonzalo Medina, Thomas Buhrmann, Leonardo Neves and Francesco Barbieri* 14:00-15:30 (Link Admin)
In this paper we present TweetNLP, an integrated platform for Natural Language Processing (NLP) in social media. TweetNLP supports a diverse set of NLP tasks, including generic focus areas such as sentiment analysis and named entity recognition, as well as social media-specific tasks such as emoji prediction and offensive language identification. Task-specific systems are powered by reasonably-sized Transformer-based language models specialized on social media text (in particular, Twitter) which can be run without the need for dedicated hardware or cloud services. The main contributions of TweetNLP are: (1) an integrated Python library for a modern toolkit supporting social media analysis using our various task-specific models adapted to the social domain; (2) an interactive online demo for codeless experimentation using our models; and (3) a tutorial covering a wide variety of typical social media applications.

# Session 4 - 16:00-17:30

## Virtual Portal 1

16:00-17:30 (Hall A, Room A)

**ReCo: Reliable Causal Chain Reasoning via Structural Causal Recurrent Neural Networks**
*Kai Xiong, Xiao Ding, Zhongyang Li, Li Du, Ting Liu, Bing Qin, Yi Zheng and baoxing Huai* 16:00-17:30 (Hall A, Room A)
Causal chain reasoning (CCR) is an essential ability for many decision-making AI systems, which requires the model to build reliable causal chains by connecting causal pairs. However, CCR suffers from two main transitive problems: threshold effect and scene drift. In other words, the causal pairs to be spliced may have a conflicting threshold boundary or scenario. To address these issues, we propose a novel Reliable Causal chain reasoning framework (ReCo), which introduces exogenous variables to represent the threshold and scene factors of each causal pair within the causal chain, and estimates the threshold and scene contradictions across exogenous variables via structural causal recurrent neural networks (SRNN). Experiments show that ReCo outperforms a series of strong baselines on both Chinese and English CCR datasets. Moreover, by injecting reliable causal chain knowledge distilled by ReCo, BERT can achieve better performances on four downstream causal-related tasks than BERT models enhanced by other kinds of knowledge.

**A Sequential Flow Control Framework for Multi-hop Knowledge Base Question Answering**
*Minghui Xie, Chuzhan Hao and Peng Zhang* 16:00-17:30 (Hall A, Room A)
One of the key challenges of knowledge base question answering (KBQA) is the multi-hop reasoning. Since in different hops, one attends to different parts of question, it is important to dynamically represent the question semantics for each hop. Existing methods, however, (i) infer the dynamic question representation only through coarse-grained attention mechanisms, which may bring information loss, (ii) and have not effectively modeled the sequential logic, which is crucial for the multi-hop reasoning process in KBQA. To address these issues, we propose a sequential reasoning self-attention mechanism to capture the crucial reasoning information of each single hop in a more fine-grained way. Based on Gated Recurrent Unit (GRU) which is good at modeling sequential process, we propose a simple but effective GRU-inspired Flow Control (GFC) framework to model sequential logic in the whole multi-hop process. Extensive experiments on three popular benchmark datasets have demonstrated the superior effectiveness of our model. In particular, GFC achieves new state-of-the-art Hits@1 of 76.8% on WebQSP and is also effective when KB is incomplete. Our code and data are available at https://github.com/Xie-Minghui/GFC.

**Identifying Physical Object Use in Sentences**
*Tianyu Jiang and Ellen Riloff* 16:00-17:30 (Hall A, Room A)
Commonsense knowledge about the typical functions of physical objects allows people to make inferences during sentence understanding. For example, we infer that "Sam enjoyed the book" means that Sam enjoyed reading the book, even though the action is implicit. Prior research has focused on learning the prototypical functions of physical objects in order to enable inferences about implicit actions. But many sentences refer to objects even when they are not used (e.g., "The book fell"). We argue that NLP systems need to recognize whether an object is being used before inferring how the object is used. We define a new task called Object Use Classification that determines whether a physical object mentioned in a sentence was used or likely will be used. We introduce a new dataset for this task and present a classification model that exploits data augmentation methods and FrameNet when fine-tuning a pre-trained language model. We also show that object use classification combined with knowledge about the prototypical functions of objects has the potential to yield very good inferences about implicit and anticipated actions.

**Sequence Models for Document Structure Identification in an Undeciphered Script**
*Logan Born, M. Monroe, Kathryn Kelley and Anoop Sarkar* 16:00-17:30 (Hall A, Room A)
This work describes the first thorough analysis of "header" signs in proto-Elamite, an undeciphered script from 3100-2900 BCE. Headers are a category of signs which have been provisionally identified through painstaking manual analysis of this script by domain experts. We use unsupervised neural and statistical sequence modeling techniques to provide new and independent evidence for the existence of headers, without supervision from domain experts. Having affirmed the existence of headers as a legitimate structural feature, we next arrive at a richer understanding of their possible meaning and purpose by (i) examining which features predict their presence; (ii) identifying correlations between these features and other document properties; and (iii) examining cases where these features predict the presence of a header in texts where domain experts do not expect one (or vice versa). We provide more concrete processes for labeling headers in this corpus and a clearer justification for existing intuitions about document structure in proto-Elamite.

**Sentence-level Media Bias Analysis Informed by Discourse Structures**
*Yuanyuan Lei, Ruihong Huang, Lu Wang and Nick Beauchamp* 16:00-17:30 (Hall A, Room A)
As polarization continues to rise among both the public and the news media, increasing attention has been devoted to detecting media bias. Most recent work in the NLP community, however, identify bias at the level of individual articles. However, each article itself comprises multiple sentences, which vary in their ideological bias. In this paper, we aim to identify sentences within an article that can illuminate and explain the overall bias of the entire article. We show that understanding the discourse role of a sentence in telling a news story, as well as its relation with nearby sentences, can reveal the ideological leanings of an author even when the sentence itself appears merely neutral. In particular, we consider using a functional news discourse structure and PDTB discourse relations to inform bias sentence identification, and distill the auxiliary knowledge from the two types of discourse structure into our bias sentence identification system. Experimental results on benchmark datasets show that incorporating both the global functional discourse structure and local rhetorical discourse relations can effectively increase the recall of bias sentence identification by 8.27% - 8.62%, as well as increase the precision by 2.82% - 3.48%.

**META-GUI: Towards Multi-modal Conversational Agents on Mobile GUI**
*Liangtai Sun, Xingyu Chen, Lu Chen, Tianle Dai, Zichen Zhu and Kai Yu* 16:00-17:30 (Hall A, Room A)
Task-oriented dialogue (TOD) systems have been widely used by mobile phone intelligent assistants to accomplish tasks such as calendar scheduling or hotel reservation. Current TOD systems usually focus on multi-turn text/speech interaction, then they would call back-end APIs designed for TODs to perform the task. However, this API-based architecture greatly limits the information-searching capability of intelligent assistants and may even lead to task failure if TOD-specific APIs are not available or the task is too complicated to be executed by the provided APIs. In this paper, we propose a new TOD architecture: GUI-based task-oriented dialogue system (GUI-TOD). A GUI-TOD system can directly perform GUI operations on real APPs and execute tasks without invoking TOD-specific backend APIs. Furthermore, we release META-GUI, a dataset for training a Multi-modal convErsaTional Agent on mobile GUI. We also propose a multi-model action prediction and response model, which show promising results on META-GUI. The dataset, codes and leaderboard are publicly available.

**UniNL: Aligning Representation Learning with Scoring Function for OOD Detection via Unified Neighborhood Learning**
*Yutao Mou, Pei Wang, Keqing He, Yanan Wu, Jingang Wang, Wei Wu and Weiran Xu* 16:00-17:30 (Hall A, Room A)
Detecting out-of-domain (OOD) intents from user queries is essential for avoiding wrong operations in task-oriented dialogue systems. The key challenge is how to distinguish in-domain (IND) and OOD intents. Previous methods ignore the alignment between representation learning and scoring function, limiting the OOD detection performance. In this paper, we propose a unified neighborhood learning framework

(UniNL) to detect OOD intents. Specifically, we design a KNCL objective for representation learning, and introduce a KNN-based scoring function for OOD detection. We aim to align representation learning with scoring function. Experiments and analysis on two benchmark datasets show the effectiveness of our method.

**Prompt Conditioned VAE: Enhancing Generative Replay for Lifelong Learning in Task-Oriented Dialogue**
*Yingxiu Zhao, Yinhe Zheng, Zhiliang Tian, Chang Gao, Jian Sun and Nevin L. Zhang* 16:00-17:30 (Hall A, Room A)
Lifelong learning (LL) is vital for advanced task-oriented dialogue (ToD) systems. To address the catastrophic forgetting issue of LL, generative replay methods are widely employed to consolidate past knowledge with generated pseudo samples. However, most existing generative replay methods use only a single task-specific token to control their models. This scheme is usually not strong enough to constrain the generative model due to insufficient information involved. In this paper, we propose a novel method, prompt conditioned VAE for lifelong learning (PCLL), to enhance generative replay by incorporating tasks' statistics. PCLL captures task-specific distributions with a conditional variational autoencoder, conditioned on natural language prompts to guide the pseudo-sample generation. Moreover, it leverages a distillation process to further consolidate past knowledge by alleviating the noise in pseudo samples. Experiments on natural language understanding tasks of ToD systems demonstrate that PCLL significantly outperforms competitive baselines in building lifelong learning models.

**End-to-End Neural Discourse Deixis Resolution in Dialogue**
*Shengjie Li and Vincent Ng* 16:00-17:30 (Hall A, Room A)
We adapt Lee et al.'s (2018) span-based entity coreference model to the task of end-to-end discourse deixis resolution in dialogue, specifically by proposing extensions to their model that exploit task-specific characteristics. The resulting model, dd-utt, achieves state-of-the-art results on the four datasets in the CODI-CRAC 2021 shared task.

**Sparse Teachers Can Be Dense with Knowledge**
*Yi Yang, Chen Zhang and Dawei Song* 16:00-17:30 (Hall A, Room A)
Recent advances in distilling pretrained language models have discovered that, besides the expressiveness of knowledge, the student-friendliness should be taken into consideration to realize a truly knowledgeable teacher. Based on a pilot study, we find that over-parameterized teachers can produce expressive yet student-unfriendly knowledge and are thus limited in overall knowledgeableness. To remove the parameters that result in student-unfriendliness, we propose a sparse teacher trick under the guidance of an overall knowledgeable score for each teacher parameter. The knowledgeable score is essentially an interpolation of the expressiveness and student-friendliness scores. The aim is to ensure that the expressive parameters are retained while the student-unfriendly ones are removed. Extensive experiments on the GLUE benchmark show that the proposed sparse teachers can be dense with knowledge and lead to students with compelling performance in comparison with a series of competitive baselines.

**Vector-Quantized Input-Contextualized Soft Prompts for Natural Language Understanding**
*Rishabh Bhardwaj, Amrita Saha, Steven C.H. Hoi and Soujanya Poria* 16:00-17:30 (Hall A, Room A)
Prompt Tuning has been largely successful as a parameter-efficient method of conditioning large-scale pre-trained language models to perform downstream tasks. Thus far, soft prompt tuning learns a fixed set of task-specific continuous vectors, i.e., soft tokens that remain static across the task samples. A fixed prompt, however, may not generalize well to the diverse kinds of inputs the task comprises. In order to address this, we propose Vector-quantized Input-contextualized Prompts (VIP) as an extension to the soft prompt tuning framework. VIP particularly focuses on two aspects—contextual prompts that learns input-specific contextualization of the soft prompt tokens through a small-scale sentence encoder and quantized prompts that maps the contextualized prompts to a set of learnable codebook vectors through a Vector quantization network. On various language understanding tasks like SuperGLUE, QA, Relation classification, NER and NLI, VIP outperforms the soft prompt tuning (PT) baseline by an average margin of 1.19%. Further, our generalization studies show that VIP learns more robust prompt representations, surpassing PT by a margin of 0.6% - 5.3% on Out-of-domain QA and NLI tasks respectively, and by 0.75% on Multi-Task setup over 4 tasks spanning across 12 domains.

**"I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset**
*Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani and Adina Williams* 16:00-17:30 (Hall A, Room A)
As language models grow in popularity, it becomes increasingly important to clearly measure all possible markers of demographic identity in order to avoid perpetuating existing societal harms. Many datasets for measuring bias currently exist, but they are restricted in their coverage of demographic axes and are commonly used with preset bias tests that presuppose which types of biases models can exhibit. In this work, we present a new, more inclusive bias measurement dataset, HolisticBias, which includes nearly 600 descriptor terms across 13 different demographic axes. HolisticBias was assembled in a participatory process including experts and community members with lived experience of these terms. These descriptors combine with a set of bias measurement templates to produce over 450,000 unique sentence prompts, which we use to explore, identify, and reduce novel forms of bias in several generative models. We demonstrate that HolisticBias is effective at measuring previously undetectable biases in token likelihoods from language models, as well as in an offensiveness classifier. We will invite additions and amendments to the dataset, which we hope will serve as a basis for more easy-to-use and standardized methods for evaluating bias in NLP models.

**BERTScore is Unfair: On Social Bias in Language Model-Based Metrics for Text Generation**
*Tianxiang Sun, Junliang He, Xipeng Qiu and Xuanjing Huang* 16:00-17:30 (Hall A, Room A)
Automatic evaluation metrics are crucial to the development of generative systems. In recent years, pre-trained language model (PLM) based metrics, such as BERTScore, have been commonly adopted in various generation tasks. However, it has been demonstrated that PLMs encode a range of stereotypical societal biases, leading to a concern about the fairness of PLMs as metrics. To that end, this work presents the first systematic study on the social bias in PLM-based metrics. We demonstrate that popular PLM-based metrics exhibit significantly higher social bias than traditional metrics on 6 sensitive attributes, namely race, gender, religion, physical appearance, age, and socioeconomic status. In-depth analysis suggests that choosing paradigms (matching, regression, or generation) of the metric has a greater impact on fairness than choosing PLMs. In addition, we develop debiasing adapters that are injected into PLM layers, mitigating bias in PLM-based metrics while retaining high performance for evaluating text generation.

**TextFusion: Privacy-Preserving Pre-trained Model Inference via Token Fusion**
*Xin Zhou, Jinzhu Lu, Tao Gui, Ruotian Ma, Zichu Fei, Yuran Wang, Yong Ding, Yibo Cheung, Qi Zhang and Xuanjing Huang* 16:00-17:30 (Hall A, Room A)
Recently, more and more pre-trained language models are released as a cloud service. It allows users who lack computing resources to perform inference with a powerful model by uploading data to the cloud. The plain text may contain private information, as the result, users prefer to do partial computations locally and upload intermediate representations to the cloud for subsequent inference. However, recent studies have shown that intermediate representations can also be recovered to plain text with reasonable accuracy, thus the risk of privacy leakage still exists. To address this issue, we propose TextFusion, a novel method for preserving inference privacy. Specifically, we train a Fusion Predictor to dynamically fuse token representations, which hides multiple private token representations behind an unrecognizable one. Furthermore, an adversarial training regime is employed to privatize these representations. In this way, the cloud only receives incomplete and perturbed

representations, making it difficult to accurately recover the complete plain text. The experimental results on diverse classification tasks show that our approach can effectively preserve inference privacy without significantly sacrificing performance in different scenarios.

[DEMO] **Twitter-Demographer: A Flow-based Tool to Enrich Twitter Data**
*Federico Bianchi, Vincenzo Cutrona and Dirk Hovy* 16:00-17:30 (Hall A, Room A)
Twitter data have become essential to Natural Language Processing (NLP) and social science research, driving various scientific discoveries in recent years. However, the textual data alone are often not enough to conduct studies: especially, social scientists need more variables to perform their analysis and control for various factors. How we augment this information, such as users' location, age, or tweet sentiment, has ramifications for anonymity and reproducibility, and requires dedicated effort. This paper describes Twitter-Demographer, a simple, flow-based tool to enrich Twitter data with additional information about tweets and users. This tool is aimed at NLP practitioners, psycho-linguists, and (computational) social scientists who want to enrich their datasets with aggregated information, facilitating reproducibility, and providing algorithmic privacy-by-design measures for pseudo-anonymity. We discuss our design choices, inspired by the flow-based programming paradigm, to use black-box components that can easily be chained together and extended. We also analyze the ethical issues related to the use of this tool, and the built-in measures to facilitate pseudo-anonymity.

[DEMO] **Hands-On Interactive Neuro-Symbolic NLP with DRaiL**
*Maria Leonor Pacheco, Shamik Roy and Dan Goldwasser* 16:00-17:30 (Hall A, Room A)
We recently introduced DRaiL, a declarative neural-symbolic modeling framework designed to support a wide variety of NLP scenarios. In this paper, we enhance DRaiL with an easy to use Python interface, equipped with methods to define, modify and augment DRaiL models interactively, as well as with methods to debug and visualize the predictions made. We demonstrate this interface with a challenging NLP task: predicting sentence and entity level moral sentiment in political tweets.

## Virtual Portal 2

16:00-17:30 (Hall A, Room B)

**Bi-Directional Iterative Prompt-Tuning for Event Argument Extraction**
*Lu Dai, Bang Wang, Wei Xiang and yijun mo* 16:00-17:30 (Hall A, Room B)
Recently, prompt-tuning has attracted growing interests in event argument extraction (EAE). However, the existing prompt-tuning methods have not achieved satisfactory performance due to the lack of consideration of entity information. In this paper, we propose a bi-directional iterative prompt-tuning method for EAE, where the EAE task is treated as a cloze-style task to take full advantage of entity information and pre-trained language models (PLMs). Furthermore, our method explores event argument interactions by introducing the argument roles of contextual entities into prompt construction. Since template and verbalizer are two crucial components in a cloze-style prompt, we propose to utilize the role label semantic knowledge to construct a semantic verbalizer and design three kind of templates for the EAE task. Experiments on the ACE 2005 English dataset with standard and low-resource settings show that the proposed method significantly outperforms the peer state-of-the-art methods.

**Attention and Edge-Label Guided Graph Convolutional Networks for Named Entity Recognition**
*Renjie Zhou, Zhongyi Xie, Jian Wan, Jilin Zhang, Yong Liao and Qiang Liu* 16:00-17:30 (Hall A, Room B)
It has been shown that named entity recognition (NER) could benefit from incorporating the long-distance structured information captured by dependency trees. However, dependency trees built by tools usually have a certain percentage of errors. Under such circumstances, how to better use relevant structured information while ignoring irrelevant or wrong structured information from the dependency trees to improve NER performance is still a challenging research problem. In this paper, we propose the Attention and Edge-Label guided Graph Convolution Network (AELGCN) model. Then, we integrate it into BiLSTM-CRF to form BiLSTM-AELGCN-CRF model. We design an edge-aware node joint update module and introduce a node-aware edge update module to explore hidden in structured information entirely and solve the wrong dependency label information to some extent. After two modules, we apply attention-guided GCN, which automatically learns how to attend to the relevant structured information selectively. We conduct extensive experiments on several standard datasets across four languages and achieve better results than previous approaches. Through experimental analysis, it is found that our proposed model can better exploit the structured information on the dependency tree to improve the recognition of long entities.

**Open Relation and Event Type Discovery with Type Abstraction**
*Sha Li, Heng Ji and Jiawei Han* 16:00-17:30 (Hall A, Room B)
Conventional "closed-world" information extraction (IE) approaches rely on human ontologies to define the scope for extraction. As a result, such approaches fall short when applied to new domains. This calls for systems that can automatically infer new types from given corpora, a task which we refer to as type discovery. To tackle this problem, we introduce the idea of type abstraction, where the model is prompted to generalize and name the type. Then we use the similarity between inferred names to induce clusters. Observing that this abstraction-based representation is often complementary to the entity/trigger token representation, we set up these two representations as two views and design our model as a co-training framework. Our experiments on multiple relation extraction and event extraction datasets consistently show the advantage of our type abstraction approach.

**WR-One2Set: Towards Well-Calibrated Keyphrase Generation**
*Binbin Xie, Xiangpeng Wei, Baosong Yang, Huan Lin, Jun Xie, Xiaoli Wang, Min Zhang and Jinsong Su* 16:00-17:30 (Hall A, Room B)
Keyphrase generation aims to automatically generate short phrases summarizing an input document. The recently emerged ONE2SET paradigm (Ye et al., 2021) generates keyphrases as a set and has achieved competitive performance. Nevertheless, we observe serious calibration errors outputted by ONE2SET, especially in the over-estimation of $\emptyset$ token (means no corresponding keyphrase). In this paper, we deeply analyze this limitation and identify two main reasons behind: 1) the parallel generation has to introduce excessive $\emptyset$ as padding tokens into training instances; and 2) the training mechanism assigning target to each slot is unstable and further aggravates the $\emptyset$ token over-estimation. To make the model well-calibrated, we propose WR-ONE2SET which extends ONE2SET with an adaptive instance-level cost Weighting strategy and a target Re-assignment mechanism. The former dynamically penalizes the over-estimated slots for different instances thus smoothing the uneven training distribution. The latter refines the original inappropriate assignment and reduces the supervisory signals of over-estimated slots. Experimental results on commonly-used datasets demonstrate the effectiveness and generality of our proposed paradigm.

**OTSeq2Set: An Optimal Transport Enhanced Sequence-to-Set Model for Extreme Multi-label Text Classification**
*Jie Cao and Yin Zhang* 16:00-17:30 (Hall A, Room B)
Extreme multi-label text classification (XMTC) is the task of finding the most relevant subset labels from an extremely large-scale label collection. Recently, some deep learning models have achieved state-of-the-art results in XMTC tasks. These models commonly predict

scores for all labels by a fully connected layer as the last layer of the model. However, such models can't predict a relatively complete and variable-length label subset for each document, because they select positive labels relevant to the document by a fixed threshold or take top k labels in descending order of scores. A less popular type of deep learning models called sequence-to-sequence (Seq2Seq) focus on predicting variable-length positive labels in sequence style. However, the labels in XMTC tasks are essentially an unordered set rather than an ordered sequence, the default order of labels restrains Seq2Seq models in training. To address this limitation in Seq2Seq, we propose an autoregressive sequence-to-set model for XMTC tasks named OTSeq2Set. Our model generates predictions in student-forcing scheme and is trained by a loss function based on bipartite matching which enables permutation-invariance. Meanwhile, we use the optimal transport distance as a measurement to force the model to focus on the closest labels in semantic label space. Experiments show that OTSeq2Set outperforms other competitive baselines on 4 benchmark datasets. Especially, on the Wikipedia dataset with 31k labels, it outperforms the state-of-the-art Seq2Seq method by 16.34% in micro-F1 score. The code is available at https://github.com/caojie54/OTSeq2Set.

**Dimension Reduction for Efficient Dense Retrieval via Conditional Autoencoder**
*Zhenghao Liu, Han Zhang, Chenyan Xiong, Zhiyuan Liu, Yu Gu and Xiaohua Li* 16:00-17:30 (Hall A, Room B)
Dense retrievers encode queries and documents and map them in an embedding space using pre-trained language models. These embeddings need to be high-dimensional to fit training signals and guarantee the retrieval effectiveness of dense retrievers. However, these high-dimensional embeddings lead to larger index storage and higher retrieval latency. To reduce the embedding dimensions of dense retrieval, this paper proposes a Conditional Autoencoder (ConAE) to compress the high-dimensional embeddings to maintain the same embedding distribution and better recover the ranking features. Our experiments show that ConAE is effective in compressing embeddings by achieving comparable ranking performance with its teacher model and making the retrieval system more efficient. Our further analyses show that ConAE can alleviate the redundancy of the embeddings of dense retrieval with only one linear layer. All codes of this work are available at https://github.com/NEUIR/ConAE.

**A Framework for Adapting Pre-Trained Language Models to Knowledge Graph Completion**
*Justin Lovelace and Carolyn Rosé* 16:00-17:30 (Hall A, Room B)
Recent work has demonstrated that entity representations can be extracted from pre-trained language models to develop knowledge graph completion models that are more robust to the naturally occurring sparsity found in knowledge graphs. In this work, we conduct a comprehensive exploration of how to best extract and incorporate those embeddings into knowledge graph completion models. We explore the suitability of the extracted embeddings for direct use in entity ranking and introduce both unsupervised and supervised processing methods that can lead to improved downstream performance. We then introduce supervised embedding extraction methods that can extract more informative representations. We then synthesize our findings and develop a knowledge graph completion model that significantly outperforms recent neural models.

**A Unified Neural Network Model for Readability Assessment with Feature Projection and Length-Balanced Loss**
*Wenbiao Li, Wang Ziyang and Yunfang Wu* 16:00-17:30 (Hall A, Room B)
Readability assessment is a basic research task in the field of education. Traditional methods mainly employ machine learning classifiers with hundreds of linguistic features. Although the deep learning model has become the prominent approach for almost all NLP tasks, it is less explored for readability assessment. In this paper, we propose a BERT-based model with feature projection and length-balanced loss (BERT-FP-LBL) to determine the difficulty level of a given text. First, we introduce topic features guided by difficulty knowledge to complement the traditional linguistic features. From the linguistic features, we extract really useful orthogonal features to supplement BERT representations by means of projection filtering. Furthermore, we design a length-balanced loss to handle the greatly varying length distribution of the readability data. We conduct experiments on three English benchmark datasets and one Chinese dataset, and the experimental results show that our proposed model achieves significant improvements over baseline models. Interestingly, our proposed model achieves comparable results with human experts in consistency test.

**Recovering Gold from Black Sand: Multilingual Dense Passage Retrieval with Hard and False Negative Samples**
*Tianhao Shen, Mingtong Liu, Ming Zhou and Deyi Xiong* 16:00-17:30 (Hall A, Room B)
Negative samples have not been efficiently explored in multilingual dense passage retrieval. In this paper, we propose a novel multilingual dense passage retrieval framework, mHFN, to recover and utilize hard and false negative samples. mHFN consists of three key components: 1) a multilingual hard negative sample augmentation module that allows knowledge of indistinguishable passages to be shared across multiple languages and synthesizes new hard negative samples by interpolating representations of queries and existing hard negative samples, 2) a multilingual negative sample cache queue that stores negative samples from previous batches in each language to increase the number of multilingual negative samples used in training beyond the batch size limit, and 3) a lightweight adaptive false negative sample filter that uses generated pseudo labels to separate unlabeled false negative samples and converts them into positive passages in training. We evaluate mHFN on Mr. TyDi, a high-quality multilingual dense passage retrieval dataset covering eleven typologically diverse languages, and experimental results show that mHFN outperforms strong sparse, dense and hybrid baselines and achieves new state-of-the-art performance on all languages. Our source code is available at https://github.com/Magnetic2014/mHFN.

**Calibration Meets Explanation: A Simple and Effective Approach for Model Confidence Estimates**
*Dongfang Li, Baotian Hu and Qingcai Chen* 16:00-17:30 (Hall A, Room B)
Calibration strengthens the trustworthiness of black-box models by producing better accurate confidence estimates on given examples. However, little is known about if model explanations can help confidence calibration. Intuitively, humans look at important features attributions and decide whether the model is trustworthy. Similarly, the explanations may tell us when the model might know and when it does not. Inspired by this, we propose a method named CME that leverages model explanations to make the model less confident with non-inductive attributions. The idea is that when the model is not highly confident, it is difficult to identify strong indications of any class, and the tokens accordingly do not have high attribution scores for any class and vice versa. We conduct extensive experiments on six datasets with two popular pre-trained language models in the in-domain and out-of-domain settings. The results show that CME improves calibration performance in all settings. The expected calibration errors are further reduced when combined with temperature scaling. Our findings highlight that model explanations can help calibrate posterior estimates.

**Towards Interactivity and Interpretability: A Rationale-based Legal Judgment Prediction Framework**
*Yiquan Wu, Yifei Liu, Weiming Lu, Yating Zhang, Jun Feng, Changlong Sun, Fei Wu and Kun Kuang* 16:00-17:30 (Hall A, Room B)
Legal judgment prediction (LJP) is a fundamental task in legal AI, which aims to assist the judge to hear the case and determine the judgment. The legal judgment usually consists of the law article, charge, and term of penalty. In the real trial scenario, the judge usually makes the decision step-by-step: first concludes the rationale according to the case's facts and then determines the judgment. Recently, many models have been proposed and made tremendous progress in LJP, but most of them adopt an end-to-end manner that cannot be manually intervened by the judge for practical use. Moreover, existing models lack interpretability due to the neglect of rationale in the prediction process. Following the judge's real trial logic, in this paper, we propose a novel Rationale-based Legal Judgment Prediction (RLJP) framework. In the RLJP framework, the LJP process is split into two steps. In the first phase, the model generates the rationales according to the fact description. Then it predicts the judgment based on the fact and the generated rationales. Extensive experiments on a real-world dataset show RLJP achieves

the best results compared to the state-of-the-art models. Meanwhile, the proposed framework provides good interactivity and interpretability which enables practical use.

**TASA: Deceiving Question Answering Models by Twin Answer Sentences Attack**
*Yu Cao, Dianqi Li, Meng Fang, Tianyi Zhou, Jun Gao, Yibing Zhan and Dacheng Tao* 16:00-17:30 (Hall A, Room B)
We present Twin Answer Sentences Attack (TASA), an adversarial attack method for question answering (QA) models that produces fluent and grammatical adversarial contexts while maintaining gold answers. Despite phenomenal progress on general adversarial attacks, few works have investigated the vulnerability and attack specifically for QA models. In this work, we first explore the biases in the existing models and discover that they mainly rely on keyword matching between the question and context, and ignore the relevant contextual relations for answer prediction. Based on two biases above, TASA attacks the target model in two folds: (1) lowering the model's confidence on the gold answer with a perturbed answer sentence; (2) misguiding the model towards a wrong answer with a distracting answer sentence. Equipped with designed beam search and filtering methods, TASA can generate more effective attacks than existing textual attack methods while sustaining the quality of contexts, in extensive experiments on five QA datasets and human evaluations.

**Improving Temporal Generalization of Pre-trained Language Models with Lexical Semantic Change**
*Zhaochen Su, Zecheng Tang, xinyan guan, Lijun Wu, Min Zhang and Juntao Li* 16:00-17:30 (Hall A, Room B)
Recent research has revealed that neural language models at scale suffer from poor temporal generalization capability, i.e., language model pre-trained on static data from past years performs worse over time on emerging data. Existing methods mainly perform continual training to mitigate such a misalignment. While effective to some extent but is far from being addressed on both the language modeling and downstream tasks. In this paper, we empirically observe that temporal generalization is closely affiliated with lexical semantic change, which is one of the essential phenomena of natural languages. Based on this observation, we propose a simple yet effective lexical-level masking strategy to post-train a converged language model. Experiments on two pre-trained language models, two different classification tasks, and four benchmark datasets demonstrate the effectiveness of our proposed method over existing temporal adaptation methods, i.e., continual training with new data. Our code is available at https://github.com/zhaochen0110/LMLM.

**Exploring Mode Connectivity for Pre-trained Language Models**
*Yujia Qin, Cheng Qian, Jing Yi, Weize Chen, Yankai Lin, Xu Han, Zhiyuan Liu, Maosong Sun and Jie Zhou* 16:00-17:30 (Hall A, Room B)
Recent years have witnessed the prevalent application of pre-trained language models (PLMs) in NLP. From the perspective of parameter space, PLMs provide generic initialization, starting from which high-performance minima could be found. Although plenty of works have studied how to effectively and efficiently adapt PLMs to high-performance minima, little is known about the connection of various minima reached under different adaptation configurations. In this paper, we investigate the geometric connections of different minima through the lens of mode connectivity, which measures whether two minima can be connected with a low-loss path. We conduct empirical analyses to investigate three questions: (1) how could hyperparameters, specific tuning methods, and training data affect PLM's mode connectivity? (2) How does mode connectivity change during pre-training? (3) How does the PLM's task knowledge change along the path connecting two minima? In general, exploring the mode connectivity of PLMs conduces to understanding the geometric connection of different minima, which may help us fathom the inner workings of PLM downstream adaptation. The codes are publicly available at https://github.com/thunlp/Mode-Connectivity-PLM.

**Parameter-Efficient Tuning Makes a Good Classification Head**
*Zhuoyi Yang, Ming Ding, Yanhui Guo, Qingsong Lv and Jie Tang* 16:00-17:30 (Hall A, Room B)
In recent years, pretrained models revolutionized the paradigm of natural language understanding (NLU), where we append a randomly initialized classification head after the pretrained backbone, e.g. BERT, and finetune the whole model. As the pretrained backbone makes a major contribution to the improvement, we naturally expect a good pretrained classification head can also benefit the training. However, the final-layer output of the backbone, i.e. the input of the classification head, will change greatly during finetuning, making the usual head-only pretraining ineffective. In this paper, we find that parameter-efficient tuning makes a good classification head, with which we can simply replace the randomly initialized heads for a stable performance gain. Our experiments demonstrate that the classification head jointly pretrained with parameter-efficient tuning consistently improves the performance on 9 tasks in GLUE and SuperGLUE.

[INDUSTRY] **A Hybrid Approach to Cross-lingual Product Review Summarization**
*Saleh Soltan, Victor Soto, Ke Tran and Wael Hamza* 16:00-17:30 (Hall A, Room B)
We present a hybrid approach for product review summarization which consists of: (i) an unsupervised extractive step to extract the most important sentences out of all the reviews, and (ii) a supervised abstractive step to summarize the extracted sentences into a coherent short summary. This approach allows us to develop an efficient cross-lingual abstractive summarizer that can generate summaries in any language, given the extracted sentences out of thousands of reviews in a source language. In order to train and test the abstractive model, we create the Cross-lingual Amazon Reviews Summarization (CARS) dataset which provides English summaries for training, and English, French, Italian, Arabic, and Hindi summaries for testing based on selected English reviews. We show that the summaries generated by our model are as good as human written summaries in coherence, informativeness, non-redundancy, and fluency.

[INDUSTRY] **Knowledge Distillation based Contextual Relevance Matching for E-commerce Product Search**
*Ziyang Liu, Chaokun Wang, Hao Feng, Lingfei Wu and Liqun Yang* 16:00-17:30 (Hall A, Room B)
Online relevance matching is an essential task of e-commerce product search to boost the utility of search engines and ensure a smooth user experience. Previous work adopts either classical relevance matching models or Transformer-style models to address it. However, they ignore the inherent bipartite graph structures that are ubiquitous in e-commerce product search logs and are too inefficient to deploy online. In this paper, we design an efficient knowledge distillation framework for e-commerce relevance matching to integrate the respective advantages of Transformer-style models and classical relevance matching models. Especially for the core student model of the framework, we propose a novel method using k-order relevance modeling. The experimental results on large-scale real-world data (the size is 6 174 million) show that the proposed method significantly improves the prediction accuracy in terms of human relevance judgment. We deploy our method to JD.com online search platform. The A/B testing results show that our method significantly improves most business metrics under price sort mode and default sort mode.

[INDUSTRY] **Tackling Temporal Questions in Natural Language Interface to Databases**
*Ngoc Phuoc An Vo, Octavian Popescu, Irene L. Manotas and Vadim Sheinin* 16:00-17:30 (Hall A, Room B)
Temporal aspect is one of the most challenging areas in Natural Language Interface to Databases (NLIDB). This paper addresses and examines how temporal questions being studied and supported by the research community at both levels: popular annotated dataset (e.g. Spider) and recent advanced models. We present a new dataset with accompanied databases supporting temporal questions in NLIDB. We experiment with two SOTA models (Picard and ValueNet) to investigate how our new dataset helps these models learn and improve performance in temporal aspect.

[INDUSTRY] **Unsupervised Dense Retrieval for Scientific Articles**

*Dan Li, Vikrant Yadav, Zubair Afzal and George Tsatsaronis* 16:00-17:30 (Hall A, Room B)
In this work, we build a dense retrieval based semantic search engine on scientific articles from Elsevier. The major challenge is that there is no labeled data for training and testing. We apply a state-of-the-art unsupervised dense retrieval model called Generative Pseudo Labeling that generates high-quality pseudo training labels. Furthermore, since the articles are unbalanced across different domains, we select passages from multiple domains to form balanced training data. For the evaluation, we create two test sets: one manually annotated and one automatically created from the meta-information of our data. We compare the semantic search engine with the currently deployed lexical search engine on the two test sets. The results of the experiment show that the semantic search engine trained with pseudo training labels can significantly improve search performance.

[INDUSTRY] **Developing Prefix-Tuning Models for Hierarchical Text Classification**
*Lei Chen, Houwei Chou and Xiaodan Zhu* 16:00-17:30 (Hall A, Room B)
Hierarchical text classification (HTC) is a key problem and task in many industrial applications, which aims to predict labels organized in a hierarchy for given input text. For example, HTC can group the descriptions of online products into a taxonomy or organizing customer reviews into a hierarchy of categories. In real-life applications, while Pre-trained Language Models (PLMs) have dominated many NLP tasks, they face significant challenges too—the conventional fine-tuning process needs to modify and save models with a huge number of parameters. This is becoming more critical for HTC in both global and local modelling—the latter needs to learn multiple classifiers at different levels/nodes in a hierarchy. The concern will be even more serious since PLM sizes are continuing to increase in order to attain more competitive performances. Most recently, prefix tuning has become a very attractive technology by only tuning and saving a tiny set of parameters. Exploring prefix turning for HTC is hence highly desirable and has timely impact. In this paper, we investigate prefix tuning on HTC in two typical setups: local and global HTC. Our experiment shows that the prefix-tuning model only needs less than 1% of parameters and can achieve performance comparable to regular full fine-tuning. We demonstrate that using contrastive learning in learning prefix vectors can further improve HTC performance.

[INDUSTRY] **Full-Stack Information Extraction System for Cybersecurity Intelligence**
*Youngja Park and Taesung Lee* 16:00-17:30 (Hall A, Room B)
Due to rapidly growing cyber-attacks and security vulnerabilities, many reports on cyber-threat intelligence (CTI) are being published daily. While these reports can help security analysts to understand on-going cyber threats, the overwhelming amount of information makes it difficult to digest the information in a timely manner. This paper presents, SecIE, an industrial-strength full-stack information extraction (IE) system for the security domain. SecIE can extract a large number of security entities, relations and the temporal information of the relations, which is critical for cyberthreat investigations. Our evaluation with 133 labeled threat reports containing 108,021 tokens shows that SecIE achieves over 92% F1-score for entity extraction and about 70% F1-score for relation extraction. We also showcase how SecIE can be used for downstream security applications.

[INDUSTRY] **Revisiting and Advancing Chinese Natural Language Understanding with Accelerated Heterogeneous Knowledge Pre-training**
*Taolin Zhang, junwei dong, Jianing Wang, Chengyu Wang, Ang Wang, Yinghui Liu, jun huang, Yong Li and XIAOFENG HE* 16:00-17:30 (Hall A, Room B)
Recently, knowledge-enhanced pre-trained language models (KEPLMs) improve context-aware representations via learning from structured relations in knowledge bases, and/or linguistic knowledge from syntactic or dependency analysis. Unlike English, there is a lack of high-performing open-source Chinese KEPLMs in the natural language processing (NLP) community to support various language understanding applications. In this paper, we revisit and advance the development of Chinese natural language understanding with a series of novel Chinese KEPLMs released in various parameter sizes, namely CKBERT (Chinese knowledge-enhanced BERT). Specifically, both relational and linguistic knowledge is effectively injected into CKBERT based on two novel pre-training tasks, i.e., linguistic-aware masked language modeling and contrastive multi-hop relation modeling. Based on the above two pre-training paradigms and our in-house implemented TorchAccelerator, we have pre-trained base (110M), large (345M) and huge (1.3B) versions of CKBERT efficiently on GPU clusters. Experiments demonstrate that CKBERT consistently outperforms strong baselines for Chinese over various benchmark NLP tasks and in terms of different model sizes.

[INDUSTRY] **PILE: Pairwise Iterative Logits Ensemble for Multi-Teacher Labeled Distillation**
*Lianshang Cai, Linhao Zhang, Dehong Ma, Jun Fan, Daiting Shi, Yi Wu, Zhicong Cheng, Simiu Gu and Dawei Yin* 16:00-17:30 (Hall A, Room B)
Pre-trained language models have become a crucial part of ranking systems and achieved very impressive effects recently. To maintain high performance while keeping efficient computations, knowledge distillation is widely used. In this paper, we focus on two key questions in knowledge distillation for ranking models: 1) how to ensemble knowledge from multi-teacher; 2) how to utilize the label information of data in the distillation process. We propose a unified algorithm called Pairwise Iterative Logits Ensemble (PILE) to tackle these two questions simultaneously. PILE ensembles multi-teacher logits supervised by label information in an iterative way and achieved competitive performance in both offline and online experiments. The proposed method has been deployed in a real-world commercial search system.

## Virtual Portal 3

16:00-17:30 (Hall A, Room C)

---

**A Localized Geometric Method to Match Knowledge in Low-dimensional Hyperbolic Space**
*Bo Hui, Tian Xia and Wei-Shinn Ku* 16:00-17:30 (Hall A, Room C)
Matching equivalent entities across Knowledge graphs is a pivotal step for knowledge fusion. Previous approaches usually study the problem in Euclidean space. However, recent works have shown that hyperbolic space has a higher capacity than Euclidean space and hyperbolic embedding can represent the hierarchical structure in a knowledge graph. In this paper, we propose a localized geometric method to find equivalent entities in hyperbolic space. Specifically, we use a hyperbolic neural network to encode the lingual information of entities and the structure of both knowledge graphs into a low-dimensional hyperbolic space. To address the asymmetry of structure on different KGs and the localized nature of relations, we learn an instance-specific geometric mapping function based on rotation to match entity pairs. A contrastive loss function is used to train the model. The experiment verifies the power of low-dimensional hyperbolic space for entity matching and shows that our method outperforms the state of the art by a large margin.

**Making Pretrained Language Models Good Long-tailed Learners**
*Chen Zhang, Lei Ren, Jingang Wang, Wei Wu and Dawei Song* 16:00-17:30 (Hall A, Room C)
Prompt-tuning has shown appealing performance in few-shot classification by virtue of its capability in effectively exploiting pre-trained knowledge. This motivates us to check the hypothesis that prompt-tuning is also a promising choice for long-tailed classification, since the

tail classes are intuitively few-shot ones. To achieve this aim, we conduct empirical studies to examine the hypothesis. The results demonstrate that prompt-tuning makes pretrained language models at least good long-tailed learners. For intuitions on why prompt-tuning can achieve good performance in long-tailed classification, we carry out in-depth analyses by progressively bridging the gap between prompt-tuning and commonly used finetuning. The summary is that the classifier structure and parameterization form the key to making good long-tailed learners, in comparison with the less important input structure. Finally, we verify the applicability of our finding to few-shot classification.

**HPT: Hierarchy-aware Prompt Tuning for Hierarchical Text Classification**
*Zihan Wang, Peiyi Wang, Tianyu Liu, Binghuai Lin, Yunbo Cao, Zhifang Sui and Houfeng Wang* 16:00-17:30 (Hall A, Room C)
Hierarchical text classification (HTC) is a challenging subtask of multi-label classification due to its complex label hierarchy. Recently, the pretrained language models (PLM) have been widely adopted in HTC through a fine-tuning paradigm. However, in this paradigm, there exists a huge gap between the classification tasks with sophisticated label hierarchy and the masked language model (MLM) pretraining tasks of PLMs and thus the potential of PLMs cannot be fully tapped. To bridge the gap, in this paper, we propose HPT, a Hierarchy-aware Prompt Tuning method to handle HTC from a multi-label MLM perspective. Specifically, we construct a dynamic virtual template and label words that take the form of soft prompts to fuse the label hierarchy knowledge and introduce a zero-bounded multi-label cross-entropy loss to harmonize the objectives of HTC and MLM. Extensive experiments show HPT achieves state-of-the-art performances on 3 popular HTC datasets and is adept at handling the imbalance and low resource situations. Our code is available at https://github.com/wzh9969/HPT.

**GA-SAM: Gradient-Strength based Adaptive Sharpness-Aware Minimization for Improved Generalization**
*Zhiyuan Zhang, Ruixuan Luo, Qi Su and Xu Sun* 16:00-17:30 (Hall A, Room C)
Recently, Sharpness-Aware Minimization (SAM) algorithm has shown state-of-the-art generalization abilities in vision tasks. It demonstrates that flat minima tend to imply better generalization abilities. However, it has some difficulty implying SAM to some natural language tasks, especially to models with drastic gradient changes, such as RNNs. In this work, we analyze the relation between the flatness of the local minimum and its generalization ability from a novel and straightforward theoretical perspective. We propose that the shift of the training and test distributions can be equivalently seen as a virtual parameter corruption or perturbation, which can explain why flat minima that are robust against parameter corruptions or perturbations have better generalization performances. On its basis, we propose a Gradient-Strength based Adaptive Sharpness-Aware Minimization (GA-SAM) algorithm to help to learn algorithms find flat minima that generalize better. Results in various language benchmarks validate the effectiveness of the proposed GA-SAM algorithm on natural language tasks.

**Active Example Selection for In-Context Learning**
*Yiming Zhang, Shi Feng and Chenhao Tan* 16:00-17:30 (Hall A, Room C)
With a handful of demonstration examples, large-scale language models demonstrate strong capability to perform various tasks by in-context learning from these examples, without any fine-tuning. We demonstrate that in-context learning performance can be highly unstable across samples of examples, indicating the idiosyncrasies of how language models acquire information. We formulate example selection for in-context learning as a sequential decision problem, and propose a reinforcement learning algorithm for identifying generalizable policies to select demonstration examples. For GPT-2, our learned policies demonstrate strong abilities of generalizing to unseen tasks in training, with a 5.8% improvement on average. Examples selected from our learned policies can even achieve a small improvement on GPT-3 Ada. However, the improvement diminishes on larger GPT-3 models, suggesting emerging capabilities of large language models.

**BBTv2: Towards a Gradient-Free Future with Large Language Models**
*Tianxiang Sun, Zhengfu He, Hong Qian, Yunhua Zhou, Xuanjing Huang and Xipeng Qiu* 16:00-17:30 (Hall A, Room C)
Most downstream adaptation methods tune all or part of the parameters of pre-trained models (PTMs) through gradient descent, where the tuning cost increases linearly with the growth of the model size. By contrast, gradient-free methods only require the forward computation of the PTM to tune the prompt, retaining the benefits of efficient tuning and deployment. Though, past work on gradient-free tuning often introduces gradient descent to seek a good initialization of prompt and lacks versatility across tasks and PTMs. In this paper, we present BBTv2, an improved version of Black-Box Tuning, to drive PTMs for few-shot learning. We prepend continuous prompts to every layer of the PTM and propose a divide-and-conquer gradient-free algorithm to optimize the prompts at different layers alternately. Extensive experiments across various tasks and PTMs show that BBTv2 can achieve comparable performance to full model tuning and state-of-the-art parameter-efficient methods (e.g., Adapter, LoRA, BitFit, etc.) under few-shot settings while maintaining much fewer tunable parameters.

**G-MAP: General Memory-Augmented Pre-trained Language Model for Domain Tasks**
*Zhongwei Wan, Yichun Yin, Wei Zhang, Jiaxin Shi, Lifeng Shang, Guangyong Chen, Xin Jiang and Qun Liu* 16:00-17:30 (Hall A, Room C)
General pre-trained language models (PLMs), such as BERT, have achieved remarkable performance on various NLP tasks. Recently, domain-specific PLMs have been proposed to boost the task performance of specific domains (e.g., biomedical and computer science) by continuing to pre-train general PLMs with domain-specific corpora. However, this domain-adaptive pre-training (DAPT [**DBLP:conf/acl/GururanganMSLBD20**]) tends to forget the previous general knowledge acquired by general PLMs, which leads to a *catastrophic forgetting* phenomenon and sub-optimal performance. To alleviate this problem, we propose a new framework of **M**emory-**A**ugmented **P**re-trained Language Model (**MAP**), which augments the domain-specific PLM by a memory built from the frozen general PLM without losing the general knowledge. Specifically, we propose a new memory-augmented layer, and based on it, different augmentation strategies are explored to build memory and fusion memory into domain-specific PLM. We demonstrate the effectiveness of MAP on different domains (biomedical and computer science publications, news, and reviews) and different kinds (text classification, QA, NER) of tasks, and the extensive results show that the proposed MAP can achieve SOTA results on these tasks.

**Textual Manifold-based Defense Against Natural Language Adversarial Examples**
*Dang Nguyen Minh and Anh Tuan Luu* 16:00-17:30 (Hall A, Room C)
Despite the recent success of large pretrained language models in NLP, they are susceptible to adversarial examples. Concurrently, several studies on adversarial images have observed an intriguing property: the adversarial images tend to leave the low-dimensional natural data manifold. In this study, we find a similar phenomenon occurs in the contextualized embedding space of natural sentences induced by pre-trained language models in which textual adversarial examples tend to have their embeddings diverge off the manifold of natural sentence embeddings. Based on this finding, we propose Textual Manifold-based Defense (TMD), a defense mechanism that learns the embedding space manifold of the underlying language model and projects novel inputs back to the approximated structure before classification. Through extensive experiments, we find that our method consistently and significantly outperforms previous defenses under various attack settings while remaining unaffected to the clean accuracy. To the best of our knowledge, this is the first kind of manifold-based defense adapted to the NLP domain.

[CL] **Enhancing Lifelong Language Learning by Improving Pseudo-Sample Generation**
*Kasidis Kanwatchara, Thanapapas Horsuwan, Piyawat Lertvittayakumjorn, Boonserm Kijsirikul and Peerapon Vateekul* 16:00-17:30 (Hall A, Room C)
To achieve lifelong language learning, pseudo-rehearsal methods leverage samples generated from a language model to refresh the knowledge of previously learned tasks. Without proper controls, however, these methods could fail to retain the knowledge of complex tasks with longer

texts since most of the generated samples are low in quality. To overcome the problem, we propose three specific contributions. First, we utilize double language models, each of which specializes on a specific part of the input, to produce high-quality pseudo samples. Second, we reduce the number of parameters used by applying adapter modules to enhance training efficiency. Third, we further improve the overall quality of pseudo samples using temporal ensembling and sample regeneration. The results show that our framework achieves significant improvement over baselines on multiple task sequences. Also, our pseudo sample analysis reveals helpful insights for designing even better pseudo-rehearsal methods in the future.

**Neural Machine Translation with Contrastive Translation Memories**
*Xin Cheng, Shen Gao, Lemao Liu, Dongyan Zhao and Rui Yan* 16:00-17:30 (Hall A, Room C)
Retrieval-augmented Neural Machine Translation models have been successful in many translation scenarios. Different from previous works that make use of mutually similar but redundant translation memories (TMs), we propose a new retrieval-augmented NMT to model contrastively retrieved translation memories that are holistically similar to the source sentence while individually contrastive to each other providing maximal information gain in three phases. First, in TM retrieval phase, we adopt contrastive retrieval algorithm to avoid redundancy and uninformativeness of similar translation pieces. Second, in memory encoding stage, given a set of TMs we propose a novel Hierarchical Group Attention module to gather both local context of each TM and global context of the whole TM set. Finally, in training phase, a Multi-TM contrastive learning objective is introduced to learn salient feature of each TM with respect to target sentence. Experimental results show that our framework obtains substantial improvements over strong baselines in the benchmark dataset.

**A Template-based Method for Constrained Neural Machine Translation**
*Shuo Wang, Peng Li, Zhixing Tan, Zhaopeng Tu, Maosong Sun and Yang Liu* 16:00-17:30 (Hall A, Room C)
Machine translation systems are expected to cope with various types of constraints in many practical scenarios. While neural machine translation (NMT) has achieved strong performance in unconstrained cases, it is non-trivial to impose pre-specified constraints into the translation process of NMT models. Although many approaches have been proposed to address this issue, most existing methods can not satisfy the following three desiderata at the same time: (1) high translation quality, (2) high match accuracy, and (3) low latency. In this work, we propose a template-based method that can yield results with high translation quality and match accuracy and the inference speed of our method is comparable with unconstrained NMT models. Our basic idea is to rearrange the generation of constrained and unconstrained tokens through a template. Our method does not require any changes in the model architecture and the decoding algorithm. Experimental results show that the proposed template-based approach can outperform several representative baselines in both lexically and structurally constrained translation tasks.

**Competency-Aware Neural Machine Translation: Can Machine Translation Know its Own Translation Quality?**
*Pei Zhang, Baosong Yang, Hao-Ran Wei, Dayiheng Liu, Kai Fan, Luo Si and Jun Xie* 16:00-17:30 (Hall A, Room C)
Neural machine translation (NMT) is often criticized for failures that happen without awareness. The lack of competency awareness makes NMT untrustworthy. This is in sharp contrast to human translators who give feedback or conduct further investigations whenever they are in doubt about predictions. To fill this gap, we propose a novel competency-aware NMT by extending conventional NMT with a self-estimator, offering abilities to translate a source sentence and estimate its competency. The self-estimator encodes the information of the decoding procedure and then examines whether it can reconstruct the original semantics of the source sentence. Experimental results on four translation tasks demonstrate that the proposed method not only carries out translation tasks intact but also delivers outstanding performance on quality estimation. Without depending on any reference or annotated data typically required by state-of-the-art metric and quality estimation methods, our model yields an even higher correlation with human quality judgments than a variety of aforementioned methods, such as BLEURT, COMET, and BERTScore. Quantitative and qualitative analyses show better robustness of competency awareness in our model.[1]

**Multi-Granularity Optimization for Non-Autoregressive Translation**
*Yafu Li, Leyang Cui, Yongjing Yin and Yue Zhang* 16:00-17:30 (Hall A, Room C)
Despite low latency, non-autoregressive machine translation (NAT) suffers severe performance deterioration due to the naive independence assumption. This assumption is further strengthened by cross-entropy loss, which encourages a strict match between the hypothesis and the reference token by token. To alleviate this issue, we propose multi-granularity optimization for NAT, which collects model behaviours on translation segments of various granularities and integrates feedback for backpropagation. Experiments on four WMT benchmarks show that the proposed method significantly outperforms the baseline models trained with cross-entropy loss, and achieves the best performance on WMT'16 En-Ro and highly competitive results on WMT'14 En-De for fully non-autoregressive translation.

**Improving Machine Translation with Phrase Pair Injection and Corpus Filtering**
*Akshay Batheja and Pushpak Bhattacharyya* 16:00-17:30 (Hall A, Room C)
In this paper, we show that the combination of Phrase Pair Injection and Corpus Filtering boosts the performance of Neural Machine Translation (NMT) systems. We extract parallel phrases and sentences from the pseudo-parallel corpus and augment it with the parallel corpus to train the NMT models. With the proposed approach, we observe an improvement in the Machine Translation (MT) system for 3 low-resource language pairs, Hindi-Marathi, English-Marathi, and English-Pashto, and 6 translation directions by up to 2.7 BLEU points, on the FLORES test data. These BLEU score improvements are over the models trained using the whole pseudo-parallel corpus augmented with the parallel corpus.

**XLM-D: Decorate Cross-lingual Pre-training Model as Non-Autoregressive Neural Machine Translation**
*Yong Wang, Shilin He, Guanhua Chen, Yun Chen and Daxin Jiang* 16:00-17:30 (Hall A, Room C)
Pre-training language models have achieved thriving success in numerous natural language understanding and autoregressive generation tasks, but non-autoregressive generation in applications such as machine translation has not sufficiently benefited from the pre-training paradigm. In this work, we establish the connection between a pre-trained masked language model (MLM) and non-autoregressive generation on machine translation. From this perspective, we present XLM-D, which seamlessly transforms an off-the-shelf cross-lingual pre-training model into a non-autoregressive translation (NAT) model with a lightweight yet effective decorator. Specifically, the decorator ensures the representation consistency of the pre-trained model and brings only one additional trainable parameter. Extensive experiments on typical translation datasets show that our models obtain state-of-the-art performance while realizing the inference speed-up by 19.9x. One striking result is that on WMT14 En-De, our XLM-D obtains 29.80 BLEU points with multiple iterations, which outperforms the previous mask-predict model by 2.77 points.

**ConsistTL: Modeling Consistency in Transfer Learning for Low-Resource Neural Machine Translation**
*Zhaocong Li, Xuebo Liu, Derek F. Wong, Lidia S. Chao and Min Zhang* 16:00-17:30 (Hall A, Room C)
Transfer learning is a simple and powerful method that can be used to boost model performance of low-resource neural machine translation (NMT). Existing transfer learning methods for NMT are static, which simply transfer knowledge from a parent model to a child model once

[1] Code and test sets are available at: https://github.com/xiaoyi0814/CANMT.

via parameter initialization. In this paper, we propose a novel transfer learning method for NMT, namely ConsistTL, which can continuously transfer knowledge from the parent model during the training of the child model. Specifically, for each training instance of the child model, ConsistTL constructs the semantically-equivalent instance for the parent model and encourages prediction consistency between the parent and child for this instance, which is equivalent to the child model learning each instance under the guidance of the parent model. Experimental results on five low-resource NMT tasks demonstrate that ConsistTL results in significant improvements over strong transfer learning baselines, with a gain up to 1.7 BLEU over the existing back-translation model on the widely-used WMT17 Turkish-English benchmark. Further analysis reveals that ConsistTL can improve the inference calibration of the child model. Code and scripts are freely available at https://github.com/NLP2CT/ConsistTL.

**RAPO: An Adaptive Ranking Paradigm for Bilingual Lexicon Induction**
*Zhoujin Tian, Chaozhuo Li, Shuo Ren, Zhiqiang Zuo, Zengxuan Wen, Xinyue Hu, Xiao Han, Haizhen Huang, Denvy Deng, Qi Zhang and Xing Xie* 16:00-17:30 (Hall A, Room C)
Bilingual lexicon induction induces the word translations by aligning independently trained word embeddings in two languages. Existing approaches generally focus on minimizing the distances between words in the aligned pairs, while suffering from low discriminative capability to distinguish the relative orders between positive and negative candidates. In addition, the mapping function is globally shared by all words, whose performance might be hindered by the deviations in the distributions of different languages. In this work, we propose a novel ranking-oriented induction model RAPO to learn personalized mapping function for each word. RAPO is capable of enjoying the merits from the unique characteristics of a single word and the cross-language isomorphism simultaneously. Extensive experimental results on public datasets including both rich-resource and low-resource languages demonstrate the superiority of our proposal. Our code is publicly available in `https://github.com/Jlfj345wf/RAPO`.

**Entropy-Based Vocabulary Substitution for Incremental Learning in Multilingual Neural Machine Translation**
*Kaiyu Huang, Peng Li, Jin Ma and Yang Liu* 16:00-17:30 (Hall A, Room C)
In a practical real-world scenario, the longstanding goal is that a universal multilingual translation model can be incrementally updated when new language pairs arrive. Specifically, the initial vocabulary only covers some of the words in new languages, which hurts the translation quality for incremental learning. Although existing approaches attempt to address this issue by replacing the original vocabulary with a rebuilt vocabulary or constructing independent language-specific vocabularies, these methods can not meet the following three demands simultaneously: (1) High translation quality for original and incremental languages, (2) low cost for model training, (3) low time overhead for preprocessing. In this work, we propose an entropy-based vocabulary substitution (EVS) method that just needs to walk through new language pairs for incremental learning in a large-scale multilingual data updating while remaining the size of the vocabulary. Our method has access to learn new knowledge from updated training samples incrementally while keeping high translation quality for original language pairs, alleviating the issue of catastrophic forgetting. Results of experiments show that EVS can achieve better performance and save excess overhead for incremental learning in the multilingual machine translation task.

**Digging Errors in NMT: Evaluating and Understanding Model Errors from Partial Hypothesis Space**
*Jianhao Yan, Chenming Wu, Fandong Meng and Jie Zhou* 16:00-17:30 (Hall A, Room C)
Solid evaluation of neural machine translation (NMT) is key to its understanding and improvement. Current evaluation of an NMT system is usually built upon a heuristic decoding algorithm (e.g., beam search) and an evaluation metric assessing similarity between the translation and golden reference. However, this system-level evaluation framework is limited by evaluating only one best hypothesis and search errors brought by heuristic decoding algorithms. To better understand NMT models, we propose a novel evaluation protocol, which defines model errors with model's ranking capability over hypothesis space. To tackle the problem of exponentially large space, we propose two approximation methods, top region evaluation along with an exact top-k decoding algorithm, which finds top-ranked hypotheses in the whole hypothesis space, and Monte Carlo sampling evaluation, which simulates hypothesis space from a broader perspective. To quantify errors, we define our NMT model errors by measuring distance between the hypothesis array ranked by the model and the ideally ranked hypothesis array. After confirming the strong correlation with human judgment, we apply our evaluation to various NMT benchmarks and model architectures. We show that the state-of-the-art Transformer models face serious ranking issues and only perform at the random chance level in the top region. We further analyze model errors on architectures with different depths and widths, as well as different data-augmentation techniques, showing how these factors affect model errors. Finally, we connect model errors with the search algorithms and provide interesting findings of beam search inductive bias and correlation with Minimum Bayes Risk (MBR) decoding.

## Virtual Portal 4

16:00-17:30 (Hall A, Room D)

---

**Cross-Align: Modeling Deep Cross-lingual Interactions for Word Alignment**
*Siyu Lai, Zhen Yang, Fandong Meng, Yufeng Chen, Jinan Xu and Jie Zhou* 16:00-17:30 (Hall A, Room D)
Word alignment which aims to extract lexicon translation equivalents between source and target sentences, serves as a fundamental tool for natural language processing. Recent studies in this area have yielded substantial improvements by generating alignments from contextualized embeddings of the pre-trained multilingual language models. However, we find that the existing approaches capture few interactions between the input sentence pairs, which degrades the word alignment quality severely, especially for the ambiguous words in the monolingual context. To remedy this problem, we propose Cross-Align to model deep interactions between the input sentence pairs, in which the source and target sentences are encoded separately with the shared self-attention modules in the shallow layers, while cross-lingual interactions are explicitly constructed by the cross-attention modules in the upper layers. Besides, to train our model effectively, we propose a two-stage training framework, where the model is trained with a simple Translation Language Modeling (TLM) objective in the first stage and then finetuned with a self-supervised alignment objective in the second stage. Experiments show that the proposed Cross-Align achieves the state-of-the-art (SOTA) performance on four out of five language pairs.

**Discovering Low-rank Subspaces for Language-agnostic Multilingual Representations**
*Zhihui Xie, Handong Zhao, Tong Yu and Shuai Li* 16:00-17:30 (Hall A, Room D)
Large pretrained multilingual language models (ML-LMs) have shown remarkable capabilities of zero-shot cross-lingual transfer, without direct cross-lingual supervision. While these results are promising, follow-up works found that, within the multilingual embedding spaces, there exists strong language identity information which hinders the expression of linguistic factors shared across languages. For semantic tasks like cross-lingual sentence retrieval, it is desired to remove such language identity signals to fully leverage semantic information. In this work, we provide a novel view of projecting away language-specific factors from a multilingual embedding space. Specifically, we discover that there exists a low-rank subspace that primarily encodes information irrelevant to semantics (e.g., syntactic information). To identify this subspace, we present a simple but effective unsupervised method based on singular value decomposition with multiple monolingual corpora

as input. Once the subspace is found, we can directly project the original embeddings into the null space to boost language agnosticism without finetuning. We systematically evaluate our method on various tasks including the challenging language-agnostic QA retrieval task. Empirical results show that applying our method consistently leads to improvements over commonly used ML-LMs.

**Intriguing Properties of Compression on Multilingual Models**
*Kelechi Ogueji, Orevaoghene Ahia, Gbemileke Onilude, Sebastian Gehrmann, Sara Hooker and Julia Kreutzer* 16:00-17:30 (Hall A, Room D)
Multilingual models are often particularly dependent on scaling to generalize to a growing number of languages. Compression techniques are widely relied upon to reconcile the growth in model size with real world resource constraints, but compression can have a disparate effect on model performance for low-resource languages. It is thus crucial to understand the trade-offs between scale, multilingualism, and compression. In this work, we propose an experimental framework to characterize the impact of sparsifying multilingual pre-trained language models during fine-tuning. Applying this framework to mBERT named entity recognition models across 40 languages, we find that compression confers several intriguing and previously unknown generalization properties. In contrast to prior findings, we find that compression may improve model robustness over dense models. We additionally observe that under certain sparsification regimes compression may aid, rather than disproportionately impact the performance of low-resource languages.

**English Contrastive Learning Can Learn Universal Cross-lingual Sentence Embeddings**
*Yaushian Wang, Ashley Wu and Graham Neubig* 16:00-17:30 (Hall A, Room D)
Universal cross-lingual sentence embeddings map semantically similar cross-lingual sentences into a shared embedding space. Aligning cross-lingual sentence embeddings usually requires supervised cross-lingual parallel sentences. In this work, we propose mSimCSE, which extends SimCSE to multilingual settings and reveal that contrastive learning on English data can surprisingly learn high-quality universal cross-lingual sentence embeddings without any parallel data. In unsupervised and weakly supervised settings, mSimCSE significantly improves previous sentence embedding methods on cross-lingual retrieval and multilingual STS tasks. The performance of unsupervised mSimCSE is comparable to fully supervised methods in retrieving low-resource languages and multilingual STS. The performance can be further enhanced when cross-lingual NLI data is available.

**PromptEHR: Conditional Electronic Healthcare Records Generation with Prompt Learning**
*Zifeng Wang and Jimeng Sun* 16:00-17:30 (Hall A, Room D)
Accessing longitudinal multimodal Electronic Healthcare Records (EHRs) is challenging due to privacy concerns, which hinders the use of ML for healthcare applications. Synthetic EHRs generation bypasses the need to share sensitive real patient records. However, existing methods generate single-modal EHRs by unconditional generation or by longitudinal inference, which falls short of low flexibility and makes unrealistic EHRs. In this work, we propose to formulate EHRs generation as a text-to-text translation task by language models (LMs), which suffices to highly flexible event imputation during generation. We also design prompt learning to control the generation conditioned by numerical and categorical demographic features. We evaluate synthetic EHRs quality by two perplexity measures accounting for their longitudinal pattern (longitudinal imputation perplexity, lpl) and the connections cross modalities (cross-modality imputation perplexity, mpl). Moreover, we utilize two adversaries: membership and attribute inference attacks for privacy-preserving evaluation. Experiments on MIMIC-III data demonstrate the superiority of our methods on realistic EHRs generation (53.1

**Rethinking Positional Encoding in Tree Transformer for Code Representation**
*Han Peng, Ge Li, Yunfei Zhao and Zhi Jin* 16:00-17:30 (Hall A, Room D)
Transformers are now widely used in code representation, and several recent works further develop tree Transformers to capture the syntactic structure in source code. Specifically, novel tree positional encodings have been proposed to incorporate inductive bias into Transformer. In this work, we propose a novel tree Transformer encoding node positions based on our new description method for tree structures. Technically, local and global soft bias shown in previous works is both introduced as positional encodings of our Transformer model. Our model finally outperforms strong baselines on code summarization and completion tasks across two languages, demonstrating our model's effectiveness. Besides, extensive experiments and ablation study shows that combining both local and global paradigms is still helpful in improving model performance. We release our code at `https://github.com/AwdHanPeng/TreeTransformer`.

**Chapter Ordering in Novels**
*Allen Kim and Steve Skiena* 16:00-17:30 (Hall A, Room D)
Understanding narrative flow and text coherence in long-form documents (novels) remains an open problem in NLP. To gain insight, we explore the task of chapter ordering, reconstructing the original order of chapters in novel given a random permutation of the text. This can be seen as extending the well-known sentence ordering task to vastly larger documents: our task deals with over 9,000 novels with an average of twenty chapters each, versus standard sentence ordering datasets averaging only 5-8 sentences. We formulate the task of reconstructing order as a constraint solving problem, using minimum feedback arc set and traveling salesman problem optimization criteria, where the weights of the graph are generated based on models for character occurrences and chapter boundary detection, using relational chapter scores derived from RoBERTa. Our best methods yield a Spearman correlation of 0.59 on this novel and challenging task, substantially above baseline.

**Open-ended Knowledge Tracing for Computer Science Education**
*Naiming Liu, Zichao Wang, Richard Baraniuk and Andrew Lan* 16:00-17:30 (Hall A, Room D)
In educational applications, knowledge tracing refers to the problem of estimating students' time-varying concept/skill mastery level from their past responses to questions and predicting their future performance. One key limitation of most existing knowledge tracing methods is that they treat student responses to questions as binary-valued, i.e., whether they are correct or incorrect. Response correctness analysis/prediction is straightforward, but it ignores important information regarding mastery, especially for open-ended questions. In contrast, exact student responses can provide much more information. In this paper, we conduct the first exploration int open-ended knowledge tracing (OKT) by studying the new task of predicting students' exact open-ended responses to questions. Our work is grounded in the domain of computer science education with programming questions. We develop an initial solution to the OKT problem, a student knowledge-guided code generation approach, that combines program synthesis methods using language models with student knowledge tracing methods. We also conduct a series of quantitative and qualitative experiments on a real-world student code dataset to validate and demonstrate the promise of OKT.

**SEEN: Structured Event Enhancement Network for Explainable Need Detection of Information Recall Assistance**
*You-En Lin, An-Zi Yen, Hen-Hsen Huang and Hsin-Hsi Chen* 16:00-17:30 (Hall A, Room D)
When recalling life experiences, people often forget or confuse life events, which necessitates information recall services. Previous work on information recall focuses on providing such assistance reactively, i.e., by retrieving the life event of a given query. Proactively detecting the need for information recall services is rarely discussed. In this paper, we use a human-annotated life experience retelling dataset to detect the right time to trigger the information recall service. We propose a pilot model—structured event enhancement network (SEEN) that detects life event inconsistency, additional information in life events, and forgotten events. A fusing mechanism is also proposed to incorporate event graphs of stories and enhance the textual representations. To explain the need detection results, SEEN simultaneously provides support evi-

dence by selecting the related nodes from the event graph. Experimental results show that SEEN achieves promising performance in detecting information needs. In addition, the extracted evidence can be served as complementary information to remind users what events they may want to recall.

**Tiny-NewsRec: Effective and Efficient PLM-based News Recommendation**
*Yang Yu, Fangzhao Wu, Chuhan Wu, Jingwei Yi and Qi Liu* 16:00-17:30 (Hall A, Room D)
News recommendation is a widely adopted technique to provide personalized news feeds for the user. Recently, pre-trained language models (PLMs) have demonstrated the great capability of natural language understanding and benefited news recommendation via improving news modeling. However, most existing works simply finetune the PLM with the news recommendation task, which may suffer from the known domain shift problem between the pre-training corpus and downstream news texts. Moreover, PLMs usually contain a large volume of parameters and have high computational overhead, which imposes a great burden on low-latency online services. In this paper, we propose Tiny-NewsRec, which can improve both the effectiveness and the efficiency of PLM-based news recommendation. We first design a self-supervised domain-specific post-training method to better adapt the general PLM to the news domain with a contrastive matching task between news titles and news bodies. We further propose a two-stage knowledge distillation method to improve the efficiency of the large PLM-based news recommendation model while maintaining its performance. Multiple teacher models originated from different time steps of our post-training procedure are used to transfer comprehensive knowledge to the student model in both its post-training stage and finetuning stage. Extensive experiments on two real-world datasets validate the effectiveness and efficiency of our method.

**Boundary-Driven Table-Filling for Aspect Sentiment Triplet Extraction**
*Yice Zhang, Yifan Yang, Yihui Li, Bin Liang, Shiwei Chen, Yixue Dang, Min Yang and Ruifeng Xu* 16:00-17:30 (Hall A, Room D)
Aspect Sentiment Triplet Extraction (ASTE) aims to extract the aspect terms along with the corresponding opinion terms and the expressed sentiments in the review, which is an important task in sentiment analysis. Previous research efforts generally address the ASTE task in an end-to-end fashion through the table-filling formalization, in which the triplets are represented by a two-dimensional (2D) table of word-pair relations. Under this formalization, a term-level relation is decomposed into multiple independent word-level relations, which leads to relation inconsistency and boundary insensitivity in the face of multi-word aspect terms and opinion terms. To overcome these issues, we propose Boundary-Driven Table-Filling (BDTF), which represents each triplet as a relation region in the 2D table and transforms the ASTE task into detection and classification of relation regions. We also notice that the quality of the table representation greatly affects the performance of BDTF. Therefore, we develop an effective relation representation learning approach to learn the table representation, which can fully exploit both word-to-word interactions and relation-to-relation interactions. Experiments on several public benchmarks show that the proposed approach achieves state-of-the-art performances.

**Cross-lingual neural fuzzy matching for exploiting target-language monolingual corpora in computer-aided translation**
*Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz and Felipe Sánchez-Martínez* 16:00-17:30 (Hall A, Room D)
Computer-aided translation (CAT) tools based on translation memories (MT) play a prominent role in the translation workflow of professional translators. However, the reduced availability of in-domain TMs, as compared to in-domain monolingual corpora, limits its adoption for a number of translation tasks. In this paper, we introduce a novel neural approach aimed at overcoming this limitation by exploiting not only TMs, but also in-domain target-language (TL) monolingual corpora, and still enabling a similar functionality to that offered by conventional TM-based CAT tools. Our approach relies on cross-lingual sentence embeddings to retrieve translation proposals from TL monolingual corpora, and on a neural model to estimate their post-editing effort. The paper presents an automatic evaluation of these techniques on four language pairs that shows that our approach can successfully exploit monolingual texts in a TM-based CAT environment, increasing the amount of useful translation proposals, and that our neural model for estimating the post-editing effort enables the combination of translation proposals obtained from monolingual corpora and from TMs in the usual way. A human evaluation performed on a single language pair confirms the results of the automatic evaluation and seems to indicate that the translation proposals retrieved with our approach are more useful than what the automatic evaluation shows.

**Improved grammatical error correction by ranking elementary edits**
*Alexey Sorokin* 16:00-17:30 (Hall A, Room D)
We offer a two-stage reranking method for grammatical error correction: the first model serves as edit generator, while the second classifies the proposed edits as correct or false. We show how to use both encoder-decoder and sequence labeling models for the first step of our pipeline. We achieve state-of-the-art quality on BEA 2019 English dataset even using weak BERT-GEC edit generator. Combining our roberta-base scorer with state-of-the-art GECToR edit generator, we surpass GECToR by 2-3

**Keyphrase Generation via Soft and Hard Semantic Corrections**
*Guangzhen Zhao, Guoshun Yin, Peng Yang and Yu Yao* 16:00-17:30 (Hall A, Room D)
Keyphrase generation aims to generate a set of condensed phrases given a source document. Although maximum likelihood estimation (MLE) based keyphrase generation methods have shown impressive performance, they suffer from the bias on the source-prediction sequence pair and the bias on the prediction-target pair. To tackle the above biases, we propose a novel correction model CorrKG on top of the MLE pipeline, where the biases are corrected via the optimal transport (OT) and a frequency-based filtering-and-sorting (FreqFS) strategy. Specifically, OT is introduced as soft correction to facilitate the alignment of salient information and rectify the semantic bias in the source document and predicted keyphrases pair. An adaptive semantic mass learning scheme is conducted on the vanilla OT to achieve a proper pair-wise optimal transport procedure, which promotes the OT learning brought by rectifying semantic masses dynamically. Besides, the FreqFS strategy is designed as hard correction to reduce the bias of predicted and ground truth keyphrases, and thus to generate accurate and sufficient keyphrases. Extensive experiments over multiple benchmark datasets show that our model achieves superior keyphrase generation as compared with the state-of-the-arts.

**JANUS: Joint Autoregressive and Non-autoregressive Training with Auxiliary Loss for Sequence Generation**
*Xiaobo Liang, Lijun Wu, Juntao Li and Min Zhang* 16:00-17:30 (Hall A, Room D)
Transformer-based autoregressive and non-autoregressive models have played an essential role in sequence generation tasks. The autoregressive model can obtain excellent performance, while the non-autoregressive model brings fast decoding speed for inference. In this paper, we propose **JANUS**, a **J**oint **A**utoregressive and **N**on-autoregressive training method using a**U**xiliary los**S** to enhance the model performance in both AR and NAR manner simultaneously and effectively alleviate the problem of distribution discrepancy. Further, we pre-train BART with JANUS on a large corpus with minimal cost (16 GPU days) and make the BART-JANUS capable of non-autoregressive generation, demonstrating that our approach can transfer the AR knowledge to NAR. Empirically, we show our approach and BART-JANUS can achieve significant improvement on multiple generation tasks, including machine translation and GLGE benchmarks. Our code is available at Github[2].

**MOCHA: A Multi-Task Training Approach for Coherent Text Generation from Cognitive Perspective**

[2] https://github.com/dropreg/JANUS

*Zhe Hu, Hou Pong Chan and Lifu Huang* 16:00-17:30 (Hall A, Room D)
Teaching neural models to generate narrative coherent texts is a critical problem. Recent pre-trained language models have achieved promising results, but there is still a gap between human written texts and machine-generated outputs. In this work, we propose a novel multi-task training strategy for long text generation grounded on the cognitive theory of writing, which empowers the model to learn essential subskills needed for writing including planning and reviewing besides end-to-end generation. We extensively evaluate our model on three open-ended generation tasks including story generation, news article writing and argument generation. Experiments show that our model achieves better results on both few-shot and fully-supervised settings than strong baselines, and human evaluations confirm that our model can generate more coherent outputs.

[DEMO] **AGReE: A system for generating Automated Grammar Reading Exercises**
*Sophia Chan, Swapna Somasundaran, Debanjan Ghosh and Mengxuan Zhao* 16:00-17:30 (Hall A, Room D)
We describe the AGReE system, which takes user-submitted passages as input and automatically generates grammar practice exercises that can be completed while reading. Multiple-choice practice items are generated for a variety of different grammar constructs: punctuation, articles, conjunctions, pronouns, prepositions, verbs, and nouns. We also conducted a large-scale human evaluation with around 4,500 multiple-choice practice items. We notice for 95

[DEMO] **Automatic Comment Generation for Chinese Student Narrative Essays**
*Zhexin Zhang, Jian Guan, Guowei Xu, Yixiang Tian and Minlie Huang* 16:00-17:30 (Hall A, Room D)
Automatic essay evaluation can help reduce teachers' workload and enable students to refine their works rapidly. Previous studies focus mainly on giving discrete scores for either the holistic quality or several distinct traits. However, real-world teachers usually provide detailed comments in natural language, which are more informative than single scores. In this paper, we present the comment generation task, which aims to generate comments for specified segments from given student narrative essays. To tackle this task, we propose a planning-based generation model, which first plans a sequence of keywords, and then expands these keywords into a complete comment. To improve the correctness and informativeness of generated comments, we adopt two following techniques: (1) training an error correction module to filter out incorrect keywords, and (2) recognizing fine-grained structured features from source essays to enrich the keywords. To support the evaluation of the task, we collect a human-written Chinese dataset, which contains 22,399 essay-comment pairs. Extensive experiments show that our model outperforms strong baselines significantly. Moreover, we exert explicit control on our model to generate comments to describe the strengths or weaknesses of inputs with a 91

## Virtual Portal 5

16:00-17:30 (Hall B)

**Eeny, meeny, miny, moe. How to choose data for morphological inflection.**
*Saliha Muradoglu and Mans Hulden* 16:00-17:30 (Hall B)
Data scarcity is a widespread problem for numerous natural language processing (NLP) tasks within low-resource languages. Within morphology, the labour-intensive task of tagging/glossing data is a serious bottleneck for both NLP and fieldwork. Active learning (AL) aims to reduce the cost of data annotation by selecting data that is most informative for the model. In this paper, we explore four sampling strategies for the task of morphological inflection using a Transformer model: a pair of oracle experiments where data is chosen based on correct/incorrect predictions by the model, model confidence, entropy, and random selection. We investigate the robustness of each sampling strategy across 30 typologically diverse languages, as well as a 10-cycle iteration using Natügu as a case study. Our results show a clear benefit to selecting data based on model confidence. Unsurprisingly, the oracle experiment, which is presented as a proxy for linguist/language informer feedback, shows the most improvement. This is followed closely by low-confidence and high-entropy forms. We also show that despite the conventional wisdom of larger data sets yielding better accuracy, introducing more instances of high-confidence, low-entropy, or forms that the model can already inflect correctly, can reduce model performance.

**Explainable Question Answering based on Semantic Graph by Global Differentiable Learning and Dynamic Adaptive Reasoning**
*Jianguo Mao, Wenbin Jiang, Xiangdong Wang, Hong Liu, Yu Xia, Yajuan Lyu and QiaoQiao She* 16:00-17:30 (Hall B)
Multi-hop Question Answering is an agent task for testing the reasoning ability. With the development of pre-trained models, the implicit reasoning ability has been surprisingly improved and can even surpass human performance. However, the nature of the black box hinders the construction of explainable intelligent systems. Several researchers have explored explainable neural-symbolic reasoning methods based on question decomposition techniques. The undifferentiable symbolic operations and the error propagation in the reasoning process lead to poor performance. To alleviate it, we propose a simple yet effective Global Differentiable Learning strategy to explore optimal reasoning paths from the latent probability space so that the model learns to solve intermediate reasoning processes without expert annotations. We further design a Dynamic Adaptive Reasoner to enhance the generalization of unseen questions. Our method achieves 17% improvements in F1-score against BreakRC and shows better interpretability. We take a step forward in building interpretable reasoning methods.

**Teaching Broad Reasoning Skills for Multi-Step QA by Generating Hard Contexts**
*Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot and Ashish Sabharwal* 16:00-17:30 (Hall B)
Question-answering datasets require a broad set of reasoning skills. We show how to use question decompositions to teach language models these broad reasoning skills in a robust fashion. Specifically, we use widely available QDMR representations to programmatically create hard-to-cheat synthetic contexts for real questions in six multi-step reasoning datasets. These contexts are carefully designed to avoid common reasoning shortcuts prevalent in real contexts that prevent models from learning the right skills. This results in a pretraining dataset, named TeaBReaC, containing 525K multi-step questions (with associated formal programs) covering about 900 reasoning patterns. We show that pretraining standard language models (LMs) on TeaBReaC before fine-tuning them on target datasets improves their performance by up to 13 F1 points across 4 multi-step QA datasets, with up to 21 point gain on more complex questions. The resulting models also demonstrate higher robustness, with a 5-8 F1 point improvement on two contrast sets. Furthermore, TeaBReaC pretraining substantially improves model performance and robustness even when starting with numerate LMs pretrained using recent methods (e.g., PReasM, POET). Our work thus shows how to effectively use decomposition-guided contexts to robustly teach multi-step reasoning.

**DuQM: A Chinese Dataset of Linguistically Perturbed Natural Questions for Evaluating the Robustness of Question Matching Models**
*Hongyu Zhu, Yan Chen, Jing Yan, Jing Liu, Yu Hong, Ying Chen, Hua Wu and Haifeng Wang* 16:00-17:30 (Hall B)
In this paper, we focus on the robustness evaluation of Chinese Question Matching (QM) models. Most of the previous work on analyzing robustness issues focus on just one or a few types of artificial adversarial examples. Instead, we argue that a comprehensive evaluation should be conducted on natural texts, which takes into account the fine-grained linguistic capabilities of QM models. For this purpose, we create

a Chinese dataset namely DuQM which contains natural questions with linguistic perturbations to evaluate the robustness of QM models. DuQM contains 3 categories and 13 subcategories with 32 linguistic perturbations. The extensive experiments demonstrate that DuQM has a better ability to distinguish different models. Importantly, the detailed breakdown of evaluation by the linguistic phenomena in DuQM helps us easily diagnose the strength and weakness of different models. Additionally, our experiment results show that the effect of artificial adversarial examples does not work on natural texts. Our baseline codes and a leaderboard are now publicly available.

**Structure-Unified M-Tree Coding Solver for Math Word Problem**
*Bin Wang, Jiangzhou Ju, Yang Fan, Xinyu Dai, Shujian Huang and Jiajun CHEN* 16:00-17:30 (Hall B)
As one of the challenging NLP tasks, designing math word problem (MWP) solvers has attracted increasing research attention for the past few years. In previous work, models designed by taking into account the properties of the binary tree structure of mathematical expressions at the output side have achieved better performance. However, the expressions corresponding to a MWP are often diverse (e.g., $n\_1 + n\_2 \times n\_3 - n\_4$, $n\_3 \times n\_2 - n\_4 + n\_1$, etc.), and so are the corresponding binary trees, which creates difficulties in model learning due to the non-deterministic output space. In this paper, we propose the Structure-Unified M-Tree Coding Solver (SUMC-Solver), which applies a tree with any M branches (M-tree) to unify the output structures. To learn the M-tree, we use a mapping to convert the M-tree into the M-tree codes, where codes store the information of the paths from tree root to leaf nodes and the information of leaf nodes themselves, and then devise a Sequence-to-Code (seq2code) model to generate the codes. Experimental results on the widely used MAWPS and Math23K datasets have demonstrated that SUMC-Solver not only outperforms several state-of-the-art models under similar experimental settings but also performs much better under low-resource conditions.

**Graph-Induced Transformers for Efficient Multi-Hop Question Answering**
*Giwon Hong, Jeonghwan Kim, Junmo Kang and Sung-Hyon Myaeng* 16:00-17:30 (Hall B)
A graph is a suitable data structure to represent the structural information of text. Recently, multi-hop question answering (MHQA) tasks, which require inter-paragraph/sentence linkages, have come to exploit such properties of a graph. Previous approaches to MHQA relied on leveraging the graph information along with the pre-trained language model (PLM) encoders. However, this trend exhibits the following drawbacks: (i) sample inefficiency while training in a low-resource setting; (ii) lack of reusability due to changes in the model structure or input. Our work proposes the Graph-Induced Transformer (GIT) that applies graph-derived attention patterns directly into a PLM, without the need to employ external graph modules. GIT can leverage the useful inductive bias of graphs while retaining the unperturbed Transformer structure and parameters. Our experiments on HotpotQA successfully demonstrate both the sample efficient characteristic of GIT and its capacity to replace the graph modules while preserving model performance.

**Pre-training Language Models with Deterministic Factual Knowledge**
*Shaobo Li, Xiaoguang Li, Lifeng Shang, Chengjie Sun, Bingquan Liu, zhenzhou Ji, Xin Jiang and Qun Liu* 16:00-17:30 (Hall B)
Previous works show that Pre-trained Language Models (PLMs) can capture factual knowledge. However, some analyses reveal that PLMs fail to perform it robustly, e.g., being sensitive to the changes of prompts when extracting factual knowledge. To mitigate this issue, we propose to let PLMs learn the deterministic relationship between the remaining context and the masked content. The deterministic relationship ensures that the masked factual content can be deterministically inferable based on the existing clues in the context. That would provide more stable patterns for PLMs to capture factual knowledge than randomly masking. Two pre-training tasks are further introduced to motivate PLMs to rely on the deterministic relationship when filling masks. Specifically, we use an external Knowledge Base (KB) to identify deterministic relationships and continuously pre-train PLMs with the proposed methods. The factual knowledge probing experiments indicate that the continuously pre-trained PLMs achieve better robustness in factual knowledge capturing. Further experiments on question-answering datasets show that trying to learn a deterministic relationship with the proposed methods can also help other knowledge-intensive tasks.

**OpenCQA: Open-ended Question Answering with Charts**
*Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Leong, Jia Qing Tan, Enamul Hoque and Shafiq Joty* 16:00-17:30 (Hall B)
Charts are very popular to analyze data and convey important insights. People often analyze visualizations to answer open-ended questions that require explanatory answers. Answering such questions are often difficult and time-consuming as it requires a lot of cognitive and perceptual efforts. To address this challenge, we introduce a new task called OpenCQA, where the goal is to answer an open-ended question about a chart with descriptive texts. We present the annotation process and an in-depth analysis of our dataset. We implement and evaluate a set of baselines under three practical settings. In the first setting, a chart and the accompanying article is provided as input to the model. The second setting provides only the relevant paragraph(s) to the chart instead of the entire article, whereas the third setting requires the model to generate an answer solely based on the chart. Our analysis of the results show that the top performing models generally produce fluent and coherent text while they struggle to perform complex logical and arithmetic reasoning.

**Title2Event: Benchmarking Open Event Extraction with a Large-scale Chinese Title Dataset**
*Haolin Deng, Yanan Zhang, Yangfan Zhang, Wangyang Ying, Changlong Yu, Jun Gao, Wei Wang, Xiaoling Bai, Nan Yang, Jin Ma, xiang chen and Tianhua Zhou* 16:00-17:30 (Hall B)
Event extraction (EE) is crucial to downstream tasks such as new aggregation and event knowledge graph construction. Most existing EE datasets manually define fixed event types and design specific schema for each of them, failing to cover diverse events emerging from the online text. Moreover, news titles, an important source of event mentions, have not gained enough attention in current EE research. In this paper, we present Title2Event, a large-scale sentence-level dataset benchmarking Open Event Extraction without restricting event types. Title2Event contains more than 42,000 news titles in 34 topics collected from Chinese web pages. To the best of our knowledge, it is currently the largest manually annotated Chinese dataset for open event extraction. We further conduct experiments on Title2Event with different models and show that the characteristics of titles make it challenging for event extraction, addressing the significance of advanced study on this problem. The dataset and baseline codes are available at https://open-event-hub.github.io/title2event.

**CN-AutoMIC: Distilling Chinese Commonsense Knowledge from Pretrained Language Models**
*Chenhao Wang, Jiachun Li, Yubo Chen, Kang Liu and Jun Zhao* 16:00-17:30 (Hall B)
Commonsense knowledge graphs (CKGs) are increasingly applied in various natural language processing tasks. However, most existing CKGs are limited to English, which hinders related research in non-English languages. Meanwhile, directly generating commonsense knowledge from pretrained language models has recently received attention, yet it has not been explored in non-English languages. In this paper, we propose a large-scale Chinese CKG generated from multilingual PLMs, named as **CN-AutoMIC**, aiming to fill the research gap of non-English CKGs. To improve the efficiency, we propose generate-by-category strategy to reduce invalid generation. To ensure the filtering quality, we develop cascaded filters to discard low-quality results. To further increase the diversity and density, we introduce a bootstrapping iteration process to reuse generated results. Finally, we conduct detailed analyses on CN-AutoMIC from different aspects. Empirical results show the proposed CKG has high quality and diversity, surpassing the direct translation version of similar English CKGs. We also find some interesting deficiency patterns and differences between relations, which reveal pending problems in commonsense knowledge generation. We share the resources and related models for further study.

**Improving Large-scale Paraphrase Acquisition and Generation**

*Yao Dou, Chao Jiang and Wei Xu* 16:00-17:30 (Hall B)
This paper addresses the quality issues in existing Twitter-based paraphrase datasets, and discusses the necessity of using two separate definitions of paraphrase for identification and generation tasks. We present a new Multi-Topic Paraphrase in Twitter (MultiPIT) corpus that consists of a total of 130k sentence pairs with crowdsoursing (MultiPIT_crowd) and expert (MultiPIT_expert) annotations using two different paraphrase definitions for paraphrase identification, in addition to a multi-reference test set (MultiPIT_NMR) and a large automatically constructed training set (MultiPIT_Auto) for paraphrase generation. With improved data annotation quality and task-specific paraphrase definition, the best pre-trained language model fine-tuned on our dataset achieves the state-of-the-art performance of 84.2 F1 for automatic paraphrase identification. Furthermore, our empirical results also demonstrate that the paraphrase generation models trained on MultiPIT_Auto generate more diverse and high-quality paraphrases compared to their counterparts fine-tuned on other corpora such as Quora, MSCOCO, and ParaNMT.

**A Survey of Computational Framing Analysis Approaches**
*Mohammad Ali and Naeemul Hassan* 16:00-17:30 (Hall B)
Framing analysis is predominantly qualitative and quantitative, examining a small dataset with manual coding. Easy access to digital data in the last two decades prompts scholars in both computation and social sciences to utilize various computational methods to explore frames in large-scale datasets. The growing scholarship, however, lacks a comprehensive understanding and resources of computational framing analysis methods. Aiming to address the gap, this article surveys existing computational framing analysis approaches and puts them together. The research is expected to help scholars and journalists gain a deeper understanding of how frames are being explored computationally, better equip them to analyze frames in large-scale datasets, and, finally, work on advancing methodological approaches.

**CRIPP-VQA: Counterfactual Reasoning about Implicit Physical Properties via Video Question Answering**
*Maitreya Patel, Tejas Gokhale, Chitta Baral and Yezhou Yang* 16:00-17:30 (Hall B)
Videos often capture objects, their visible properties, their motion, and the interactions between different objects. Objects also have physical properties such as mass, which the imaging pipeline is unable to directly capture. However, these properties can be estimated by utilizing cues from relative object motion and the dynamics introduced by collisions. In this paper, we introduce CRIPP-VQA, a new video question answering dataset for reasoning about the implicit physical properties of objects in a scene. CRIPP-VQA contains videos of objects in motion, annotated with questions that involve counterfactual reasoning about the effect of actions, questions about planning in order to reach a goal, and descriptive questions about visible properties of objects. The CRIPP-VQA test set enables evaluation under several out-of-distribution settings – videos with objects with masses, coefficients of friction, and initial velocities that are not observed in the training distribution. Our experiments reveal a surprising and significant performance gap in terms of answering questions about implicit properties (the focus of this paper) and explicit properties of objects (the focus of prior work).

**GuoFeng: A Benchmark for Zero Pronoun Recovery and Translation**
*Mingzhou Xu, Longyue Wang, Derek F. Wong, Hongye Liu, Linfeng Song, Lidia S. Chao, Shuming Shi and Zhaopeng Tu*16:00-17:30 (Hall B)
The phenomenon of zero pronoun (ZP) has attracted increasing interest in the machine translation (MT) community due to its importance and difficulty. However, previous studies generally evaluate the quality of translating ZPs with BLEU scores on MT testsets, which is not expressive or sensitive enough for accurate assessment. To bridge the data and evaluation gaps, we propose a benchmark testset for target evaluation on Chinese-English ZP translation. The human-annotated testset covers five challenging genres, which reveal different characteristics of ZPs for comprehensive evaluation. We systematically revisit eight advanced models on ZP translation and identify current challenges for future exploration. We release data, code, models and annotation guidelines, which we hope can significantly promote research in this field (https://github.com/longyuewangdcu/mZPRT).

**Effective and Efficient Query-aware Snippet Extraction for Web Search**
*Jingwei Yi, Fangzhao Wu, Chuhan Wu, Xiaolong Huang, Binxing Jiao, Guangzhong Sun and Xing Xie* 16:00-17:30 (Hall B)
Query-aware webpage snippet extraction is widely used in search engines to help users better understand the content of the returned webpages before clicking. The extracted snippet is expected to summarize the webpage in the context of the input query. Existing snippet extraction methods mainly rely on handcrafted features of overlapping words, which cannot capture deep semantic relationships between the query and webpages. Another idea is to extract the sentences which are most relevant to queries as snippets with existing text matching methods. However, these methods ignore the contextual information of webpages, which may be sub-optimal. In this paper, we propose an effective query-aware webpage snippet extraction method named DeepQSE. In DeepQSE, the concatenation of title, query and each candidate sentence serves as an input of query-aware sentence encoder, aiming to capture the fine-grained relevance between the query and sentences. Then, these query-aware sentence representations are modeled jointly through a document-aware relevance encoder to capture contextual information of the webpage. Since the query and each sentence are jointly modeled in DeepQSE, its online inference may be slow. Thus, we further propose an efficient version of DeepQSE, named Efficient-DeepQSE, which can significantly improve the inference speed of DeepQSE without affecting its performance. The core idea of Efficient-DeepQSE is to decompose the query-aware snippet extraction task into two stages, i.e., a coarse-grained candidate sentence selection stage where sentence representations can be cached, and a fine-grained relevance modeling stage. Experiments on two datasets validate the effectiveness and efficiency of our methods.

**Opinion Summarization by Weak-Supervision from Mix-structured Data**
*Yizhu Liu, Qi Jia and Kenny Zhu* 16:00-17:30 (Hall B)
Opinion summarization of multiple reviews suffers from the lack of reference summaries for training. Most previous approaches construct multiple reviews and their summary based on textual similarities between reviews, resulting in information mismatch between the review input and the summary. In this paper, we convert each review into a mix of structured and unstructured data, which we call opinion-aspect pairs (OAs) and implicit sentences (ISs). We propose a new method to synthesize training pairs of such mix-structured data as input and the textual summary as output, and design a summarization model with OA encoder and IS encoder. Experiments show that our approach outperforms previous methods on Yelp, Amazon and RottenTomatos datasets.

**Improving Faithfulness by Augmenting Negative Summaries from Fake Documents**
*Tianshu Wang, Faisal Ladhak, Esin Durmus and He He* 16:00-17:30 (Hall B)
Current abstractive summarization systems tend to hallucinate content that is unfaithful to the source document, posing a risk of misinformation. To mitigate hallucination, we must teach the model to distinguish hallucinated summaries from faithful ones. However, the commonly used maximum likelihood training does not disentangle factual errors from other model errors. To address this issue, we propose a back-translation-style approach to augment negative samples that mimic factual errors made by the model. Specifically, we train an elaboration model that generates hallucinated documents given the reference summaries, and then generates negative summaries from the fake documents. We incorporate the negative samples into training through a controlled generator, which produces faithful/unfaithful summaries conditioned on the control codes. Additionally, we find that adding textual entailment data through multitasking further boosts the performance. Experiments on three datasets (XSum, Gigaword, and WikiHow) show that our method consistently improves faithfulness without sacrificing informativeness according to both human and automatic evaluation

[DEMO] **MIC: A Multi-task Interactive Curation Tool**

*Shi Yu, Mingfeng Yang, Jerrod Parker and Stephen Brock* 16:00-17:30 (Hall B)
This paper introduces MIC, a Multi-task Interactive Curation tool, a human-machine collaborative curation tool for multiple NLP tasks. The tool aims to borrow recent advances in literature to solve pain-points in real NLP tasks. Firstly, it supports multiple projects with multiple users which enables collaborative annotations. Secondly, MIC allows easy integration of pre-trained models, rules, and dictionaries to auto label the text and speed up the labeling process. Thirdly, MIC supports annotation at different scales (span of characters and words, tokens and lines, or document) and different types (free text, sentence labels, entity labels, and relationship triplets) with easy GUI operations.

[DEMO] **POTATO: The Portable Text Annotation Tool**
*Jiaxin Pei, Aparna Kamakshi Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent and David Jurgens*
16:00-17:30 (Hall B)
We present POTATO, the Portable text annotation tool, a free, fully open-sourced annotation system that 1) supports labeling many types of text and multimodal data; 2) offers easy-to-configure features to maximize the productivity of both deployers and annotators (convenient templates for common ML/NLP tasks, active learning, keypress shortcuts, keyword highlights, tooltips); and 3) supports a high degree of customization (editable UI, inserting pre-screening questions, attention and qualification tests). Experiments over two annotation tasks suggest that POTATO improves labeling speed through its specially-designed productivity features, especially for long documents and complex tasks. POTATO is available at https://github.com/davidjurgens/potato and will continue to be updated.

## Virtual Portal 6

16:00-17:30 (Collaboratorium)

**Kernel-Whitening: Overcome Dataset Bias with Isotropic Sentence Embedding**
*SongYang Gao, Shihan Dou, Qi Zhang and Xuanjing Huang* 16:00-17:30 (Collaboratorium)
Dataset bias has attracted increasing attention recently for its detrimental effect on the generalization ability of fine-tuned models. The current mainstream solution is designing an additional shallow model to pre-identify biased instances. However, such two-stage methods scale up the computational complexity of training process and obstruct valid feature information while mitigating bias. To address this issue, we utilize the representation normalization method which aims at disentangling the correlations between features of encoded sentences. We find it also promising in eliminating the bias problem by providing isotropic data distribution. We further propose Kernel-Whitening, a Nystrom kernel approximation method to achieve more thorough debiasing on nonlinear spurious correlations. Our framework is end-to-end with similar time consumption to fine-tuning. Experiments show that Kernel-Whitening significantly improves the performance of BERT on out-of-distribution datasets while maintaining in-distribution accuracy.

**Neural-Symbolic Inference for Robust Autoregressive Graph Parsing via Compositional Uncertainty Quantification**
*Zi Lin, Jeremiah Liu and Jingbo Shang* 16:00-17:30 (Collaboratorium)
Pre-trained seq2seq models excel at graph semantic parsing with rich annotated data, but generalize worse to out-of-distribution (OOD) and long-tail examples. In comparison, symbolic parsers under-perform on population-level metrics, but exhibit unique strength in OOD and tail generalization. In this work, we study compositionality-aware approach to neural-symbolic inference informed by model confidence, performing fine-grained neural-symbolic reasoning at subgraph level (i.e., nodes and edges) and precisely targeting subgraph components with high uncertainty in the neural parser. As a result, the method combines the distinct strength of the neural and symbolic approaches in capturing different aspects of the graph prediction, leading to well-rounded generalization performance both across domains and in the tail. We empirically investigate the approach in the English Resource Grammar (ERG) parsing problem on a diverse suite of standard in-domain and seven OOD corpora. Our approach leads to 35.26% and 35.60% error reduction in aggregated SMATCH score over neural and symbolic approaches respectively, and 14% absolute accuracy gain in key tail linguistic categories over the neural model, outperforming prior state-of-art methods that do not account for compositionality or uncertainty.

**Leveraging Affirmative Interpretations from Negation Improves Natural Language Understanding**
*Md Mosharaf Hossain and Eduardo Blanco* 16:00-17:30 (Collaboratorium)
Negation poses a challenge in many natural language understanding tasks. Inspired by the fact that understanding a negated statement often requires humans to infer affirmative interpretations, in this paper we show that doing so benefits models for three natural language understanding tasks. We present an automated procedure to collect pairs of sentences with negation and their affirmative interpretations, resulting in over 150,000 pairs. Experimental results show that leveraging these pairs helps (a) T5 generate affirmative interpretations from negations in a previous benchmark, and (b) a RoBERTa-based classifier solve the task of natural language inference. We also leverage our pairs to build a plug-and-play neural generator that given a negated statement generates an affirmative interpretation. Then, we incorporate the pretrained generator into a RoBERTa-based classifier for sentiment analysis and show that doing so improves the results. Crucially, our proposal does not require any manual effort.

**PromptBERT: Improving BERT Sentence Embeddings with Prompts**
*Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, deqing wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng and Qi Zhang*
16:00-17:30 (Collaboratorium)
We propose PromptBERT, a novel contrastive learning method for learning better sentence representation. We firstly analysis the drawback of current sentence embedding from original BERT and find that it is mainly due to the static token embedding bias and ineffective BERT layers. Then we propose the first prompt-based sentence embeddings method and discuss two prompt representing methods and three prompt searching methods to make BERT achieve better sentence embeddings .Moreover, we propose a novel unsupervised training objective by the technology of template denoising, which substantially shortens the performance gap between the supervised and unsupervised settings. Extensive experiments show the effectiveness of our method. Compared to SimCSE, PromptBert achieves 2.29 and 2.58 points of improvement based on BERT and RoBERTa in the unsupervised setting.

**An Empirical Revisiting of Linguistic Knowledge Fusion in Language Understanding Tasks**
*Changlong Yu, Tianyi Xiao, Lingpeng Kong, Yangqiu Song and Wilfred Ng* 16:00-17:30 (Collaboratorium)
Though linguistic knowledge emerges during large-scale language model pretraining, recent work attempt to explicitly incorporate human-defined linguistic priors into task-specific fine-tuning. Infusing language models with syntactic or semantic knowledge from parsers has shown improvements on many language understanding tasks. To further investigate the effectiveness of structural linguistic priors, we conduct empirical study of replacing parsed graphs or trees with trivial ones (rarely carrying linguistic knowledge e.g., balanced tree) for tasks in the GLUE benchmark. Encoding with trivial graphs achieves competitive or even better performance in fully-supervised and few-shot settings. It reveals that the gains might not be significantly attributed to explicit linguistic priors but rather to more feature interactions brought by fusion layers. Hence we call for attention to using trivial graphs as necessary baselines to design advanced knowledge fusion methods in the future.

**Cross-domain Generalization for AMR Parsing**
*Xuefeng Bai, Sen Yang, Leyang Cui, Linfeng Song and Yue Zhang* 16:00-17:30 (Collaboratorium)
Abstract Meaning Representation (AMR) parsing aims to predict an AMR graph from textual input. Recently, there has been notable growth in AMR parsing performance. However, most existing work focuses on improving the performance in the specific domain, ignoring the potential domain dependence of AMR parsing systems. To address this, we extensively evaluate five representative AMR parsers on five domains and analyze challenges to cross-domain AMR parsing. We observe that challenges to cross-domain AMR parsing mainly arise from the distribution shift of words and AMR concepts. Based on our observation, we investigate two approaches to reduce the domain distribution divergence of text and AMR features, respectively. Experimental results on two out-of-domain test sets show the superiority of our method.

**Mutual Exclusivity Training and Primitive Augmentation to Induce Compositionality**
*Yichen Jiang, Xiang Zhou and Mohit Bansal* 16:00-17:30 (Collaboratorium)
Recent datasets expose the lack of the systematic generalization ability in standard sequence-to-sequence models. In this work, we analyze this behavior of seq2seq models and identify two contributing factors: a lack of mutual exclusivity bias (one target sequence can only be mapped to one source sequence), and the tendency to memorize whole examples rather than separating structures from contents. We propose two techniques to address these two issues respectively: Mutual Exclusivity Training that prevents the model from producing seen generations when facing novel examples via an unlikelihood-based loss, and prim2primX data augmentation that automatically diversifies the arguments of every syntactic function to prevent memorizing and provide a compositional inductive bias without exposing test-set data. Combining these two techniques, we show substantial empirical improvements using standard sequence-to-sequence models (LSTMs and Transformers) on two widely-used compositionality datasets: SCAN and COGS. Finally, we provide analysis characterizing the improvements as well as the remaining challenges, and provide detailed ablations of our method.

**Mitigating Inconsistencies in Multimodal Sentiment Analysis under Uncertain Missing Modalities**
*Jiandian Zeng, Jiantao Zhou and Tianyi Liu* 16:00-17:30 (Collaboratorium)
For the missing modality problem in Multimodal Sentiment Analysis (MSA), the inconsistency phenomenon occurs when the sentiment changes due to the absence of a modality. The absent modality that determines the overall semantic can be considered as a key missing modality. However, previous works all ignored the inconsistency phenomenon, simply discarding missing modalities or solely generating associated features from available modalities. The neglect of the key missing modality case may lead to incorrect semantic results. To tackle the issue, we propose an Ensemble-based Missing Modality Reconstruction (EMMR) network to detect and recover semantic features of the key missing modality. Specifically, we first learn joint representations with remaining modalities via a backbone encoder-decoder network. Then, based on the recovered features, we check the semantic consistency to determine whether the absent modality is crucial to the overall sentiment polarity. Once the inconsistency problem due to the key missing modality exists, we integrate several encoder-decoder approaches for better decision making. Extensive experiments and analyses are conducted on CMU-MOSI and IEMOCAP datasets, validating the superiority of the proposed method.

**Pair-Based Joint Encoding with Relational Graph Convolutional Networks for Emotion-Cause Pair Extraction**
*Junlong Liu, Xichen Shang and Qianli Ma* 16:00-17:30 (Collaboratorium)
Emotion-cause pair extraction (ECPE) aims to extract emotion clauses and corresponding cause clauses, which have recently received growing attention. Previous methods sequentially encode features with a specified order. They first encode the emotion and cause features for clause extraction and then combine them for pair extraction. This lead to an imbalance in inter-task feature interaction where features extracted later have no direct contact with the former. To address this issue, we propose a novel **P**air-**B**ased **J**oint **E**ncoding (**PBJE**) network, which generates pairs and clauses features simultaneously in a joint feature encoding manner to model the causal relationship in clauses. PBJE can balance the information flow among emotion clauses, cause clauses and pairs. From a multi-relational perspective, we construct a heterogeneous undirected graph and apply the Relational Graph Convolutional Network (RGCN) to capture the multiplex relationship between clauses and the relationship between pairs and clauses. Experimental results show that PBJE achieves state-of-the-art performance on the Chinese benchmark corpus.

**UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition**
*Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu and Yongbin Li* 16:00-17:30 (Collaboratorium)
Multimodal sentiment analysis (MSA) and emotion recognition in conversation (ERC) are key research topics for computers to understand human behaviors. From a psychological perspective, emotions are the expression of affect or feelings during a short period, while sentiments are formed and held for a longer period. However, most existing works study sentiment and emotion separately and do not fully exploit the complementary knowledge behind the two. In this paper, we propose a multimodal sentiment knowledge-sharing framework (UniMSE) that unifies MSA and ERC tasks from features, labels, and models. We perform modality fusion at the syntactic and semantic levels and introduce contrastive learning between modalities and samples to better capture the difference and consistency between sentiments and emotions. Experiments on four public benchmark datasets, MOSI, MOSEI, MELD, and IEMOCAP, demonstrate the effectiveness of the proposed method and achieve consistent improvements compared with state-of-the-art methods.

**Argument Mining for Review Helpfulness Prediction**
*Zaiqian Chen, Daniel Verdi do Amarante, Jenna Donaldson, Yohan Jo and Joonsuk Park* 16:00-17:30 (Collaboratorium)
The importance of reliably determining the helpfulness of product reviews is rising as both helpful and unhelpful reviews continue to accumulate on e-commerce websites. And argumentational features—such as the structure of arguments and the types of underlying elementary units—have shown to be promising indicators of product review helpfulness. However, their adoption has been limited due to the lack of sufficient resources and large-scale experiments investigating their utility. To this end, we present the AMazon Argument Mining ($AM^2$) corpus—a corpus of 878 Amazon reviews on headphones annotated according to a theoretical argumentation model designed to evaluate argument quality. Experiments show that employing argumentational features leads to statistically significant improvements over the state-of-the-art review helpfulness predictors under both text-only and text-and-image settings.

**Prompt-based Distribution Alignment for Domain Generalization in Text Classification**
*Chen Jia and Yue Zhang* 16:00-17:30 (Collaboratorium)
Prompt-based learning (a.k.a. prompting) achieves high performance by bridging the gap between the objectives of language modeling and downstream tasks. Domain generalization ability can be improved by prompting since classification across different domains can be unified into the prediction of the same set of label words. The remaining challenge for domain generalization by prompting comes from discrepancies between the data distribution of different domains. To improve domain generalization with prompting, we learn distributional invariance across source domains via two alignment regularization loss functions. The first is vocabulary distribution alignment, which uses a Kullback-Leibler divergence regularization on source-domain vocabulary distributions. The second is feature distribution alignment, which uses a novel adversarial training strategy to learn domain invariant representation across source domains. Experiments on sentiment analysis and natural language inference show the effectiveness of our method and achieve state-of-the-art results on six datasets.

**A Generative Model for End-to-End Argument Mining with Reconstructed Positional Encoding and Constrained Pointer Mechanism**
*Jianzhu Bao, Yuhang He, Yang Sun, Bin Liang, Jiachen Du, Bing Qin, Min Yang and Ruifeng Xu* 16:00-17:30 (Collaboratorium)
Argument mining (AM) is a challenging task as it requires recognizing the complex argumentation structures involving multiple subtasks. To handle all subtasks of AM in an end-to-end fashion, previous works generally transform AM into a dependency parsing task. However, such methods largely require complex pre- and post-processing to realize the task transformation. In this paper, we investigate the end-to-end AM task from a novel perspective by proposing a generative framework, in which the expected outputs of AM are framed as a simple target sequence. Then, we employ a pre-trained sequence-to-sequence language model with a constrained pointer mechanism (CPM) to model the clues for all the subtasks of AM in the light of the target sequence. Furthermore, we devise a reconstructed positional encoding (RPE) to alleviate the order biases induced by the autoregressive generation paradigm. Experimental results show that our proposed framework achieves new state-of-the-art performance on two AM benchmarks.

**Semantic Simplification for Sentiment Classification**
*Xiaotong Jiang, Zhongqing Wang and Guodong Zhou* 16:00-17:30 (Collaboratorium)
Recent work on document-level sentiment classification has shown that the sentiment in the original text is often hard to capture, since the sentiment is usually either expressed implicitly or shifted due to the occurrences of negation and rhetorical words. To this end, we enhance the original text with a sentiment-driven simplified clause to intensify its sentiment. The simplified clause shares the same opinion with the original text but expresses the opinion much more simply. Meanwhile, we employ Abstract Meaning Representation (AMR) for generating simplified clauses, since AMR explicitly provides core semantic knowledge, and potentially offers core concepts and explicit structures of original texts. Empirical studies show the effectiveness of our proposed model over several strong baselines. The results also indicate the importance of simplified clauses for sentiment classification.

[CL] **Dimensional Modeling of Emotions in Text with Appraisal Theories: Corpus Creation, Annotation Reliability, and Prediction**
*Enrica Troiano, Laura Oberlaender and Roman Klinger* 16:00-17:30 (Collaboratorium)
The most prominent tasks in emotion analysis are to assign emotions to texts and to understand how emotions manifest in language. An observation for NLP is that emotions can be communicated implicitly by referring to events, appealing to an empathetic, intersubjective understanding of events, even without explicitly mentioning an emotion name. In psychology, the class of emotion theories known as appraisal theories aims at explaining the link between events and emotions. Appraisals can be formalized as variables that measure a cognitive evaluation by people living through an event that they consider relevant. They include the assessment if an event is novel, if the person considers themselves to be responsible, if it is in line with the own goals, and many others. Such appraisals explain which emotions are developed based on an event, e.g., that a novel situation can induce surprise or one with uncertain consequences could evoke fear. We analyze the suitability of appraisal theories for emotion analysis in text with the goal of understanding if appraisal concepts can reliably be reconstructed by annotators, if they can be predicted by text classifiers, and if appraisal concepts help to identify emotion categories. To achieve that, we compile a corpus by asking people to textually describe events that triggered particular emotions and to disclose their appraisals. Then, we ask readers to reconstruct emotions and appraisals from the text. This setup allows us to measure if emotions and appraisals can be recovered purely from text and provides a human baseline. Our comparison of text classification methods to human annotators shows that both can reliably detect emotions and appraisals with similar performance. Therefore, appraisals constitute an alternative computational emotion analysis paradigm and further improve the categorization of emotions in text with joint models.

**LiteVL: Efficient Video-Language Learning with Enhanced Spatial-Temporal Modeling**
*Dongsheng Chen, Chaofan Tao, Lu Hou, Lifeng Shang, Xin Jiang and Qun Liu* 16:00-17:30 (Collaboratorium)
Recent large-scale video-language pre-trained models have shown appealing performance on various downstream tasks. However, the pre-training process is computationally expensive due to the requirement of millions of video-text pairs and the redundant data structure of each video. To mitigate these problems, we propose LiteVL, which adapts a pre-trained image-language model BLIP into a video-text model directly on downstream tasks, without heavy pre-training. To enhance the temporal modeling lacking in the image-language model, we propose to add temporal attention modules in the image encoder of BLIP with dynamic temporal scaling. Besides the model-wise adaptation, we also propose a non-parametric pooling mechanism to adaptively reweight the fine-grained video embedding conditioned on the text. Experimental results on text-video retrieval and video question answering show that the proposed LiteVL even outperforms previous video-language pre-trained models by a clear margin, though without any video-language pre-training.

**ALFRED-L: Investigating the Role of Language for Action Learning in Interactive Visual Environments**
*Arjun Akula, Spandana Gella, Aishwarya Padmakumar, Mahdi Namazifar, Mohit Bansal, Jesse Thomason and Dilek Hakkani-Tur* 16:00-17:30 (Collaboratorium)
Embodied Vision and Language Task Completion requires an embodied agent to interpret natural language instructions and egocentric visual observations to navigate through and interact with environments. In this work, we examine ALFRED, a challenging benchmark for embodied task completion, with the goal of gaining insight into how effectively models utilize language. We find evidence that sequence-to-sequence and transformer-based models trained on this benchmark are not sufficiently sensitive to changes in input language instructions. Next, we construct a new test split – ALFRED-L to test whether ALFRED models can generalize to task structures not seen during training that intuitively require the same types of language understanding required in ALFRED. Evaluation of existing models on ALFRED-L suggests that (a) models are overly reliant on the sequence in which objects are visited in typical ALFRED trajectories and fail to adapt to modifications of this sequence and (b) models trained with additional augmented trajectories are able to adapt relatively better to such changes in input language instructions.

**Directions for NLP Practices Applied to Online Hate Speech Detection**
*Paula Fortuna, Monica Dominguez, Leo Wanner and Zeerak Talat* 16:00-17:30 (Collaboratorium)
Addressing hate speech in online spaces has been conceptualized as a classification task that uses Natural Language Processing (NLP) techniques. Through this conceptualization, the hate speech detection task has relied on common conventions and practices from NLP. For instance, inter-annotator agreement is conceptualized as a way to measure dataset quality and certain metrics and benchmarks are used to assure model generalization. However, hate speech is a deeply complex and situated concept that eludes such static and disembodied practices. In this position paper, we critically reflect on these methodologies for hate speech detection, we argue that many conventions in NLP are poorly suited for the problem and encourage researchers to develop methods that are more appropriate for the task.

[DEMO] **An Explainable Toolbox for Evaluating Pre-trained Vision-Language Models**
*Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu and Jianwei Yin* 16:00-17:30 (Collaboratorium)
We introduce VL-CheckList, a toolbox for evaluating Vision-Language Pretraining (VLP) models, including the preliminary datasets that deepen the image-texting ability of a VLP model. Most existing VLP works evaluated their systems by comparing the fine-tuned downstream task performance. However, only average downstream task accuracy provides little information about the pros and cons of each VLP method. In this paper, we demonstrate how minor input changes in language and vision will affect the prediction outputs. Then, we describe the

detailed user guidelines to utilize and contribute to the community. We show new findings on one of the representative VLP models to provide an example analysis. The data/code is available at https://github.com/om-ai-lab/VL-CheckList

## Poster Sessions 5 & 6

16:00-17:30 (Atrium)

---

**Structural Constraints and Natural Language Inference for End-to-End Flowchart Grounded Dialog Response Generation**
*Dinesh Raghu, Suraj Joshi, Sachindra Joshi and Mausam* - 16:00-17:30 (Atrium)
Flowchart grounded dialog systems converse with users by following a given flowchart and a corpus of FAQs. The existing state-of-the-art approach (Raghu et al, 2021) for learning such a dialog system, named FLONET, has two main limitations. (1) It uses a Retrieval Augmented Generation (RAG) framework which represents a flowchart as a bag of nodes. By doing so, it loses the connectivity structure between nodes that can aid in better response generation. (2) Typically dialogs progress with the agent asking polar (Y/N) questions, but users often respond indirectly without the explicit use of polar words. In such cases, it fails to understand the correct polarity of the answer. To overcome these issues, we propose Structure-Aware FLONET (SA-FLONET) which infuses structural constraints derived from the connectivity structure of flowcharts into the RAG framework. It uses natural language inference to better predict the polarity of indirect Y/N answers. We find that SA-FLONET outperforms FLONET, with a success rate improvement of 68% and 123% in flowchart grounded response generation and zero-shot flowchart grounded response generation tasks respectively.

**Should We Ban English NLP for a Year?**
*Anders Søgaard* 16:00-17:30 (Atrium)
Around two thirds of NLP research at top venues is devoted exclusively to developing technology for speakers of English, most speech data comes from young urban speakers, and most texts used to train language models come from male writers. These biases feed into consumer technologies to widen existing inequality gaps, not only within, but also across, societies. Many have argued that it is almost impossible to mitigate inequality amplification. I argue that, on the contrary, it is quite simple to do so, and that counter-measures would have little-to-no negative impact, except for, perhaps, in the very short term.

**That's the Wrong Lung! Evaluating and Improving the Interpretability of Unsupervised Multimodal Encoders for Medical Data**
*Jered McInerney, Geoffrey Young, Jan-Willem van de Meent and Byron Wallace* 16:00-17:30 (Atrium)
Pretraining multimodal models on Electronic Health Records (EHRs) provides a means of learning representations that can transfer to downstream tasks with minimal supervision. Recent multimodal models induce soft local alignments between image regions and sentences. This is of particular interest in the medical domain, where alignments might highlight regions in an image relevant to specific phenomena described in free-text. While past work has suggested that attention "heatmaps" can be interpreted in this manner, there has been little evaluation of such alignments. We compare alignments from a state-of-the-art multimodal (image and text) model for EHR with human annotations that link image regions to sentences. Our main finding is that the text has an often weak or unintuitive influence on attention; alignments do not consistently reflect basic anatomical information. Moreover, synthetic modifications — such as substituting "left" for "right" — do not substantially influence highlights. Simple techniques such as allowing the model to opt out of attending to the image and few-shot finetuning show promise in terms of their ability to improve alignments with very little or no supervision. We make our code and checkpoints open-source.

**Adversarial Concept Erasure in Kernel Space**
*Shauli Ravfogel, Francisco Vargas, Yoav Goldberg and Ryan Cotterell* 16:00-17:30 (Atrium)
The representation space of neural models for textual data emerges in an unsupervised manner during training. Understanding how human-interpretable concepts, such as gender, are encoded in these representations would improve the ability of users to control the content of these representations and analyze the working of the models that rely on them. One prominent approach to the control problem is the identification and removal of linear concept subspaces – subspaces in the representation space that correspond to a given concept. While those are tractable and interpretable, neural network do not necessarily represent concepts in linear subspaces.
We propose a kernelization of the recently-proposed linear concept-removal objective, and show that it is effective in guarding against the ability of certain nonlinear adversaries to recover the concept. Interestingly, our findings suggest that the division between linear and nonlinear models is overly simplistic: when considering the concept of binary gender and its neutralization, we do not find a single kernel space that exclusively contains all the concept-related information. It is therefore challenging to protect against all nonlinear adversaries at once.

**One size does not fit all: Investigating strategies for differentially-private learning across NLP tasks**
*Manuel Senge, Timour Igamberdiev and Ivan Habernal* 16:00-17:30 (Atrium)
Preserving privacy in contemporary NLP models allows us to work with sensitive data, but unfortunately comes at a price. We know that stricter privacy guarantees in differentially-private stochastic gradient descent (DP-SGD) generally degrade model performance. However, previous research on the efficiency of DP-SGD in NLP is inconclusive or even counter-intuitive. In this short paper, we provide an extensive analysis of different privacy preserving strategies on seven downstream datasets in five different 'typical' NLP tasks with varying complexity using modern neural models based on BERT and XtremeDistil architectures. We show that unlike standard non-private approaches to solving NLP tasks, where bigger is usually better, privacy-preserving strategies do not exhibit a winning pattern, and each task and privacy regime requires a special treatment to achieve adequate performance.

**Towards Teachable Reasoning Systems: Using a Dynamic Memory of User Feedback for Continual System Improvement**
*Bhavana Dalvi Mishra, Oyvind Tafjord and Peter Clark* 16:00-17:30 (Atrium)
Our goal is a teachable reasoning system for question-answering (QA), where a user can interact with faithful answer explanations, and correct its errors so that the system improves over time. Our approach is to augment a QA model with a dynamic memory of user feedback, containing user-supplied corrections to erroneous model beliefs that users identify during interaction. Retrievals from memory are used as additional context for QA, to help avoid previous mistakes in similar new situations - a novel application of memory-based continuous learning. With simulated feedback, we find that our system (called TeachMe) continually improves with time, and without model retraining, requiring feedback on only 25% of training examples to reach within 1% of the upper-bound (feedback on all examples). Similarly, in experiments with real users, we observe a similar trend, with performance improving by over 15% on a hidden test set after teaching. This suggests new opportunities for using frozen language models in an interactive setting where users can inspect, debug, and correct the model's beliefs, leading to improved system's performance over time.

**Mixed-effects transformers for hierarchical adaptation**
*Julia White, Noah Goodman and Robert Hawkins* 16:00-17:30 (Atrium)
Language differs dramatically from context to context. To some degree, large language models like GPT-3 account for such variation by

conditioning on strings of initial input text, or prompts. However, prompting can be ineffective when contexts are sparse, out-of-sample, or extra-textual. In this paper, we introduce the mixed-effects transformer (MET), a novel approach for learning hierarchically-structured prefixes— lightweight modules prepended to an input sequence— to account for structured variation in language use. Specifically, we show how the popular class of mixed-effects regression models may be extended to transformer-based architectures using a regularized prefix-tuning procedure with dropout. We evaluate this approach on several domain-adaptation benchmarks, finding that it learns contextual variation from minimal data while generalizing well to unseen contexts.

**Adapting a Language Model While Preserving its General Knowledge**
*Zixuan Ke, Yijia Shao, Haowei Lin, Hu Xu, Lei Shu and Bing Liu* 16:00-17:30 (Atrium)
Domain-adaptive pre-training (or DA-training for short), also known as post-training, aims to train a pre-trained general-purpose language model (LM) using an unlabeled corpus of a particular domain to adapt the LM so that end-tasks in the domain can give improved performances. However, existing DA-training methods are in some sense blind as they do not explicitly identify what knowledge in the LM should be preserved and what should be changed by the domain corpus. This paper shows that the existing methods are suboptimal and proposes a novel method to perform a more informed adaptation of the knowledge in the LM by (1) soft-masking the attention heads based on their importance to best preserve the general knowledge in the LM and (2) contrasting the representations of the general and the full (both general and domain knowledge) to learn an integrated representation with both general and domain-specific knowledge. Experimental results will demonstrate the effectiveness of the proposed approach.

**Towards Opening the Black Box of Neural Machine Translation: Source and Target Interpretations of the Transformer**
*Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano and Marta R. Costa-jussà* 16:00-17:30 (Atrium)
In Neural Machine Translation (NMT), each token prediction is conditioned on the source sentence and the target prefix (what has been previously translated at a decoding step). However, previous work on interpretability in NMT has mainly focused solely on source sentence tokens' attributions. Therefore, we lack a full understanding of the influences of every input token (source sentence and target prefix) in the model predictions. In this work, we propose an interpretability method that tracks input tokens' attributions for both contexts. Our method, which can be extended to any encoder-decoder Transformer-based model, allows us to better comprehend the inner workings of current NMT models. We apply the proposed method to both bilingual and multilingual Transformers and present insights into their behaviour.

**Polyglot Prompt: Multilingual Multitask Prompt Training**
*Jinlan Fu, See-Kiong Ng and Pengfei Liu* 16:00-17:30 (Atrium)
This paper aims for a potential architectural improvement for multilingual learning and asks: Can different tasks from different languages be modeled in a monolithic framework, i.e. without any task/language-specific module? The benefit of achieving this could open new doors for future multilingual research, including allowing systems trained on low resources to be further assisted by other languages as well as other tasks. We approach this goal by developing a learning framework named Polyglot Prompting to exploit prompting methods for learning a unified semantic space for different languages and tasks with multilingual prompt engineering. We performed a comprehensive evaluation of 6 tasks, namely topic classification, sentiment classification, named entity recognition, question answering, natural language inference, and summarization, covering 24 datasets and 49 languages. The experimental results demonstrated the efficacy of multilingual multitask prompt-based learning and led to inspiring observations. We also present an interpretable multilingual evaluation methodology and show how the proposed framework, multilingual multitask prompt training, works. We release all datasets prompted in the best setting and code.

**Context-Situated Pun Generation**
*Jiao Sun, Anjali Narayan-Chen, Shereen Oraby, Shuyang Gao, Tagyoung Chung, Jing Huang, Yang Liu and Nanyun Peng* 16:00-17:30 (Atrium)
Previous work on pun generation commonly begins with a given pun word (a pair of homophones for heterographic pun generation and a polyseme for homographic pun generation) and seeks to generate an appropriate pun. While this may enable efficient pun generation, we believe that a pun is most entertaining if it fits appropriately within a given context, e.g., a given situation or dialogue. In this work, we propose a new task, context-situated pun generation, where a specific context represented by a set of keywords is provided, and the task is to first identify suitable pun words that are appropriate for the context, then generate puns based on the context keywords and the identified pun words. We collect a new dataset, CUP (Context-sitUated Pun), containing 4.5k tuples of context words and pun pairs. Based on the new data and setup, we propose a pipeline system for context-situated pun generation, including a pun word retrieval module that identifies suitable pun words for a given context, and a pun generation module that generates puns from context keywords and pun words. Human evaluation shows that 69% of our top retrieved pun words can be used to generate context-situated puns, and our generation module yields successful puns 31% of the time given a plausible tuple of context words and pun pair, almost tripling the yield of a state-of-the-art pun generation model. With an end-to-end evaluation, our pipeline system with the top-1 retrieved pun pair for a given context can generate successful puns 40% of the time, better than all other modeling variations but 32% lower than the human success rate. This highlights the difficulty of the task, and encourages more research in this direction.

**Twist Decoding: Diverse Generators Guide Each Other**
*Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Hao Peng, Ximing Lu, Dragomir Radev, Yejin Choi and Noah A. Smith* 16:00-17:30 (Atrium)
Many language generation models are now available for a wide range of generation tasks, including machine translation and summarization. Combining such diverse models may lead to further progress, but ensembling generation models is challenging during inference: conventional ensembling methods (e.g., shallow fusion) require that the models share vocabulary/tokenization schemes. We introduce Twist decoding, a simple and general text generation algorithm that benefits from diverse models at inference time. Our method does not assume the vocabulary, tokenization or even generation order is shared. Our extensive evaluations on machine translation and scientific paper summarization demonstrate that Twist decoding substantially outperforms each model decoded in isolation over various scenarios, including cases where domain-specific and general-purpose models are both available. Twist decoding also consistently outperforms the popular reranking heuristic where output candidates from one model are rescored by another. We hope that our work will encourage researchers and practitioners to examine generation models collectively, not just independently, and to seek out models with complementary strengths to the currently available models.

**T-STAR: Truthful Style Transfer using AMR Graph as Intermediate Representation**
*Anubhav Jangra, Preksha Nema and Aravindan Raghuveer* 16:00-17:30 (Atrium)
Unavailability of parallel corpora for training text style transfer (TST) models is a very challenging yet common scenario. Also, TST models implicitly need to preserve the content while transforming a source sentence into the target style. To tackle these problems, an intermediate representation is often constructed that is devoid of style while still preserving the meaning of the source sentence. In this work, we study the usefulness of Abstract Meaning Representation (AMR) graph as the intermediate style agnostic representation. We posit that semantic notations like AMR are a natural choice for an intermediate representation. Hence, we propose T-STAR: a model comprising of two components, text-to-AMR encoder and a AMR-to-text decoder. We propose several modeling improvements to enhance the style agnosticity of the generated AMR. To the best of our knowledge, T-STAR is the first work that uses AMR as an intermediate representation for TST. With

thorough experimental evaluation we show T-STAR significantly outperforms state of the art techniques by achieving on an average 15.2% higher content preservation with negligible loss ( 3%) in style accuracy. Through detailed human evaluation with 90,000 ratings, we also show that T-STAR has upto 50% lesser hallucinations compared to state of the art TST models.

**LILA: A Unified Benchmark for Mathematical Reasoning**
*Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark and Ashwin Kalyan* 16:00-17:30 (Atrium)
Mathematical reasoning skills are essential for general-purpose intelligent systems to perform tasks from grocery shopping to climate modeling. Towards evaluating and improving AI systems in this domain, we propose LILA, a unified mathematical reasoning benchmark consisting of 23 diverse tasks along four dimensions: (i) mathematical abilities e.g., arithmetic, calculus (ii) language format e.g., question-answering, fill-in-the-blanks (iii) language diversity e.g., no language, simple language (iv) external knowledge e.g., commonsense, physics. We construct our benchmark by extending 20 datasets benchmark by collecting task instructions and solutions in the form of Python programs, thereby obtaining explainable solutions in addition to the correct answer. We additionally introduce two evaluation datasets to measure out-of-distribution performance and robustness to language perturbation. Finally, we introduce BHASKARA, a general-purpose mathematical reasoning model trained on LILA. Importantly, we find that multi-tasking leads to significant improvements (average relative improvement of 21.83% F1 score vs. single-task models), while the best performing model only obtains 60.40%, indicating the room for improvement in general mathematical reasoning and understanding.

**Character-centric Story Visualization via Visual Planning and Token Alignment**
*Hong Chen, Rujun Han, Te-Lin Wu, Hideki Nakayama and Nanyun Peng* 16:00-17:30 (Atrium)
Story visualization advances the traditional text-to-image generation by enabling multiple image generation based on a complete story. This task requires machines to 1) understand long text inputs, and 2) produce a globally consistent image sequence that illustrates the contents of the story. A key challenge of consistent story visualization is to preserve characters that are essential in stories. To tackle the challenge, we propose to adapt a recent work that augments VQ-VAE with a text-to-visual-token (transformer) architecture. Specifically, we modify the text-to-visual-token module with a two-stage framework: 1) character token planning model that predicts the visual tokens for characters only; 2) visual token completion model that generates the remaining visual token sequence, which is sent to VQ-VAE for finalizing image generations. To encourage characters to appear in the images, we further train the two-stage framework with a character-token alignment objective. Extensive experiments and evaluations demonstrate that the proposed method excels at preserving characters and can produce higher quality image sequences compared with the strong baselines.

**Algorithms for Acyclic Weighted Finite-State Automata with Failure Arcs**
*Anej Svete, Benjamin Dayan, Ryan Cotterell, Tim Vieira and Jason Eisner* 16:00-17:30 (Atrium)
Weighted finite-state automata (WSFAs) are commonly used in NLP. Failure transitions are a useful extension for compactly representing backoffs or interpolation in n-gram models and CRFs, which are special cases of WFSAs. Unfortunately, applying standard algorithms for computing the pathsum requires expand- ing these compact failure transitions. As a result, naive computation of the pathsum in acyclic WFSAs with failure transitions runs in $O(|Q|2|\Sigma|)$ ($O(|Q||\Sigma|)$ for deterministic WF- SAs) while the equivalent algorithm in normal WFSAs runs in O(|El), where E represents the set of transitions, Q the set of states, and $\Sigma$ the alphabet. In this work, we present more efficient algorithms for computing the pathsum in sparse acyclic WFSAs, i.e., WFSAs with av- erage out symbol fraction $s \ll 1$. In those, backward runs in $O(s|Q||\Sigma|)$. We propose an algorithm for semiring-weighted automata which runs in $O(|E| + s|\Sigma||Q||Tmax|log|\Sigma|)$, where |Tmax| is the size of the largest con- nected component of failure transitions. Ad- ditionally, we propose faster algorithms for two specific cases. For ring-weighted WF- SAs we propose an algorithm with complex- ity $O(|E| + s|\Sigma||Q||\pi max|)$, where $|\pi max|$ denotes the longest path length of failure transi- tions stemming from q and $\Sigma(q)$ the set of sym- bols on the outgoing transitions from q. For semiring-weighted WFSAs whose failure tran- sition topology satisfies a condition exemplified by CRFs, we propose an algorithm with complexity $O(|E| + s|\Sigma||Q|log|\Sigma|)$.

**The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation**
*Barbara Plank* 16:00-17:30 (Atrium)
Human variation in labeling is often considered noise. Annotation projects for machine learning (ML) aim at minimizing human label variation, with the assumption to maximize data quality and in turn optimize and maximize machine learning metrics. However, this conventional practice assumes that there exists a *ground truth*, and neglects that there exists genuine human variation in labeling due to disagreement, subjectivity in annotation or multiple plausible answers. In this position paper, we argue that this big open problem of *human label variation* persists and critically needs more attention to move our field forward. This is because human label variation impacts all stages of the ML pipeline: *data, modeling and evaluation*. However, few works consider all of these dimensions jointly; and existing research is fragmented. We reconcile different previously proposed notions of human label variation, provide a repository of publicly-available datasets with un-aggregated labels, depict approaches proposed so far, identify gaps and suggest ways forward. As datasets are becoming increasingly available, we hope that this synthesized view on the "problem" will lead to an open discussion on possible strategies to devise fundamentally new directions.

# Main Conference: Saturday, December 10, 2022

# Session 6 - 09:00-10:30

## Dialog and Interactive Systems 1

09:00-10:30 (Hall A, Room A)

**CDConv: A Benchmark for Contradiction Detection in Chinese Conversations**
*Chujie Zheng, Jinfeng Zhou, Yinhe Zheng, Libiao Peng, Zhen Guo, Wenquan Wu, Zheng-Yu Niu, Hua Wu and Minlie Huang* 09:00-09:15 (Hall A, Room A)
Dialogue contradiction is a critical issue in open-domain dialogue systems. The contextualization nature of conversations makes dialogue contradiction detection rather challenging. In this work, we propose a benchmark for Contradiction Detection in Chinese Conversations, namely CDConv. It contains 12K multi-turn conversations annotated with three typical contradiction categories: Intra-sentence Contradiction, Role Confusion, and History Contradiction. To efficiently construct the CDConv conversations, we devise a series of methods for automatic conversation generation, which simulate common user behaviors that trigger chatbots to make contradictions. We conduct careful manual quality screening of the constructed conversations and show that state-of-the-art Chinese chatbots can be easily goaded into making contradictions. Experiments on CDConv show that properly modeling contextual information is critical for dialogue contradiction detection, but there are still unresolved challenges that require future research.

**Co-guiding Net: Achieving Mutual Guidances between Multiple Intent Detection and Slot Filling via Heterogeneous Semantics-Label Graphs**
*Bowen Xing and Ivor Tsang* 09:15-09:30 (Hall A, Room A)
Recent graph-based models for joint multiple intent detection and slot filling have obtained promising results through modeling the guidance from the prediction of intents to the decoding of slot filling. However, existing methods (1) only model the *unidirectional guidance* from intent to slot; (2) adopt *homogeneous graphs* to model the interactions between the slot semantics nodes and intent label nodes, which limit the performance. In this paper, we propose a novel model termed Co-guiding Net, which implements a two-stage framework achieving the *mutual guidances* between the two tasks. In the first stage, the initial estimated labels of both tasks are produced, and then they are leveraged in the second stage to model the mutual guidances. Specifically, we propose two *heterogeneous graph attention networks* working on the proposed two *heterogeneous semantics-label graphs*, which effectively represent the relations among the semantics nodes and label nodes. Experiment results show that our model outperforms existing models by a large margin, obtaining a relative improvement of 19.3

**Estimating Soft Labels for Out-of-Domain Intent Detection**
*Hao Lang, Yinhe Zheng, Jian Sun, Fei Huang, Luo Si and Yongbin Li* 09:30-09:45 (Hall A, Room A)
Out-of-Domain (OOD) intent detection is important for practical dialog systems. To alleviate the issue of lacking OOD training samples, some works propose synthesizing pseudo OOD samples and directly assigning one-hot OOD labels to these pseudo samples. However, these one-hot labels introduce noises to the training process because some "hard" pseudo OOD samples may coincide with In-Domain (IND) intents. In this paper, we propose an adaptive soft pseudo labeling (ASoul) method that can estimate soft labels for pseudo OOD samples when training OOD detectors. Semantic connections between pseudo OOD samples and IND intents are captured using an embedding graph. A co-training framework is further introduced to produce resulting soft labels following the smoothness assumption, i.e., close samples are likely to have similar labels. Extensive experiments on three benchmark datasets show that ASoul consistently improves the OOD detection performance and outperforms various competitive baselines.

**InstructDial: Improving Zero and Few-shot Generalization in Dialogue through Instruction Tuning**
*Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi and Jeffrey Bigham* 09:45-10:00 (Hall A, Room A)
Instruction tuning is an emergent paradigm in NLP wherein natural language instructions are leveraged with language models to induce zero-shot performance on unseen tasks. Dialogue is an especially interesting area in which to explore instruction tuning because dialogue systems perform multiple kinds of tasks related to language (e.g., natural language understanding and generation, domain-specific interaction), yet instruction tuning has not been systematically explored for dialogue-related tasks. We introduce InstructDial, an instruction tuning framework for dialogue, which consists of a repository of 48 diverse dialogue tasks in a unified text-to-text format created from 59 openly available dialogue datasets. We explore cross-task generalization ability on models tuned on InstructDial across diverse dialogue tasks. Our analysis reveals that InstructDial enables good zero-shot performance on unseen datasets and tasks such as dialogue evaluation and intent detection, and even better performance in a few-shot setting. To ensure that models adhere to instructions, we introduce novel meta-tasks. We establish benchmark zero-shot and few-shot performance of models trained using the proposed framework on multiple dialogue tasks.

**Aligning Recommendation and Conversation via Dual Imitation**
*Jinfeng Zhou, Bo Wang, Minlie Huang, Dongming Zhao, Kun Huang, Ruifang He and Yuexian Hou* 10:00-10:15 (Hall A, Room A)
Human conversations of recommendation naturally involve the shift of interests which can align the recommendation actions and conversation process to make accurate recommendations with rich explanations. However, existing conversational recommendation systems (CRS) ignore the advantage of user interest shift in connecting recommendation and conversation, which leads to an ineffective loose coupling structure of CRS. To address this issue, by modeling the recommendation actions as recommendation paths in a knowledge graph (KG), we propose DICR (**D**ual **I**mitation for **C**onversational **R**ecommendation), which designs a dual imitation to explicitly align the recommendation paths and user interest shift paths in a recommendation module and a conversation module, respectively. By exchanging alignment signals, DICR achieves bidirectional promotion between recommendation and conversation modules and generates high-quality responses with accurate recommendations and coherent explanations. Experiments demonstrate that DICR outperforms the state-of-the-art models on recommendation and conversation performance with automatic, human, and novel explainability metrics.

**Correctable-DST: Mitigating Historical Context Mismatch between Training and Inference for Improved Dialogue State Tracking**
*Hongyan Xie, Haoxiang Su, Shuangyong Song, Hao Huang, Bo Zou, Kun Deng, Jianghua Lin, Zhihui Zhang and Xiaodong He* 10:15-10:30 (Hall A, Room A)
Recently proposed dialogue state tracking (DST) approaches predict the dialogue state of a target turn sequentially based on the previous dialogue state. During the training time, the ground-truth previous dialogue state is utilized as the historical context. However, only the previously predicted dialogue state can be used in inference. This discrepancy might lead to error propagation, i.e., mistakes made by the model in the current turn are likely to be carried over to the following turns. To solve this problem, we propose Correctable Dialogue State Tracking

(Correctable-DST). Specifically, it consists of three stages: (1) a Predictive State Simulator is exploited to generate a previously "predicted" dialogue state based on the ground-truth previous dialogue state during training; (2) a Slot Detector is proposed to determine the slots with an incorrect value in the previously "predicted" state and the slots whose values are to be updated in the current turn; (3) a State Generator takes the name of the above-selected slots as a prompt to generate the current state. Empirical results show that our approach achieves 67.51%, 68.24%, 70.30%, 71.38%, and 81.27% joint goal accuracy on MultiWOZ 2.0-2.4 datasets, respectively, and achieves a new state-of-the-art performance with significant improvements.

## Multilinguality

09:00-10:30 (Hall A, Room B)

**Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset**
*Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen and Radu Soricut* 09:00-09:15 (Hall A, Room B)
Research in massively multilingual image captioning has been severely hampered by a lack of high-quality evaluation datasets. In this paper we present the Crossmodal-3600 dataset (XM3600 in short), a geographically diverse set of 3600 images annotated with human-generated reference captions in 36 languages. The images were selected from across the world, covering regions where the 36 languages are spoken, and annotated with captions that achieve consistency in terms of style across all languages, while avoiding annotation artifacts due to direct translation. We apply this benchmark to model selection for massively multilingual image captioning models, and show superior correlation results with human evaluations when using XM3600 as golden references for automatic metrics.

**Graph-Based Multilingual Label Propagation for Low-Resource Part-of-Speech Tagging**
*Ayyoob ImaniGooghari, Silvia Severini, Masoud Jalili Sabet, François Yvon and Hinrich Schütze* 09:15-09:30 (Hall A, Room B)
Part-of-Speech (POS) tagging is an important component of the NLP pipeline, but many low-resource languages lack labeled data for training. An established method for training a POS tagger in such a scenario is to create a labeled training set by transferring from high-resource languages. In this paper, we propose a novel method for transferring labels from multiple high-resource source to low-resource target languages. We formalize POS tag projection as graph-based label propagation. Given translations of a sentence in multiple languages, we create a graph with words as nodes and alignment links as edges by aligning words for all language pairs. We then propagate node labels from source to target using a Graph Neural Network augmented with transformer layers. We show that our propagation creates training sets that allow us to train POS taggers for a diverse set of languages. When combined with enhanced contextualized embeddings, our method achieves a new state-of-the-art for unsupervised POS tagging of low-resource languages.

**AfroLID: A Neural Language Identification Tool for African Languages**
*Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed and Alcides Inciarte* 09:30-09:45 (Hall A, Room B)
Language identification (LID) is a crucial precursor for NLP, especially for mining web data. Problematically, most of the world's 7000+ languages today are not covered by LID technologies. We address this pressing issue for Africa by introducing AfroLID, a neural LID toolkit for 517 African languages and varieties. AfroLID exploits a multi-domain web dataset manually curated from across 14 language families utilizing five orthographic systems. When evaluated on our blind Test set, AfroLID achieves 95.89 F_1-score. We also compare AfroLID to five existing LID tools that each cover a small number of African languages, finding it to outperform them on most languages. We further show the utility of AfroLID in the wild by testing it on the acutely under-served Twitter domain. Finally, we offer a number of controlled case studies and perform a linguistically-motivated error analysis that allow us to both showcase AfroLID's powerful capabilities and limitations[3]

**The (Undesired) Attenuation of Human Biases by Multilinguality**
*Cristina España-Bonet and Alberto Barrón-Cedeño* 09:45-10:00 (Hall A, Room B)
Some human preferences are universal. The odor of vanilla is perceived as pleasant all around the world. We expect neural models trained on human texts to exhibit these kind of preferences, i.e. biases, but we show that this is not always the case. We explore 16 static and contextual embedding models in 9 languages and, when possible, compare them under similar training conditions. We introduce and release CA-WEAT, multilingual cultural aware tests to quantify biases, and compare them to previous English-centric tests. Our experiments confirm that monolingual static embeddings do exhibit human biases, but values differ across languages, being far from universal. Biases are less evident in contextual models, to the point that the original human association might be reversed. Multilinguality proves to be another variable that attenuates and even reverses the effect of the bias, specially in contextual multilingual models. In order to explain this variance among models and languages, we examine the effect of asymmetries in the training corpus, departures from isomorphism in multilingual embedding spaces and discrepancies in the testing measures between languages.

**CoCoa: An Encoder-Decoder Model for Controllable Code-switched Generation**
*Sneha Mondal, Ritika ., Shreya Pathak, Preethi Jyothi and Aravindan Raghuveer* 10:00-10:15 (Hall A, Room B)
Code-switching has seen growing interest in recent years as an important multilingual NLP phenomenon. Generating code-switched text for data augmentation has been sufficiently well-explored. However, there is no prior work on generating code-switched text with fine-grained control on the degree of code-switching and the lexical choices used to convey formality. We present CoCoa, an encoder-decoder translation model that converts monolingual Hindi text to Hindi-English code-switched text with both encoder-side and decoder-side interventions to achieve fine-grained controllable generation. CoCoa can be invoked at test-time to synthesize code-switched text that is simultaneously faithful to syntactic and lexical attributes relevant to code-switching. CoCoa outputs were subjected to rigorous subjective and objective evaluations. Human evaluations establish that our outputs are of superior quality while being faithful to desired attributes. We show significantly improved BLEU scores when compared with human-generated code-switched references. Compared to competitive baselines, we show 10% reduction in perplexity on a language modeling task and also demonstrate clear improvements on a downstream code-switched sentiment analysis task.

**Calibrating Zero-shot Cross-lingual (Un-)structured Predictions**
*Zhengping Jiang, Anqi Liu and Benjamin Van Durme* 10:15-10:30 (Hall A, Room B)
We investigate model calibration in the setting of zero-shot cross-lingual transfer with large-scale pre-trained language models. The level of model calibration is an important metric for evaluating the trustworthiness of predictive models. There exists an essential need for model calibration when natural language models are deployed in critical tasks. We study different post-training calibration methods in structured and unstructured prediction tasks. We find that models trained with data from the source language become less calibrated when applied to the target language and that calibration errors increase with intrinsic task difficulty and relative sparsity of training data. Moreover, we observe a

[3] AfroLID is publicly available at https://github.com/UBC-NLP/afrolid.

potential connection between the level of calibration error and an earlier proposed measure of the distance from English to other languages. Finally, our comparison demonstrates that among other methods Temperature Scaling (TS) generalizes well to distant languages, but TS fails to calibrate more complex confidence estimation in structured predictions compared to more expressive alternatives like Gaussian Process Calibration.

## Natural Language Generation 2 & TACL

09:00-10:30 (Hall A, Room C)

**Visual Spatial Description: Controlled Spatial-Oriented Image-to-Text Generation**
*Yu Zhao, Jianguo Wei, ZhiChao Lin, Yueheng Sun, Meishan Zhang and Min Zhang* 09:00-09:15 (Hall A, Room C)
Image-to-text tasks such as open-ended image captioning and controllable image description have received extensive attention for decades. Here we advance this line of work further, presenting Visual Spatial Description (VSD), a new perspective for image-to-text toward spatial semantics. Given an image and two objects inside it, VSD aims to produce one description focusing on the spatial perspective between the two objects. Accordingly, we annotate a dataset manually to facilitate the investigation of the newly-introduced task, and then build several benchmark encoder-decoder models by using VL-BART and VL-T5 as backbones. In addition, we investigate visual spatial relationship classification (VSRC) information into our model by pipeline and end-to-end architectures. Finally, we conduct experiments on our benchmark dataset to evaluate all our models. Results show that our models are awe-inspiring, offering accurate and human-like spatial-oriented text descriptions. Besides, VSRC has great potential for VSD, and the joint end-to-end architecture is the better choice for their integration. We will make the dataset and codes publicly available for research purposes.

**SubeventWriter: Iterative Sub-event Sequence Generation with Coherence Controller**
*Zhaowei Wang, Hongming Zhang, Tianqing Fang, Yangqiu Song, Ginny Wong and Simon See* 09:15-09:30 (Hall A, Room C)
In this paper, we propose a new task of sub-event generation for an unseen process to evaluate the understanding of the coherence of sub-event actions and objects. To solve the problem, we design SubeventWriter, a sub-event sequence generation framework with a coherence controller. Given an unseen process, the framework can iteratively construct the sub-event sequence by generating one sub-event at each iteration. We also design a very effective coherence controller to decode more coherent sub-events. As our extensive experiments and analysis indicate, SubeventWriter can generate more reliable and meaningful sub-event sequences for unseen processes.

**Towards a Unified Multi-Dimensional Evaluator for Text Generation**
*Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji and Jiawei Han*09:30-09:45 (Hall A, Room C)
Multi-dimensional evaluation is the dominant paradigm for human evaluation in Natural Language Generation (NLG), i.e., evaluating the generated text from multiple explainable dimensions, such as coherence and fluency. However, automatic evaluation in NLG is still dominated by similarity-based metrics, and we lack a reliable framework for a more comprehensive evaluation of advanced models. In this paper, we propose a unified multi-dimensional evaluator UniEval for NLG. We re-frame NLG evaluation as a Boolean Question Answering (QA) task, and by guiding the model with different questions, we can use one evaluator to evaluate from multiple dimensions. Furthermore, thanks to the unified Boolean QA format, we are able to introduce an intermediate learning phase that enables UniEval to incorporate external knowledge from multiple related tasks and gain further improvement. Experiments on three typical NLG tasks show that UniEval correlates substantially better with human judgments than existing metrics. Specifically, compared to the top-performing unified evaluators, UniEval achieves a 23% higher correlation on text summarization, and over 43% on dialogue response generation. Also, UniEval demonstrates a strong zero-shot learning ability for unseen evaluation dimensions and tasks. Source code, data, and all pre-trained evaluators are available at https://github.com/maszhongming/UniEval.

**Prompt-and-Rerank: A Method for Zero-Shot and Few-Shot Arbitrary Textual Style Transfer with Small Language Models**
*Mirac Suzgun, Luke Melas-Kyriazi and Dan Jurafsky* 09:45-10:00 (Hall A, Room C)
We propose a method for arbitrary textual style transfer (TST)—the task of transforming a text into any given style—utilizing general-purpose pre-trained language models. Our method, Prompt-and-Rerank, is based on a mathematical formulation of the TST task, decomposing it into three constituent components: textual similarity, target style strength, and fluency. Our method uses zero-shot or few-shot prompting to obtain a set of candidate generations in the target style, and then re-ranks them according to the three components. Our method enables small pre-trained language models to perform on par with state-of-the-art large-scale models while using two orders of magnitude less compute and memory. We also investigate the effect of model size and prompt design (e.g., prompt paraphrasing and delimiter-pair choice) on style transfer quality across seven diverse textual style transfer datasets, finding, among other things, that delimiter-pair choice has a large impact on performance, and that models have biases on the direction of style transfer.

**Gradient-based Constrained Sampling from Language Models**
*Sachin Kumar, Biswajit Paria and Yulia Tsvetkov* 10:00-10:15 (Hall A, Room C)
Large pretrained language models are successful at generating fluent text but are notoriously hard to controllably sample from. In this work, we study constrained sampling from such language models, i.e., generating text that satisfies user-defined constraints, while maintaining fluency and model's performance in a downstream task. We propose MuCoLa—a sampling procedure that combines the log-likelihood of the language model with arbitrary (differentiable) constraints in a single energy function, and then generates samples in a non-autoregressive manner. Specifically, it initializes the entire output sequence with noise and follows a Markov chain defined by Langevin Dynamics using the gradients of this energy. We evaluate MuCoLa on text generation with soft and hard constraints as well as their combinations, obtaining significant improvements over competitive baselines for toxicity avoidance, sentiment control, and keyword-guided generation.

[TACL] **Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open-Domain Question-Answering**
*Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana and Suranga Nanayakkara*10:15-10:30 (Hall A, Room C)
Retrieval Augment Generation (RAG) is a recent advancement in Open Domain Ques tion Answering (ODQA). RAG has only been trained and explored with a Wikipedia- based external knowledge base and not opti mized for use in other specialized domains such as healthcare and news. In this pa per, we propose a novel approach to adapt RAG for domain-specific question answer ing. For this, we present RAG-end2end , an extension to RAG, that can adapt to a domain-specific knowledge base by updating all components of the external knowledge base during training. In addition, we intro duce an auxiliary training signal to infuse more domain knowledge. This auxiliary signal forces RAG-end2end to reconstruct a given sentence by accessing the relevant information from the external knowledge base. We evaluate our approach with datasets from three domains: COVID-19, News, and Conversations, and achieve significant perfor mance improvements compared to the orig inal RAG model. Our work, RAG-end2end , has been open-sourced through the Hugging face Transformers library.

## Efficient Methods for NLP

09:00-10:30 (Hall A, Room D)

**Inducer-tuning: Connecting Prefix-tuning and Adapter-tuning**
*Yifan Chen, Devamanyu Hazarika, Mahdi Namazifar, Yang Liu, Di Jin and Dilek Hakkani-Tur* 09:00-09:15 (Hall A, Room D)
Prefix-tuning, or more generally continuous prompt tuning, has become an essential paradigm of parameter-efficient transfer learning. Using a large pre-trained language model (PLM), prefix-tuning can obtain strong performance by training only a small portion of parameters. In this paper, we propose to understand and further develop prefix-tuning through the kernel lens. Specifically, we make an analogy between *prefixes* and *inducing variables* in kernel methods and hypothesize that *prefixes* serving as *inducing variables* would improve their overall mechanism. From the kernel estimator perspective, we suggest a new variant of prefix-tuning—*inducer-tuning*, which shares the exact mechanism as prefix-tuning while leveraging the residual form found in adapter-tuning. This mitigates the initialization issue in prefix-tuning. Through comprehensive empirical experiments on natural language understanding and generation tasks, we demonstrate that inducer-tuning can close the performance gap between prefix-tuning and fine-tuning.

**LightEA: A Scalable, Robust, and Interpretable Entity Alignment Framework via Three-view Label Propagation**
*Xin Mao, wenting wang, Yuanbin Wu and Man Lan* 09:15-09:30 (Hall A, Room D)
Entity Alignment (EA) aims to find equivalent entity pairs between KGs, which is the core step to bridging and integrating multi-source KGs. In this paper, we argue that existing complex EA methods inevitably inherit the inborn defects from their neural network lineage: poor interpretability and weak scalability. Inspired by recent studies, we reinvent the classical Label Propagation algorithm to effectively run on KGs and propose a neural-free EA framework — LightEA, consisting of three efficient components: (i) Random Orthogonal Label Generation, (ii) Three-view Label Propagation, and (iii) Sparse Sinkhorn Operation. According to the extensive experiments on public datasets, LightEA has impressive scalability, robustness, and interpretability. With a mere tenth of time consumption, LightEA achieves comparable results to state-of-the-art methods across all datasets and even surpasses them on many. Besides, due to the computational process of LightEA being entirely linear, we could trace the propagation process at each step and clearly explain how the entities are aligned.

**VIRT: Improving Representation-based Text Matching via Virtual Interaction**
*Dan Li, Yang Yang, Hongyin Tang, Jiahao Liu, Qifan Wang, Jingang Wang, Tong Xu, Wei Wu and Enhong Chen* 09:30-09:45 (Hall A, Room D)
Text matching is a fundamental research problem in natural language understanding. Interaction-based approaches treat the text pair as a single sequence and encode it through cross encoders, while representation-based models encode the text pair independently with siamese or dual encoders. Interaction-based models require dense computations and thus are impractical in real-world applications. Representation-based models have become the mainstream paradigm for efficient text matching. However, these models suffer from severe performance degradation due to the lack of interactions between the pair of texts. To remedy this, we propose a Virtual InteRacTion mechanism (VIRT) for improving representation-based text matching while maintaining its efficiency. In particular, we introduce an interactive knowledge distillation module that is only applied during training. It enables deep interaction between texts by effectively transferring knowledge from the interaction-based model. A light interaction strategy is designed to fully leverage the learned interactive knowledge. Experimental results on six text matching benchmarks demonstrate the superior performance of our method over several state-of-the-art representation-based models. We further show that VIRT can be integrated into existing methods as plugins to lift their performances.

**Learning Label Modular Prompts for Text Classification in the Wild**
*Hailin Chen, Amrita Saha, Shafiq Joty and Steven C.H. Hoi* 09:45-10:00 (Hall A, Room D)
Machine learning models usually assume i.i.d data during training and testing, but data and tasks in real world often change over time. To emulate the transient nature of real world, we propose a challenging but practical task: text classification in-the-wild, which introduces different non-stationary training/testing stages. Decomposing a complex task into modular components can enable robust generalisation under such non-stationary environment. However, current modular approaches in NLP do not take advantage of recent advances in parameter efficient tuning of pretrained language models. To close this gap, we propose ModularPrompt, a label-modular prompt tuning framework for text classification tasks. In ModularPrompt, the input prompt consists of a sequence of soft label prompts, each encoding modular knowledge related to the corresponding class label. In two of most formidable settings, ModularPrompt outperforms relevant baselines by a large margin demonstrating strong generalisation ability. We also conduct comprehensive analysis to validate whether the learned prompts satisfy properties of a modular representation.

**COST-EFF: Collaborative Optimization of Spatial and Temporal Efficiency with Slenderized Multi-exit Language Models**
*Bowen Shen, Zheng Lin, Yuanxin LIU, Zhengxiao Liu, Lei Wang and Weiping Wang* 10:00-10:15 (Hall A, Room D)
Transformer-based pre-trained language models (PLMs) mostly suffer from excessive overhead despite their advanced capacity. For resource-constrained devices, there is an urgent need for a spatially and temporally efficient model which retains the major capacity of PLMs. However, existing statically compressed models are unaware of the diverse complexities between input instances, potentially resulting in redundancy and inadequacy for simple and complex inputs. Also, miniature models with early exiting encounter challenges in the trade-off between making predictions and serving the deeper layers. Motivated by such considerations, we propose a collaborative optimization for PLMs that integrates static model compression and dynamic inference acceleration. Specifically, the PLM is slenderized in width while the depth remains intact, complementing layer-wise early exiting to speed up inference dynamically. To address the trade-off of early exiting, we propose a joint training approach that calibrates slenderization and preserves contributive structures to each exit instead of only the final layer. Experiments are conducted on GLUE benchmark and the results verify the Pareto optimality of our approach at high compression and acceleration rate with 1/8 parameters and 1/19 FLOPs of BERT.

**Training Dynamics for Curriculum Learning: A Study on Monolingual and Cross-lingual NLU**
*Fenia Christopoulou, Gerasimos Lampouras and Ignacio Iacobacci* 10:15-10:30 (Hall A, Room D)
Curriculum Learning (CL) is a technique of training models via ranking examples in a typically increasing difficulty trend with the aim of accelerating convergence and improving generalisability. Current approaches for Natural Language Understanding (NLU) tasks use CL to improve in-distribution data performance often via heuristic-oriented or task-agnostic difficulties. In this work, instead, we employ CL for NLU by taking advantage of training dynamics as difficulty metrics, i.e., statistics that measure the behavior of the model at hand on specific task-data instances during training and propose modifications of existing CL schedulers based on these statistics. Differently from existing works, we focus on evaluating models on in-distribution (ID), out-of-distribution (OOD) as well as zero-shot (ZS) cross-lingual transfer datasets. We show across several NLU tasks that CL with training dynamics can result in better performance mostly on zero-shot cross-lingual transfer and OOD settings with improvements up by 8.5% in certain cases. Overall, experiments indicate that training dynamics can lead to better performing models with smoother training compared to other difficulty metrics while being 20% faster on average. In addition, through analysis we shed light on the correlations of task-specific versus task-agnostic metrics.

## Information Retrieval and Text Mining

09:00-10:30 (Hall B)

**Certified Error Control of Candidate Set Pruning for Two-Stage Relevance Ranking**
*Minghan Li, Xinyu Zhang, Ji Xin, Hongyang Zhang and Jimmy Lin* 09:00-09:15 (Hall B)
In information retrieval (IR), candidate set pruning has been commonly used to speed up two-stage relevance ranking. However, such an approach lacks accurate error control and often trades accuracy against computational efficiency in an empirical fashion, missing theoretical guarantees. In this paper, we propose the concept of certified error control of candidate set pruning for relevance ranking, which means that the test error after pruning is guaranteed to be controlled under a user-specified threshold with high probability. Both in-domain and out-of-domain experiments show that our method successfully prunes the first-stage retrieved candidate sets to improve the second-stage reranking speed while satisfying the pre-specified accuracy constraints in both settings. For example, on MS MARCO Passage v1, our method reduces the average candidate set size from 1000 to 27, increasing reranking speed by about 37 times, while keeping MRR@10 greater than a pre-specified value of 0.38 with about 90% empirical coverage. In contrast, empirical baselines fail to meet such requirements. Code and data are available at: https://github.com/alexlimh/CEC-Ranking.

**RetroMAE: Pre-Training Retrieval-oriented Language Models Via Masked Auto-Encoder**
*Shitao Xiao, Zheng Liu, Yingxia Shao and Zhao Cao* 09:15-09:30 (Hall B)
Despite pre-training's progress in many important NLP tasks, it remains to explore effective pre-training strategies for dense retrieval. In this paper, we propose RetroMAE, a new retrieval oriented pre-training paradigm based on Masked Auto-Encoder (MAE). RetroMAE is highlighted by three critical designs. 1) A novel MAE workflow, where the input sentence is polluted for encoder and decoder with different masks. The sentence embedding is generated from the encoder's masked input; then, the original sentence is recovered based on the sentence embedding and the decoder's masked input via masked language modeling. 2) Asymmetric model structure, with a full-scale BERT like transformer as encoder, and a one-layer transformer as decoder. 3) Asymmetric masking ratios, with a moderate ratio for encoder: 15 30%, and an aggressive ratio for decoder: 50 70%. Our framework is simple to realize and empirically competitive: the pre-trained models dramatically improve the SOTA performances on a wide range of dense retrieval benchmarks, like BEIR and MS MARCO. The source code and pre-trained models are made publicly available at https://github.com/staoxiao/RetroMAE so as to inspire more interesting research.

**Efficient Document Retrieval by End-to-End Refining and Quantizing BERT Embedding with Contrastive Product Quantization**
*Zexuan Qiu, Qinliang Su, Jianxing Yu and Shijing Si* 09:30-09:45 (Hall B)
Efficient document retrieval heavily relies on the technique of semantic hashing, which learns a binary code for every document and employs Hamming distance to evaluate document distances. However, existing semantic hashing methods are mostly established on outdated TFIDF features, which obviously do not contain lots of important semantic information about documents. Furthermore, the Hamming distance can only be equal to one of several integer values, significantly limiting its representational ability for document distances. To address these issues, in this paper, we propose to leverage BERT embeddings to perform efficient retrieval based on the product quantization technique, which will assign for every document a real-valued codeword from the codebook, instead of a binary code as in semantic hashing. Specifically, we first transform the original BERT embeddings via a learnable mapping and feed the transformed embedding into a probabilistic product quantization module to output the assigned codeword. The refining and quantizing modules can be optimized in an end-to-end manner by minimizing the probabilistic contrastive loss. A mutual information maximization based method is further proposed to improve the representativeness of codewords, so that documents can be quantized more accurately. Extensive experiments conducted on three benchmarks demonstrate that our proposed method significantly outperforms current state-of-the-art baselines.

**Prompt-Based Meta-Learning For Few-shot Text Classification**
*Haoxing Zhang, Xiaofeng Zhang, Haibo Huang and Lei Yu* 09:45-10:00 (Hall B)
Few-shot Text Classification predicts the semantic label of a given text with a handful of supporting instances. Current meta-learning methods have achieved satisfying results in various few-shot situations. Still, they often require a large amount of data to construct many few-shot tasks for meta-training, which is not practical in real-world few-shot scenarios. Prompt-tuning has recently proved to be another effective few-shot learner by bridging the gap between pre-train and downstream tasks. In this work, we closely combine the two promising few-shot learning methodologies in structure and propose a Prompt-Based Meta-Learning (PBML) model to overcome the above meta-learning problem by adding the prompting mechanism. PBML assigns label word learning to base-learners and template learning to meta-learner, respectively. Experimental results show state-of-the-art performance on four text classification datasets under few-shot settings, with higher accuracy and good robustness. We demonstrate through low-resource experiments that our method alleviates the shortcoming that meta-learning requires too much data for meta-training. In the end, we use the visualization to interpret and verify that the meta-learning framework can help the prompting method converge better. We release our code to reproduce our experiments.

**Generative Multi-hop Retrieval**
*Hyunji Lee, Sohee Yang, Hanseok Oh and Minjoon Seo* 10:00-10:15 (Hall B)
A common practice for text retrieval is to use an encoder to map the documents and the query to a common vector space and perform a nearest neighbor search (NNS); multi-hop retrieval also often adopts the same paradigm, usually with a modification of iteratively reformulating the query vector so that it can retrieve different documents at each hop. However, such a bi-encoder approach has limitations in multi-hop settings; (1) the reformulated query gets longer as the number of hops increases, which further tightens the embedding bottleneck of the query vector, and (2) it is prone to error propagation. In this paper, we focus on alleviating these limitations in multi-hop settings by formulating the problem in a fully generative way. We propose an encoder-decoder model that performs multi-hop retrieval by simply generating the entire text sequences of the retrieval targets, which means the query and the documents interact in the language model's parametric space rather than L2 or inner product space as in the bi-encoder approach. Our approach, Generative Multi-hop Retrieval (GMR), consistently achieves comparable or higher performance than bi-encoder models in five datasets while demonstrating superior GPU memory and storage footprint.

**COCO-DR: Combating the Distribution Shift in Zero-Shot Dense Retrieval with Contrastive and Distributionally Robust Learning**
*Yue Yu, Chenyan Xiong, Si Sun, Chao Zhang and Arnold Overwijk* 10:15-10:30 (Hall B)
We present a new zero-shot dense retrieval (ZeroDR) method, COCO-DR, to improve the generalization ability of dense retrieval by combating the distribution shifts between source training tasks and target scenarios. To mitigate the impact of document differences, COCO-DR continues pretraining the language model on the target corpora to adapt the model to target distributions via COtinuous COtrastive learning. To prepare for unseen target queries, COCO-DR leverages implicit Distributionally Robust Optimization (iDRO) to reweight samples from different source query clusters for improving model robustness over rare queries during fine-tuning. COCO-DR achieves superior average performance on BEIR, the zero-shot retrieval benchmark. At BERT_Base scale, COCO-DR Base outperforms other ZeroDR models with 60x larger size. At BERT_Large scale, COCO-DR Large outperforms the giant GPT-3 embedding model which has 500x more parameters. Our analysis shows the correlation between COCO-DR's effectiveness in combating distribution shifts and improving zero-shot accuracy. Our

code and model can be found at https://github.com/OpenMatch/COCO-DR.

## Industry 2

09:00-10:30 (Collaboratorium)

[INDUSTRY] **Grafting Pre-trained Models for Multimodal Headline Generation**
*Lingfeng Qiao, Chen Wu, Ye Liu, haoyuan peng, di yin and Bo Ren* 09:00-09:15 (Collaboratorium)
Multimodal headline utilizes both video frames and transcripts to generate the natural language title of the videos. Due to a lack of large-scale, manually annotated data, the task of annotating grounded headlines for video is labor intensive and impractical. Previous researches on pre-trained language models and video-language models have achieved significant progress in related downstream tasks. However, none of them can be directly applied to multimodal headline architecture where we need both multimodal encoder and sentence decoder. A major challenge in simply gluing language model and video-language model is the modality balance, which is aimed at combining visual-language complementary abilities. In this paper, we propose a novel approach to graft the video encoder from the pre-trained video-language model on the generative pre-trained language model. We also present a consensus fusion mechanism for the integration of different components, via inter/intra modality relation. Empirically, experiments show that the grafted model achieves strong results on a brand-new dataset collected from real-world applications.

[INDUSTRY] **Named Entity Recognition in Industrial Tables using Tabular Language Models**
*Aneta Koleva, Martin Ringsquandl, Mark Buckley, Rakeb Hasan and Volker Tresp* 09:15-09:30 (Collaboratorium)
Specialized transformer-based models for encoding tabular data have gained interest in academia. Although tabular data is omnipresent in industry, applications of table transformers are still missing. In this paper, we study how these models can be applied to an industrial Named Entity Recognition (NER) problem where the entities are mentioned in tabular-structured spreadsheets. The highly technical nature of spreadsheets as well as the lack of labeled data present major challenges for fine-tuning transformer-based models. Therefore, we develop a dedicated table data augmentation strategy based on available domain-specific knowledge graphs. We show that this boosts performance in our low-resource scenario considerably. Further, we investigate the benefits of tabular structure as inductive bias compared to tables as linearized sequences. Our experiments confirm that a table transformer outperforms other baselines and that its tabular inductive bias is vital for convergence of transformer-based models.

[INDUSTRY] **Knowledge Distillation Transfer Sets and their Impact on Downstream NLU Tasks**
*Charith Peris, Lizhen Tan, Thomas Gueudre, Turan Gojayev, Pan Wei and Gokmen Oz* 09:30-09:45 (Collaboratorium)
Teacher-student knowledge distillation is a popular technique for compressing today's prevailing large language models into manageable sizes that fit low-latency downstream applications. Both the teacher and the choice of transfer set used for distillation are crucial ingredients in creating a high quality student. Yet, the generic corpora used to pretrain the teacher and the corpora associated with the downstream target domain are often significantly different, which raises a natural question: should the student be distilled over the generic corpora, so as to learn from high-quality teacher predictions, or over the downstream task corpora to align with finetuning? Our study investigates this trade-off using Domain Classification (DC) and Intent Classification/Named Entity Recognition (ICNER) as downstream tasks. We distill several multilingual students from a larger multilingual LM with varying proportions of generic and task-specific datasets, and report their performance after finetuning on DC and ICNER. We observe significant improvements across tasks and test sets when only task-specific corpora is used. We also report on how the impact of adding task-specific data to the transfer set correlates with the similarity between generic and task-specific data. Our results clearly indicate that, while distillation from a generic LM benefits downstream tasks, students learn better using target domain data even if it comes at the price of noisier teacher predictions. In other words, target domain data still trumps teacher knowledge.

[INDUSTRY] **Iterative Stratified Testing and Measurement for Automated Model Updates**
*Elizabeth Dekeyser, Nicholas Comment, Shermin Pei, Rajat Kumar, Shruti Rai, Fengtao Wu, Lisa Haverty and Kanna Shimizu* 09:45-10:00 (Collaboratorium)
Automating updates to machine learning systems is an important but understudied challenge in AutoML. The high model variance of many cutting-edge deep learning architectures means that retraining a model provides no guarantee of accurate inference on all sample types. To address this concern, we present Automated Data-Shape Stratified Model Updates (ADSMU), a novel framework that relies on iterative model building coupled with data-shape stratified model testing and improvement. Using ADSMU, we observed a 26% (relative) improvement in accuracy for new model use cases on a large-scale NLU system, compared to a naive (manually) retrained baseline and current cutting-edge methods.

[INDUSTRY] **Augmenting Operations Research with Auto-Formulation of Optimization Models From Problem Descriptions**
*Rindra Ramamonjison, Haley Li, Timothy TL Yu, Shiqi HE, Vishnu Rengan, Amin Banitalebi-Dehkordi, Zirui Zhou and Yong Zhang* 10:00-10:15 (Collaboratorium)
We describe an augmented intelligence system for simplifying and enhancing the modeling experience for operations research. Using this system, the user receives a suggested formulation of an optimization problem based on its description. To facilitate this process, we build an intuitive user interface system that enables the users to validate and edit the suggestions. We investigate controlled generation techniques to obtain an automatic suggestion of formulation. Then, we evaluate their effectiveness with a newly created dataset of linear programming problems drawn from various application domains.

[INDUSTRY] **Distilling Multilingual Transformers into CNNs for Scalable Intent Classification**
*Besnik Fetahu, Akash Veeragouni, Oleg Rokhlenko and Shervin Malmasi* 10:15-10:30 (Collaboratorium)
We describe an application of Knowledge Distillation used to distill and deploy multilingual Transformer models for voice assistants, enabling text classification for customers globally.Transformers have set new state-of-the-art results for tasks like intent classification, and multilingual models exploit cross-lingual transfer to allow serving requests across 100+ languages. However, their prohibitive inference time makes them impractical to deploy in real-world scenarios with low latency requirements, such as is the case of voice assistants. We address the problem of cross-architecture distillation of multilingual Transformers to simpler models, while maintaining multilinguality without performance degradation. Training multilingual student models has received little attention, and is our main focus. We show that a teacher-student framework, where the teacher's unscaled activations (logits) on unlabelled data are used to supervise student model training, enables distillation of Transformers into efficient multilingual CNN models. Our student model achieves equivalent performance as the teacher, and outperforms a similar model trained on the labelled data used to train the teacher model. This approach has enabled us to accurately serve global customer requests at speed (18x improvement), scale, and low cost.

## Poster Sessions 7 & 8

09:00-10:30 (Atrium)

**Memory-assisted prompt editing to improve GPT-3 after deployment**
*Aman Madaan, Niket Tandon, Peter Clark and Yiming Yang* 09:00-10:30 (Atrium)
Large LMs such as GPT-3 are powerful, but can commit mistakes that are obvious to humans. For example, GPT-3 would mistakenly interpret "What word is similar to good?" to mean a homophone, while the user intended a synonym. Our goal is to effectively correct such errors via user interactions with the system but without retraining, which will be prohibitively costly. We pair GPT-3 with a growing memory of recorded cases where the model misunderstood the user's intents, along with user feedback for clarification. Such a memory allows our system to produce enhanced prompts for any new query based on the user feedback for error correction on similar cases in the past. On four tasks (two lexical tasks, two advanced ethical reasoning tasks), we show how a (simulated) user can interactively teach a deployed GPT-3, substantially increasing its accuracy over the queries with different kinds of misunderstandings by the GPT-3. Our approach is a step towards the low-cost utility enhancement for very large pre-trained LMs.

**Two is Better than Many? Binary Classification as an Effective Approach to Multi-Choice Question Answering**
*Deepanway Ghosal, Navonil Majumder, Rada Mihalcea and Soujanya Poria* 09:00-10:30 (Atrium)
We propose a simple refactoring of multi-choice question answering (MCQA) tasks as a series of binary classifications. The MCQA task is generally performed by scoring each (question, answer) pair normalized over all the pairs, and then selecting the answer from the pair that yield the highest score. For n answer choices, this is equivalent to an n-class classification setup where only one class (true answer) is correct. We instead show that classifying (question, true answer) as positive instances and (question, false answer) as negative instances is significantly more effective across various models and datasets. We show the efficacy of our proposed approach in different tasks – abductive reasoning, commonsense question answering, science question answering, and sentence completion. Our DeBERTa binary classification model reaches the top or close to the top performance on public leaderboards for these tasks. The source code of the proposed approach is available at https://github.com/declare-lab/TEAM.

**Discovering Differences in the Representation of People using Contextualized Semantic Axes**
*Li Lucy, Divya Tadimeti and David Bamman* 09:00-10:30 (Atrium)
A common paradigm for identifying semantic differences across social and temporal contexts is the use of static word embeddings and their distances. In particular, past work has compared embeddings against "semantic axes" that represent two opposing concepts. We extend this paradigm to BERT embeddings, and construct contextualized axes that mitigate the pitfall where antonyms have neighboring representations. We validate and demonstrate these axes on two people-centric datasets: occupations from Wikipedia, and multi-platform discussions in extremist, men's communities over fourteen years. In both studies, contextualized semantic axes can characterize differences among instances of the same word type. In the latter study, we show that references to women and the contexts around them have become more detestable over time.

**Borrowing Human Senses: Comment-Aware Self-Training for Social Media Multimodal Classification**
*Chunpu Xu and Jing Li* 09:00-10:30 (Atrium)
Social media is daily creating massive multimedia content with paired image and text, presenting the pressing need to automate the vision and language understanding for various multimodal classification tasks. Compared to the commonly researched visual-lingual data, social media posts tend to exhibit more implicit image-text relations. To better glue the cross-modal semantics therein, we capture hinting features from user comments, which are retrieved via jointly leveraging visual and lingual similarity. Afterwards, the classification tasks are explored via self-training in a teacher-student framework, motivated by the usually limited labeled data scales in existing benchmarks. Substantial experiments are conducted on four multimodal social media benchmarks for image-text relation classification, sarcasm detection, sentiment classification, and hate speech detection. The results show that our method further advances the performance of previous state-of-the-art models, which do not employ comment modeling or self-training.

**Reflect, Not Reflex: Inference-Based Common Ground Improves Dialogue Response Quality**
*Pei Zhou, Hyundong Cho, Pegah Jandaghi, Dong-Ho Lee, Bill Yuchen Lin, Jay Pujara and Xiang Ren* 09:00-10:30 (Atrium)
Human communication relies on common ground (CG), the mutual knowledge and beliefs shared by participants, to produce coherent and interesting conversations. In this paper, we demonstrate that current response generation (RG) models produce generic and dull responses in dialogues because they act reflexively, failing to explicitly model CG, both due to the lack of CG in training data and the standard RG training procedure. We introduce Reflect, a dataset that annotates dialogues with explicit CG (materialized as inferences approximating shared knowledge and beliefs) and solicits 9k diverse human-generated responses each following one common ground. Using Reflect, we showcase the limitations of current dialogue data and RG models: less than half of the responses in current data is rated as high quality (sensible, specific, and interesting) and models trained using this data have even lower quality, while most Reflect responses are judged high quality. Next, we analyze whether CG can help models produce better quality responses by using Reflect CG to guide RG models. Surprisingly, we find that simply prompting GPT3 to "think" about CG generates 30% more quality responses, showing promising benefits to integrating CG into the RG process.

**Eliciting Knowledge from Large Pre-Trained Models for Unsupervised Knowledge-Grounded Conversation**
*Yanyang Li, Jianqiao Zhao, Michael Lyu and Liwei Wang* 09:00-10:30 (Atrium)
Recent advances in large-scale pre-training provide large models with the potential to learn knowledge from the raw text. It is thus natural to ask whether it is possible to leverage these large models as knowledge bases for downstream tasks. In this work, we answer the aforementioned question in unsupervised knowledge-grounded conversation. We explore various methods that best elicit knowledge from large models. Our human study indicates that, though hallucinations exist, large models post the unique advantage of being able to output common sense and summarize facts that cannot be directly retrieved from the search engine. To better exploit such generated knowledge in dialogue generation, we treat the generated knowledge as a noisy knowledge source and propose the posterior-based reweighing as well as the noisy training strategy. Empirical results on two benchmarks show advantages over the state-of-the-art methods.

**Discourse Comprehension: A Question Answering Framework to Represent Sentence Connections**
*Wei-Jen Ko, Cutter Dalton, Mark Simmons, Eliza Fisher, Greg Durrett and Junyi Jessy Li* 09:00-10:30 (Atrium)
While there has been substantial progress in text comprehension through simple factoid question answering, more holistic comprehension of a discourse still presents a major challenge (Dunietz et al., 2020). Someone critically reflecting on a text as they read it will pose curiosity-driven, often open-ended questions, which reflect deep understanding of the content and require complex reasoning to answer (Ko et al., 2020; Westera et al., 2020). A key challenge in building and evaluating models for this type of discourse comprehension is the lack of annotated data, especially since collecting answers to such questions requires high cognitive load for annotators.
This paper presents a novel paradigm that enables scalable data collection targeting the comprehension of news documents, viewing these

questions through the lens of discourse. The resulting corpus, DCQA (Discourse Comprehension by Question Answering), captures both discourse and semantic links between sentences in the form of free-form, open-ended questions. On an evaluation set that we annotated on questions from Ko et al. (2020), we show that DCQA provides valuable supervision for answering open-ended questions. We additionally design pre-training methods utilizing existing question-answering resources, and use synthetic data to accommodate unanswerable questions.

**The Optimal BERT Surgeon: Scalable and Accurate Second-Order Pruning for Large Language Models**
*Eldar Kurtic, Daniel Campos, Tuan Nguyen, Elias Frantar, Mark Kurtz, Benjamin Fineran, Michael Goin and Dan Alistarh* 09:00-10:30 (Atrium)
In this paper, we consider the problem of sparsifying BERT models, which are a key building block for natural language processing, in order to reduce their storage and computational cost. We introduce the Optimal BERT Surgeon (oBERT), an efficient and accurate pruning method based on approximate second-order information, which we show to yield state-of-the-art results in both stages of language tasks: pre-training and fine-tuning. Specifically, oBERT extends existing work on second-order pruning by allowing for pruning weight blocks, and is the first such method that is applicable at BERT scale. Second, we investigate compounding compression approaches to obtain highly compressed but accurate models for deployment on edge devices. These models significantly push boundaries of the current state-of-the-art sparse BERT models with respect to all metrics: model size, inference speed and task accuracy. For example, relative to the dense BERT-base, we obtain 10x model size compression with < 1% accuracy drop, 10x CPU-inference speedup with < 2% accuracy drop, and 29x CPU-inference speedup with < 7.5% accuracy drop. Our code, fully integrated with Transformers and SparseML, is available at https://github.com/neuralmagic/sparseml/tree/main/research/optimal_BERT_surgeon_oBERT.

**Syntactic Multi-view Learning for Open Information Extraction**
*Kuicai Dong, Aixin Sun, Jung-Jae Kim and Xiaoli Li* 09:00-10:30 (Atrium)
Open Information Extraction (OpenIE) aims to extract relational tuples from open-domain sentences. Traditional rule-based or statistical models were developed based on syntactic structure of sentence, identified by syntactic parsers. However, previous neural OpenIE models under-explored the useful syntactic information. In this paper, we model both constituency and dependency trees into word-level graphs, and enable neural OpenIE to learn from the syntactic structures. To better fuse heterogeneous information from the two graphs, we adopt multi-view learning to capture multiple relationships from them. Finally, the finetuned constituency and dependency representations are aggregated with sentential semantic representations for tuple generation. Experiments show that both constituency and dependency information, and the multi-view learning are effective.

**DuReader-Retrieval: A Large-scale Chinese Benchmark for Passage Retrieval from Web Search Engine**
*Yifu Qiu, Hongyu Li, Yingqi Qu, Ying Chen, QiaoQiao She, Jing Liu, Hua Wu and Haifeng Wang* 09:00-10:30 (Atrium)
In this paper, we present DuReader-retrieval, a large-scale Chinese dataset for passage retrieval. DuReader-retrieval contains more than 90K queries and over 8M unique passages from a commercial search engine. To alleviate the shortcomings of other datasets and ensure the quality of our benchmark, we (1) reduce the false negatives in development and test sets by manually annotating results pooled from multiple retrievers, and (2) remove the training queries that are semantically similar to the development and testing queries. Additionally, we provide two out-of-domain testing sets for cross-domain evaluation, as well as a set of human translated queries for for cross-lingual retrieval evaluation. The experiments demonstrate that DuReader-retrieval is challenging and a number of problems remain unsolved, such as the salient phrase mismatch and the syntactic mismatch between queries and paragraphs. These experiments also show that dense retrievers do not generalize well across domains, and cross-lingual retrieval is essentially challenging. DuReader-retrieval is publicly available at https://github.com/baidu/DuReader/tree/master/DuReader-Retrieval.

**CODER: An efficient framework for improving retrieval through COntextual Document Embedding Reranking**
*George Zerveas, Navid Rekabsaz, Daniel Cohen and Carsten Eickhoff* 09:00-10:30 (Atrium)
Contrastive learning has been the dominant approach to training dense retrieval models. In this work, we investigate the impact of ranking context - an often overlooked aspect of learning dense retrieval models. In particular, we examine the effect of its constituent parts: jointly scoring a large number of negatives per query, using retrieved (query-specific) instead of random negatives, and a fully list-wise loss.

To incorporate these factors into training, we introduce Contextual Document Embedding Reranking (CODER), a highly efficient retrieval framework. When reranking, it incurs only a negligible computational overhead on top of a first-stage method at run time (approx. 5 ms delay per query), allowing it to be easily combined with any state-of-the-art dual encoder method. Models trained through CODER can also be used as stand-alone retrievers.

Evaluating CODER in a large set of experiments on the MS MARCO and TripClick collections, we show that the contextual reranking of precomputed document embeddings leads to a significant improvement in retrieval performance. This improvement becomes even more pronounced when more relevance information per query is available, shown in the TripClick collection, where we establish new state-of-the-art results by a large margin.

**Finding Dataset Shortcuts with Grammar Induction**
*Dan Friedman, Alexander Wettig and Danqi Chen* 09:00-10:30 (Atrium)
Many NLP datasets have been found to contain shortcuts: simple decision rules that achieve surprisingly high accuracy. However, it is difficult to discover shortcuts automatically. Prior work on automatic shortcut detection has focused on enumerating features like unigrams or bigrams, which can find only low-level shortcuts, or relied on post-hoc model interpretability methods like saliency maps, which reveal qualitative patterns without a clear statistical interpretation. In this work, we propose to use probabilistic grammars to characterize and discover shortcuts in NLP datasets. Specifically, we use a context-free grammar to model patterns in sentence classification datasets and use a synchronous context-free grammar to model datasets involving sentence pairs. The resulting grammars reveal interesting shortcut features in a number of datasets, including both simple and high-level features, and automatically identify groups of test examples on which conventional classifiers fail. Finally, we show that the features we discover can be used to generate diagnostic contrast examples and incorporated into standard robust optimization methods to improve worst-group accuracy.

**SLING: Sino Linguistic Evaluation of Large Language Models**
*Yixiao Song, Kalpesh Krishna, Rajesh Bhatt and Mohit Iyyer* 09:00-10:30 (Atrium)
To understand what kinds of linguistic knowledge are encoded by pretrained Chinese language models (LMs), we introduce the benchmark of Sino LINGuistics (SLING), which consists of 38K minimal sentence pairs in Mandarin Chinese grouped into 9 high-level linguistic phenomena. Each pair demonstrates the acceptability contrast of a specific syntactic or semantic phenomenon (e.g., The keys are lost vs. The keys is lost), and an LM should assign lower perplexity to the acceptable sentence. In contrast to the CLiMP dataset (Xiang et al., 2021), which also contains Chinese minimal pairs and was created by translating the vocabulary of the English BLiMP dataset, the minimal pairs in SLING are derived primarily by applying syntactic and lexical transformations to naturally-occurring, linguist-annotated sentences from the Chinese Treebank 9.0, thus addressing severe issues in CLiMP's data generation process. We test 18 publicly available pretrained monolingual (e.g., BERT-base-zh, CPM) and multi-lingual (e.g., mT5, XLM) language models on SLING. Our experiments show that the average accuracy for LMs is far below human performance (69.7% vs. 97.1%), while BERT-base-zh achieves the highest accuracy (84.8%) of all tested LMs, even much larger ones. Additionally, we find that most LMs have a strong gender and number (singular/plural) bias, and they perform better on

local phenomena than hierarchical ones.

**Textual Backdoor Attacks Can Be More Harmful via Two Simple Tricks**
*Yangyi Chen, Fanchao Qi, Hongcheng Gao, Zhiyuan Liu and Maosong Sun* 09:00-10:30 (Atrium)
Backdoor attacks are a kind of emergent security threat in deep learning. After being injected with a backdoor, a deep neural model will behave normally on standard inputs but give adversary-specified predictions once the input contains specific backdoor triggers. In this paper, we find two simple tricks that can make existing textual backdoor attacks much more harmful. The first trick is to add an extra training task to distinguish poisoned and clean data during the training of the victim model, and the second one is to use all the clean training data rather than remove the original clean data corresponding to the poisoned data. These two tricks are universally applicable to different attack models. We conduct experiments in three tough situations including clean data fine-tuning, low-poisoning-rate, and label-consistent attacks. Experimental results show that the two tricks can significantly improve attack performance. This paper exhibits the great potential harmfulness of backdoor attacks. All the code and data can be obtained at `https://github.com/thunlp/StyleAttack`.

**AdaMix: Mixture-of-Adaptations for Parameter-efficient Model Tuning**
*Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah and Jianfeng Gao* 09:00-10:30 (Atrium)
Standard fine-tuning of large pre-trained language models (PLMs) for downstream tasks requires updating hundreds of millions to billions of parameters, and storing a large copy of the PLM weights for every task resulting in increased cost for storing, sharing and serving the models. To address this, parameter-efficient fine-tuning (PEFT) techniques were introduced where small trainable components are injected in the PLM and updated during fine-tuning. We propose AdaMix as a general PEFT method that tunes a mixture of adaptation modules – given the underlying PEFT method of choice – introduced in each Transformer layer while keeping most of the PLM weights frozen. For instance, AdaMix can leverage a mixture of adapters like Houlsby or a mixture of low rank decomposition matrices like LoRA to improve downstream task performance over the corresponding PEFT methods for fully supervised and few-shot NLU and NLG tasks. Further, we design AdaMix such that it matches the same computational cost and the number of tunable parameters as the underlying PEFT method. By only tuning 0.1-0.2% of PLM parameters, we show that AdaMix outperforms SOTA parameter-efficient fine-tuning and full model fine-tuning for both NLU and NLG tasks.

**Spectral Probing**
*Max Müller-Eberstein, Rob van der Goot and Barbara Plank* 09:00-10:30 (Atrium)
Linguistic information is encoded at varying timescales (subwords, phrases, etc.) and communicative levels, such as syntax and semantics. Contextualized embeddings have analogously been found to capture these phenomena at distinctive layers and frequencies. Leveraging these findings, we develop a fully learnable frequency filter to identify spectral profiles for any given task. It enables vastly more granular analyses than prior handcrafted filters, and improves on efficiency. After demonstrating the informativeness of spectral probing over manual filters in a monolingual setting, we investigate its multilingual characteristics across seven diverse NLP tasks in six languages. Our analyses identify distinctive spectral profiles which quantify cross-task similarity in a linguistically intuitive manner, while remaining consistent across languages—highlighting their potential as robust, lightweight task descriptors.

**Cascading Biases: Investigating the Effect of Heuristic Annotation Strategies on Data and Models**
*Chaitanya Malaviya, Sudeep Bhatia and Mark Yatskar* 09:00-10:30 (Atrium)
Cognitive psychologists have documented that humans use cognitive heuristics, or mental shortcuts, to make quick decisions while expending less effort. While performing annotation work on crowdsourcing platforms, we hypothesize that such heuristic use among annotators cascades on to data quality and model robustness. In this work, we study cognitive heuristic use in the context of annotating multiple-choice reading comprehension datasets. We propose tracking annotator heuristic traces, where we tangibly measure low-effort annotation strategies that could indicate usage of various cognitive heuristics. We find evidence that annotators might be using multiple such heuristics, based on correlations with a battery of psychological tests. Importantly, heuristic use among annotators determines data quality along several dimensions: (1) known biased models, such as partial input models, more easily solve examples authored by annotators that rate highly on heuristic use, (2) models trained on annotators scoring highly on heuristic use don't generalize as well, and (3) heuristic-seeking annotators tend to create qualitatively less challenging examples. Our findings suggest that tracking heuristic usage among annotators can potentially help with collecting challenging datasets and diagnosing model biases.

**Quality Scoring of Source Words in Neural Translation Models**
*Priyesh Jain, Sunita Sarawagi and Tushar Tomar* 09:00-10:30 (Atrium)
Word-level quality scores on input source sentences can provide useful feedback to an end-user when translating into an unfamiliar target language. Recent approaches either require training special word-scoring models based on synthetic data or require repeated invocation of the translation model. We propose a simple approach based on comparing the difference of probabilities from two language models. The basic premise of our method is to reason how well each source word is explained by the target sentence as against the source language model. Our approach provides up to five points higher F1 scores and is significantly faster than the state of the art methods on three language pairs. Also, our method does not require training any new model. We release a public dataset on word omissions and mistranslations on a new language pair.

**Language Contamination Helps Explains the Cross-lingual Capabilities of English Pretrained Models**
*Terra Blevins and Luke Zettlemoyer* 09:00-10:30 (Atrium)
English pretrained language models, which make up the backbone of many modern NLP systems, require huge amounts of unlabeled training data. These models are generally presented as being trained only on English text but have been found to transfer surprisingly well to other languages. We investigate this phenomenon and find that common English pretraining corpora actually contain significant amounts of non-English text: even when less than 1% of data is not English (well within the error rate of strong language classifiers), this leads to hundreds of millions of foreign language tokens in large-scale datasets. We then demonstrate that even these small percentages of non-English data facilitate cross-lingual transfer for models trained on them, with target language performance strongly correlated to the amount of in-language data seen during pretraining. In light of these findings, we argue that no model is truly monolingual when pretrained at scale, which should be considered when evaluating cross-lingual transfer.

**PRO-CS : An Instance-Based Prompt Composition Technique for Code-Switched Tasks**
*Srijan Bansal, Suraj Tripathi, Sumit Agarwal, Teruko Mitamura and Eric Nyberg* 09:00-10:30 (Atrium)
Code-switched (CS) data is ubiquitous in today's globalized world, but the dearth of annotated datasets in code-switching poses a significant challenge for learning diverse tasks across different language pairs. Parameter-efficient prompt-tuning approaches conditioned on frozen language models have shown promise for transfer learning in limited-resource setups. In this paper, we propose a novel instance-based prompt composition technique, PRO-CS, for CS tasks that combine language and task knowledge. We compare our approach with prompt-tuning and fine-tuning for code-switched tasks on 10 datasets across 4 language pairs. Our model outperforms the prompt-tuning approach by significant margins across all datasets and outperforms or remains at par with fine-tuning by using just 0.18% of total parameters. We also achieve com-

petitive results when compared with the fine-tuned model in the low-resource cross-lingual and cross-task setting, indicating the effectiveness of our approach to incorporate new code-switched tasks.

**Multitask Instruction-based Prompting for Fallacy Recognition**
*Tariq Alhindi, Tuhin Chakrabarty, Elena Musi and Smaranda Muresan* 09:00-10:30 (Atrium)
Fallacies are used as seemingly valid arguments to support a position and persuade the audience about its validity. Recognizing fallacies is an intrinsically difficult task both for humans and machines. Moreover, a big challenge for computational models lies in the fact that fallacies are formulated differently across the datasets with differences in the input format (e.g., question-answer pair, sentence with fallacy fragment), genre (e.g., social media, dialogue, news), as well as types and number of fallacies (from 5 to 18 types per dataset). To move towards solving the fallacy recognition task, we approach these differences across datasets as multiple tasks and show how instruction-based prompting in a multitask setup based on the T5 model improves the results against approaches built for a specific dataset such as T5, BERT or GPT-3. We show the ability of this multitask prompting approach to recognize 28 unique fallacies across domains and genres and study the effect of model size and prompt choice by analyzing the per-class (i.e., fallacy type) results. Finally, we analyze the effect of annotation quality on model performance, and the feasibility of complementing this approach with external knowledge.

**Unsupervised Non-transferable Text Classification**
*Guangtao Zeng and Wei Lu* 09:00-10:30 (Atrium)
Training a good deep learning model requires substantial data and computing resources, which makes the resulting neural model a valuable intellectual property. To prevent the neural network from being undesirably exploited, non-transferable learning has been proposed to reduce the model generalization ability in specific target domains. However, existing approaches require labeled data for the target domain which can be difficult to obtain. Furthermore, they do not have the mechanism to still recover the model's ability to access the target domain. In this paper, we propose a novel unsupervised non-transferable learning method for the text classification task that does not require annotated target domain data. We further introduce a secret key component in our approach for recovering the access to the target domain, where we design both an explicit and an implicit method for doing so. Extensive experiments demonstrate the effectiveness of our approach.

**Data-Efficient Playlist Captioning With Musical and Linguistic Knowledge**
*Giovanni Gabbolini, Romain Hennequin and Elena Epure* 09:00-10:30 (Atrium)
Music streaming services feature billions of playlists created by users, professional editors or algorithms. In this content overload scenario, it is crucial to characterise playlists, so that music can be effectively organised and accessed. Playlist titles and descriptions are proposed in natural language either manually by music editors and users or automatically from pre-defined templates. However, the former is time-consuming while the latter is limited by the vocabulary and covered music themes. In this work, we propose PlayNTell, a data-efficient multi-modal encoder-decoder model for automatic playlist captioning. Compared to existing music captioning algorithms, PlayNTell leverages also linguistic and musical knowledge to generate correct and thematic captions. We benchmark PlayNTell on a new editorial playlists dataset collected from two major music streaming services. PlayNTell yields 2x-3x higher BLEU@4 and CIDEr than state of the art captioning algorithms.

[CL] **How Much Does Lookahead Matter for Disambiguation? Partial Arabic Diacritization Case Study**
*Saeed Esmail, Kfir Bar and Nachum Dershowitz* 09:00-10:30 (Atrium)
We suggest a model for partial diacritization of deep orthographies. We focus on Arabic, where the optional indication of selected vowels by means of diacritics can resolve ambiguity and improve readability. Our partial diacritizer restores short vowels only when they contribute to the ease of understandability during reading a given running text. The idea is to identify those uncertainties of absent vowels that require the reader to look ahead to disambiguate. To achieve this, two independent neural networks are employed for predicting diacritics, one that takes the entire sentence as input and another that considers only the text that has been read thus far. Partial diacritization is then determined by retaining precisely those vowels on which the two networks disagree, preferring the reading based on consideration of the whole sentence over the more naive reading-order diacritization. For evaluation, we prepared a new dataset of Arabic texts with both full and partial vowelization. In addition to facilitating readability, we find that our partial diacritizer improves translation quality compared either to their total absence or to random selection. Lastly, we study the benefit of knowing the text that follows the word in focus towards the restoration of short vowels during reading, and we measure the degree to which lookahead contributes to resolving ambiguities encountered while reading.

[CL] **Revise and Resubmit: An Intertextual Model of Text-based Collaboration in Peer Review**
*Ilia Kuznetsov, Jan Buchmann, Max Eichler and Iryna Gurevych* 09:00-10:30 (Atrium)
Peer review is a key component of the publishing process in most fields of science. Increasing submission rates put a strain on reviewing quality and efficiency, motivating the development of applications to support the reviewing and editorial work. While existing NLP studies focus on the analysis of individual texts, editorial assistance often requires modeling interactions between pairs of texts—yet general frameworks and datasets to support this scenario are missing. Relationships between texts are the core object of the intertextuality theory—a family of approaches in literary studies not yet operationalized in NLP. Inspired by prior theoretical work, we propose the first intertextual model of text-based collaboration, which encompasses three major phenomena that make up a full iteration of the review–revise–and–resubmit cycle – pragmatic tagging, linking, and long-document version alignment. While peer review is used across the fields of science and publication formats, existing datasets solely focus on conference-style review in computer science. Addressing this, we instantiate our proposed model in the first annotated multidomain corpus in journal-style post-publication open peer review, and provide detailed insights into the practical aspects of intertextual annotation. Our resource is a major step toward multidomain, fine-grained applications of NLP in editorial support for peer review, and our intertextual framework paves the path for general-purpose modeling of text-based collaboration. We make our corpus, detailed annotation guidelines, and accompanying code publicly available.

**Re3: Generating Longer Stories With Recursive Reprompting and Revision**
*Kevin Yang, Yuandong Tian, Nanyun Peng and Dan Klein* 09:00-10:30 (Atrium)
We consider the problem of automatically generating longer stories of over two thousand words. Compared to prior work on shorter stories, long-range plot coherence and relevance are more central challenges here. We propose the Recursive Reprompting and Revision framework (Re3) to address these challenges by (a) prompting a general-purpose language model to construct a structured overarching plan, and (b) generating story passages by repeatedly injecting contextual information from both the plan and current story state into a language model prompt. We then revise by (c) reranking different continuations for plot coherence and premise relevance, and finally (d) editing the best continuation for factual consistency. Compared to similar-length stories generated directly from the same base model, human evaluators judged substantially more of Re3's stories as having a coherent overarching plot (by 14% absolute increase), and relevant to the given initial premise (by 20%).

**IndicNLG Benchmark: Multilingual Datasets for Diverse NLG Tasks in Indic Languages**
*Aman Kumar, Himani Shrotriya, Prachi Sahu, Amogh Mishra, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Mitesh M. Khapra and Pratyush Kumar* 09:00-10:30 (Atrium)
Natural Language Generation (NLG) for non-English languages is hampered by the scarcity of datasets in these languages. We present the In-

dicNLG Benchmark, a collection of datasets for benchmarking NLG for 11 Indic languages. We focus on five diverse tasks, namely, biography generation using Wikipedia infoboxes, news headline generation, sentence summarization, paraphrase generation and, question generation. We describe the created datasets and use them to benchmark the performance of several monolingual and multilingual baselines that leverage pre-trained sequence-to-sequence models. Our results exhibit the strong performance of multilingual language-specific pre-trained models, and the utility of models trained on our dataset for other related NLG tasks. Our dataset creation methods can be easily applied to modest-resource languages as they involve simple steps such as scraping news articles and Wikipedia infoboxes, light cleaning, and pivoting through machine translation data. To the best of our knowledge, the IndicNLG Benchmark is the first NLG benchmark for Indic languages and the most diverse multilingual NLG dataset, with approximately 8M examples across 5 tasks and 11 languages. The datasets and models will be publicly available.

**You Only Need One Model for Open-domain Question Answering**
*Haejun Lee, Akhil Kedia, Jongwon Lee, Ashwin Paranjape, Christopher Manning and Kyoung-Gu Woo* 09:00-10:30 (Atrium)
Recent approaches to Open-domain Question Answering refer to an external knowledge base using a retriever model, optionally rerank passages with a separate reranker model and generate an answer using another reader model. Despite performing related tasks, the models have separate parameters and are weakly-coupled during training. We propose casting the retriever and the reranker as internal passage-wise attention mechanisms applied sequentially within the transformer architecture and feeding computed representations to the reader, with the hidden representations progressively refined at each stage. This allows us to use a single question answering model trained end-to-end, which is a more efficient use of model capacity and also leads to better gradient flow. We present a pre-training method to effectively train this architecture and evaluate our model on the Natural Questions and TriviaQA open datasets. For a fixed parameter budget, our model outperforms the previous state-of-the-art model by 1.0 and 0.7 exact match scores.

**ASQA: Factoid Questions Meet Long-Form Answers**
*Ivan Stelmakh, Yi Luan, Bhuwan Dhingra and Ming-Wei Chang* 09:00-10:30 (Atrium)
Recent progress on open domain factoid question answering (QA) does not easily transfer to the task of long-form QA, where the goal is to answer questions that require in-depth explanations. The hurdles include a lack of high-quality data and the absence of a well-defined notion of an answer's quality. In this work, we address these problems by releasing a novel dataset and a task that we call ASQA (Answer Summaries for Questions which are Ambiguous); and proposing a reliable metric for measuring performance on ASQA. Our task focuses on ambiguous factoid questions which have different correct answers depending on the interpretation. Answers to ambiguous questions should combine factual information from multiple sources into a coherent long-form summary that resolves the ambiguity. In contrast to existing long-form QA tasks (such as ELI5), ASQA admits a clear notion of correctness: a user faced with a good summary should be able to answer different interpretations of the original ambiguous question. Our analysis demonstrates an agreement between this metric and human judgments, and reveals a considerable gap between human performance and strong baselines.

**ReasTAP: Injecting Table Reasoning Skills During Pre-training via Synthetic Reasoning Examples**
*Yilun Zhao, Linyong Nan, Zhenting Qi, Rui Zhang and Dragomir Radev* 09:00-10:30 (Atrium)
Reasoning over tabular data requires both table structure understanding and a broad set of table reasoning skills. Current models with table-specific architectures and pre-training methods perform well on understanding table structures, but they still struggle with tasks that require various table reasoning skills. In this work, we develop ReasTAP to show that high-level table reasoning skills can be injected into models during pre-training without a complex table-specific architecture design. We define 7 table reasoning skills, such as numerical operation, temporal comparison, and conjunction. Each reasoning skill is associated with one example generator, which synthesizes questions over semi-structured tables according to the sampled templates. We model the table pre-training task as a sequence generation task and pre-train ReasTAP to generate precise answers of the synthetic examples. ReasTAP is evaluated on four benchmarks covering three downstream tasks including 1) WikiSQL-Weak and WikiTQ for Table Question Answering, 2) TabFact for Table Fact Verification, and 3) LogicNLG for Faithful Table-to-Text Generation. Experimental results demonstrate that ReasTAP achieves new state-of-the-art results on all of them and delivers a significant improvement under low-resource setting. Our code is publicly available at https://github.com/Yale-LILY/ReasTAP.

**Knowledge Transfer from Answer Ranking to Answer Generation**
*Matteo Gabburo, Rik Koncel-Kedziorski, Siddhant Garg, Luca Soldaini and Alessandro Moschitti* 09:00-10:30 (Atrium)
Recent studies show that Question Answering (QA) based on Answer Sentence Selection (AS2) can be improved by generating an improved answer from the top-k ranked answer sentences (termed GenQA). This allows for synthesizing the information from multiple candidates into a concise, natural-sounding answer. However, creating large-scale supervised training data for GenQA models is very challenging. In this paper, we propose to train a GenQA model by transferring knowledge from a trained AS2 model, to overcome the aforementioned issue. First, we use an AS2 model to produce a ranking over answer candidates for a set of questions. Then, we use the top ranked candidate as the generation target, and the next k top ranked candidates as context for training a GenQA model. We also propose to use the AS2 model prediction scores for loss weighting and score-conditioned input/output shaping, to aid the knowledge transfer. Our evaluation on three public and one large industrial datasets demonstrates the superiority of our approach over the AS2 baseline, and GenQA trained using supervised data.

**Pre-training Transformer Models with Sentence-Level Objectives for Answer Sentence Selection**
*Luca Di Liello, Siddhant Garg, Luca Soldaini and Alessandro Moschitti* 09:00-10:30 (Atrium)
An important task for designing QA systems is answer sentence selection (AS2): selecting the sentence containing (or constituting) the answer to a question from a set of retrieved relevant documents. In this paper, we propose three novel sentence-level transformer pre-training objectives that incorporate paragraph-level semantics within and across documents, to improve the performance of transformers for AS2, and mitigate the requirement of large labeled datasets. Specifically, the model is tasked to predict whether: (i) two sentences are extracted from the same paragraph, (ii) a given sentence is extracted from a given paragraph, and (iii) two paragraphs are extracted from the same document. Our experiments on three public and one industrial AS2 datasets demonstrate the empirical superiority of our pre-trained transformers over baseline models such as RoBERTa and ELECTRA for AS2.

**Towards Knowledge-Intensive Text-to-SQL Semantic Parsing with Formulaic Knowledge**
*Longxu Dou, Yan Gao, Xuqi Liu, Mingyang Pan, Dingzirui Wang, Wanxiang Che, Dechen Zhan, Min-Yen Kan and Jian-Guang LOU* 09:00-10:30 (Atrium)
In this paper, we study the problem of knowledge-intensive text-to-SQL, in which domain knowledge is necessary to parse expert questions into SQL queries over domain-specific tables. We formalize this scenario by building a new benchmark KnowSQL consisting of domain-specific questions covering various domains. We then address this problem by representing formulaic knowledge rather than by annotating additional data examples. More concretely, we construct a formulaic knowledge bank as a domain knowledge base and propose a framework (ReGrouP) to leverage this formulaic knowledge during parsing. Experiments using ReGrouP demonstrate a significant 28.2% improvement overall on KnowSQL.

**Exploring Document-Level Literary Machine Translation with Parallel Paragraphs from World Literature**
*Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting and Mohit Iyyer* 09:00-10:30 (Atrium)

Literary translation is a culturally significant task, but it is bottlenecked by the small number of qualified literary translators relative to the many untranslated works published around the world. Machine translation (MT) holds potential to complement the work of human translators by improving both training procedures and their overall efficiency. Literary translation is less constrained than more traditional MT settings since translators must balance meaning equivalence, readability, and critical interpretability in the target language. This property, along with the complex discourse-level context present in literary texts, also makes literary MT more challenging to computationally model and evaluate. To explore this task, we collect a dataset (Par3) of non-English language novels in the public domain, each aligned at the paragraph level to both human and automatic English translations. Using Par3, we discover that expert literary translators prefer reference human translations over machine-translated paragraphs at a rate of 84

**Making Science Simple: Corpora for the Lay Summarisation of Scientific Literature**
*Tomas Goldsack, Zhihao Zhang, Chenghua Lin and Carolina Scarton* 09:00-10:30 (Atrium)
Lay summarisation aims to jointly summarise and simplify a given text, thus making its content more comprehensible to non-experts. Automatic approaches for lay summarisation can provide significant value in broadening access to scientific literature, enabling a greater degree of both interdisciplinary knowledge sharing and public understanding when it comes to research findings. However, current corpora for this task are limited in their size and scope, hindering the development of broadly applicable data-driven approaches. Aiming to rectify these issues, we present two novel lay summarisation datasets, PLOS (large-scale) and eLife (medium-scale), each of which contains biomedical journal articles alongside expert-written lay summaries. We provide a thorough characterisation of our lay summaries, highlighting differing levels of readability and abstractiveness between datasets that can be leveraged to support the needs of different applications. Finally, we benchmark our datasets using mainstream summarisation approaches and perform a manual evaluation with domain experts, demonstrating their utility and casting light on the key challenges of this task.

**Generate, Discriminate and Contrast: A Semi-Supervised Sentence Representation Learning Framework**
*Yiming Chen, Yan Zhang, Bin Wang, ZUOZHU LIU and Haizhou Li* 09:00-10:30 (Atrium)
Most sentence embedding techniques heavily rely on expensive human-annotated sentence pairs as the supervised signals. Despite the use of large-scale unlabeled data, the performance of unsupervised methods typically lags far behind that of the supervised counterparts in most downstream tasks. In this work, we propose a semi-supervised sentence embedding framework, GenSE, that effectively leverages large-scale unlabeled data. Our method include three parts: 1) Generate: A generator/discriminator model is jointly trained to synthesize sentence pairs from open-domain unlabeled corpus; 2) Discriminate: Noisy sentence pairs are filtered out by the discriminator to acquire high-quality positive and negative sentence pairs; 3) Contrast: A prompt-based contrastive approach is presented for sentence representation learning with both annotated and synthesized data. Comprehensive experiments show that GenSE achieves an average correlation score of 85.19 on the STS datasets and consistent performance improvement on four domain adaptation tasks, significantly surpassing the state-of-the-art methods and convincingly corroborating its effectiveness and generalization ability.

**Looking at the Overlooked: An Analysis on the Word-Overlap Bias in Natural Language Inference**
*Sara Rajaee, Yadollah Yaghoobzadeh and Mohammad Taher Pilehvar* 09:00-10:30 (Atrium)
It has been shown that NLI models are usually biased with respect to the word-overlap between the premise and the hypothesis, as they take this feature as a primary cue for predicting the entailment label. In this paper, we focus on an overlooked aspect of the overlap bias in the NLI models: the reverse word-overlap bias. Our experimental results demonstrate that current NLI systems are also highly biased towards the non-entailment label on instances with low overlap and that existing debiasing methods, which are reportedly successful on challenge datasets, are generally ineffective in addressing this category of bias. Through a set of analyses, we investigate the reasons for the emergence of the overlap bias and the role of minority examples in mitigating this bias. For the former, we find that the word overlap bias does not stem from pre-training, and in the latter, we observe that in contrast to the accepted assumption, eliminating minority examples does not affect the generalizability of debiasing methods with respect to the overlap bias.

**CPL: Counterfactual Prompt Learning for Vision and Language Models**
*Xuehai He, Diji Yang, Weixi Feng, Tsu-Jui Fu, Arjun Akula, Varun Jampani, Pradyumna Narayana, Sugato Basu, William Yang Wang and Xin Wang* 09:00-10:30 (Atrium)
Prompt tuning is a new few-shot transfer learning technique that only tunes the learnable prompt for pre-trained vision and language models such as CLIP. However, existing prompt tuning methods tend to learn spurious or entangled representations, which leads to poor generalization to unseen concepts. Towards non-spurious and efficient prompt learning from limited examples, this paper presents a novel Counterfactual Prompt Learning (CPL) method for vision and language models, which simultaneously employs counterfactual generation and contrastive learning in a joint optimization framework. Particularly, CPL constructs counterfactual by identifying minimal non-spurious feature change between semantically-similar positive and negative samples that causes concept change, and learns more generalizable prompt representation from both factual and counterfactual examples via contrastive learning. Extensive experiments demonstrate that CPL can obtain superior few-shot performance on different vision and language tasks than previous prompt tuning methods on CLIP. On image classification, we achieve 3.55% average relative improvement on unseen classes across seven datasets; on image-text retrieval and visual question answering, we gain up to 4.09% and 25.08% relative improvements across three few-shot scenarios on unseen test sets respectively.

**MGDoc: Pre-training with Multi-granular Hierarchy for Document Image Understanding**
*Zilong Wang, Jiuxiang Gu, Chris Tensmeyer, Nikolaos Barmpalios, Ani Nenkova, Tong Sun, Jingbo Shang and Vlad Morariu* 09:00-10:30 (Atrium)
Document images are a ubiquitous source of data where the text is organized in a complex hierarchical structure ranging from fine granularity (e.g., words), medium granularity (e.g., regions such as paragraphs or figures), to coarse granularity (e.g., the whole page). The spatial hierarchical relationships between content at different levels of granularity are crucial for document image understanding tasks. Existing methods learn features from either word-level or region-level but fail to consider both simultaneously. Word-level models are restricted by the fact that they originate from pure-text language models, which only encode the word-level context. In contrast, region-level models attempt to encode regions corresponding to paragraphs or text blocks into a single embedding, but they perform worse with additional word-level features. To deal with these issues, we propose MGDoc, a new multi-modal multi-granular pre-training framework that encodes page-level, region-level, and word-level information at the same time. MGDoc uses a unified text-visual encoder to obtain multi-modal features across different granularities, which makes it possible to project the multi-granular features into the same hyperspace. To model the region-word correlation, we design a cross-granular attention mechanism and specific pre-training tasks for our model to reinforce the model of learning the hierarchy between regions and words. Experiments demonstrate that our proposed model can learn better features that perform well across granularities and lead to improvements in downstream tasks.

**Cross-Modal Similarity-Based Curriculum Learning for Image Captioning**
*Hongkuan Zhang, Saku Sugawara, Akiko Aizawa, Lei Zhou, Ryohei Sasano and Koichi Takeda* 09:00-10:30 (Atrium)
Image captioning models require the high-level generalization ability to describe the contents of various images in words. Most existing approaches treat the image–caption pairs equally in their training without considering the differences in their learning difficulties. Several image captioning approaches introduce curriculum learning methods that present training data with increasing levels of difficulty. However, their

difficulty measurements are either based on domain-specific features or prior model training. In this paper, we propose a simple yet efficient difficulty measurement for image captioning using cross-modal similarity calculated by a pretrained vision–language model. Experiments on the COCO and Flickr30k datasets show that our proposed approach achieves superior performance and competitive convergence speed to baselines without requiring heuristics or incurring additional training costs. Moreover, the higher model performance on difficult examples and unseen data also demonstrates the generalization ability.

**Textless Speech Emotion Conversion using Discrete & Decomposed Representations**
*Felix Kreuk, Adam Polyak, Jade Copet, Eugene Kharitonov, Tu Anh Nguyen, Morgan Rivière, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux and Yossi Adi* 09:00-10:30 (Atrium)
Speech emotion conversion is the task of modifying the perceived emotion of a speech utterance while preserving the lexical content and speaker identity. In this study, we cast the problem of emotion conversion as a spoken language translation task. We use a decomposition of the speech signal into discrete learned representations, consisting of phonetic-content units, prosodic features, speaker, and emotion. First, we modify the speech content by translating the phonetic-content units to a target emotion, and then predict the prosodic features based on these units. Finally, the speech waveform is generated by feeding the predicted representations into a neural vocoder. Such a paradigm allows us to go beyond spectral and parametric changes of the signal, and model non-verbal vocalizations, such as laughter insertion, yawning removal, etc. We demonstrate objectively and subjectively that the proposed method is vastly superior to current approaches and even beats text-based systems in terms of perceived emotion and audio quality. We rigorously evaluate all components of such a complex system and conclude with an extensive model analysis and ablation study to better emphasize the architectural choices, strengths and weaknesses of the proposed method. Samples are available under the following link: https://speechbot.github.io/emotion

**X-FACTOR: A Cross-metric Evaluation of Factual Correctness in Abstractive Summarization**
*Subhajit Chaudhury, Sarathkrishna Swaminathan, Chulaka Gunasekara, Maxwell Crouse, Srinivas Ravishankar, Daiki Kimura, Keerthiram Murugesan, Ramón Fernandez Astudillo, Tahira Naseem, Pavan Kapanipathi and Alexander Gray* 09:00-10:30 (Atrium)
Abstractive summarization models often produce factually inconsistent summaries that are not supported by the original article. Recently, a number of fact-consistent evaluation techniques have been proposed to address this issue; however, a detailed analysis of how these metrics agree with one another has yet to be conducted. In this paper, we present X-FACTOR, a cross-evaluation of three high-performing fact-aware abstractive summarization methods. First, we show that summarization models are often fine-tuned on datasets that contain factually inconsistent summaries and propose a fact-aware filtering mechanism that improves the quality of training data and, consequently, the factuality of these models. Second, we propose a corrector module that can be used to improve the factual consistency of generated summaries. Third, we present a re-ranking technique that samples summary instances from the output distribution of a summarization model and re-ranks the sampled instances based on their factuality. Finally, we provide a detailed cross-metric agreement analysis that shows how tuning a model to output summaries based on a particular factuality metric influences factuality as determined by the other metrics. Our goal in this work is to facilitate research that improves the factuality and faithfulness of abstractive summarization models.

**Unsupervised Opinion Summarisation in the Wasserstein Space**
*Jiayu Song, Iman Munire Bilal, Adam Tsakalidis, Rob Procter and Maria Liakata* 09:00-10:30 (Atrium)
Opinion summarisation synthesises opinions expressed in a group of documents discussing the same topic to produce a single summary. Recent work has looked at opinion summarisation of clusters of social media posts. Such posts are noisy and have unpredictable structure, posing additional challenges for the construction of the summary distribution and the preservation of meaning compared to online reviews, which has been so far the focus on opinion summarisation. To address these challenges we present WassOS, an unsupervised abstractive summarization model which makes use of the Wasserstein distance. A Variational Autoencoder is first used to obtain the distribution of documents/posts, and the summary distribution is obtained as the Wasserstein barycenter. We create separate disentangled latent semantic and syntactic representations of the summary, which are fed into a GRU decoder with a transformer layer to produce the final summary. Our experiments on multiple datasets including reviews, Twitter clusters and Reddit threads show that WassOS almost always outperforms the state-of-the-art on ROUGE metrics and consistently produces the best summaries with respect to meaning preservation according to human evaluations.

**Referee: Reference-Free Sentence Summarization with Sharper Controllability through Symbolic Knowledge Distillation**
*Melanie Sclar, Peter West, Sachin Kumar, Yulia Tsvetkov and Yejin Choi* 09:00-10:30 (Atrium)
We present Referee, a novel framework for sentence summarization that can be trained reference-free (i.e., requiring no gold summaries for supervision), while allowing direct control for compression ratio. Our work is the first to demonstrate that reference-free, controlled sentence summarization is feasible via the conceptual framework of Symbolic Knowledge Distillation (West et al., 2022), where latent knowledge in pre-trained language models is distilled via explicit examples sampled from the teacher models, further purified with three types of filters: length, fidelity, and Information Bottleneck. Moreover, we uniquely propose iterative distillation of knowledge, where student models from the previous iteration of distillation serve as teacher models in the next iteration. Starting off from a relatively modest set of GPT3-generated summaries, we demonstrate how iterative knowledge distillation can lead to considerably smaller, but better summarizers with sharper controllability. A useful by-product of this iterative distillation process is a high-quality dataset of sentence-summary pairs with varying degrees of compression ratios. Empirical results demonstrate that the final student models vastly outperform the much larger GPT3-Instruct model in terms of the controllability of compression ratios, without compromising the quality of resulting summarization.

**Learning to Generate Overlap Summaries through Noisy Synthetic Data**
*Naman Bansal, Mousumi Akter and Shubhra Kanti Karmaker Santu* 09:00-10:30 (Atrium)
Semantic Overlap Summarization (SOS) is a novel and relatively under-explored seq-to-seq task which entails summarizing common information from multiple alternate narratives. One of the major challenges for solving this task is the lack of existing datasets for supervised training. To address this challenge, we propose a novel data augmentation technique, which allows us to create large amount of synthetic data for training a seq-to-seq model that can perform the SOS task. Through extensive experiments using narratives from the news domain, we show that the models fine-tuned using the synthetic dataset provide significant performance improvements over the pre-trained vanilla summarization techniques and are close to the models fine-tuned on the golden training data; which essentially demonstrates the effectiveness of out proposed data augmentation technique for training seq-to-seq models on the SOS task.

**Questioning the Validity of Summarization Datasets and Improving Their Factual Consistency**
*Yanzhu Guo, Chloé Clavel, Moussa Kamal Eddine and michalis vazirgiannis* 09:00-10:30 (Atrium)
The topic of summarization evaluation has recently attracted a surge of attention due to the rapid development of abstractive summarization systems. However, the formulation of the task is rather ambiguous, neither the linguistic nor the natural language processing communities have succeeded in giving a mutually agreed-upon definition. Due to this lack of well-defined formulation, a large number of popular abstractive summarization datasets are constructed in a manner that neither guarantees validity nor meets one of the most essential criteria of summarization: factual consistency. In this paper, we address this issue by combining state-of-the-art factual consistency models to identify the problematic instances present in popular summarization datasets. We release SummFC, a filtered summarization dataset with improved factual consistency, and demonstrate that models trained on this dataset achieve improved performance in nearly all quality aspects. We argue

that our dataset should become a valid benchmark for developing and evaluating summarization systems.

**The Authenticity Gap in Human Evaluation**
*Kawin Ethayarajh and Dan Jurafsky* 09:00-10:30 (Atrium)
Human ratings are the gold standard in NLG evaluation. The standard protocol is to collect ratings of generated text, average across annotators, and rank NLG systems by their average scores. However, little consideration has been given as to whether this approach faithfully captures human preferences. Analyzing this standard protocol through the lens of utility theory in economics, we identify the implicit assumptions it makes about annotators. These assumptions are often violated in practice, in which case annotator ratings cease to reflect their preferences. The most egregious violations come from using Likert scales, which provably reverse the direction of the true preference in certain cases. We suggest improvements to the standard protocol to make it more theoretically sound, but even in its improved form, it cannot be used to evaluate open-ended tasks like story generation. For the latter, we propose a new human evaluation protocol called system-level probabilistic assessment (SPA). When human evaluation of stories is done with SPA, we can recover the ordering of GPT-3 models by size, with statistically significant results. However, when human evaluation is done with the standard protocol, less than half of the expected preferences can be recovered (e.g., there is no significant difference between curie and davinci, despite using a highly powered test).

**Towards Robust Numerical Question Answering: Diagnosing Numerical Capabilities of NLP Systems**
*Jialiang Xu, Mengyu Zhou, Xinyi He, Shi Han and Dongmei Zhang* 09:00-10:30 (Atrium)
Numerical Question Answering is the task of answering questions that require numerical capabilities. Previous works introduce general adversarial attacks to Numerical Question Answering, while not systematically exploring numerical capabilities specific to the topic. In this paper, we propose to conduct numerical capability diagnosis on a series of Numerical Question Answering systems and datasets. A series of numerical capabilities are highlighted, and corresponding dataset perturbations are designed. Empirical results indicate that existing systems are severely challenged by these perturbations. E.g., Graph2Tree experienced a 53.83% absolute accuracy drop against the "Extra" perturbation on ASDiv-a, and BART experienced 13.80% accuracy drop against the "Language" perturbation on the numerical subset of DROP. As a counteracting approach, we also investigate the effectiveness of applying perturbations as data augmentation to relieve systems' lack of robust numerical capabilities. With experiment analysis and empirical studies, it is demonstrated that Numerical Question Answering with robust numerical capabilities is still to a large extent an open question. We discuss future directions of Numerical Question Answering and summarize guidelines on future dataset collection and system design.

## Demo Session 3

09:00-10:30 (Atrium)

[DEMO] **ELEVANT: A Fully Automatic Fine-Grained Entity Linking Evaluation and Analysis Tool**
*Hannah Bast, Matthias Hertel and Natalie Prange* 09:00-10:30 (Atrium)
We present Elevant, a tool for the fully automatic fine-grained evaluation of a set of entity linkers on a set of benchmarks. Elevant provides an automatic breakdown of the performance by various error categories and by entity type. Elevant also provides a rich and compact, yet very intuitive and self-explanatory visualization of the results of a linker on a benchmark in comparison to the ground truth. A live demo, the link to the complete code base on GitHub and a link to a demo video are provided under https://elevant.cs.uni-freiburg.de .

[DEMO] **Evaluate & Evaluation on the Hub: Better Best Practices for Data and Model Measurements**
*Leandro von Werra, Lewis Tunstall, abhishek kumar thakur, Sasha Alexandra Luccioni, Tristan Thrush, Aleksandra Piktus, Felix Marty, Nazneen Rajani, Victor Mustar, Helen Ngo, Omar Sanseviero, Mario Šaško, Albert Villanova, quentin lhoest, Julien Chaumond, Margaret Mitchell, Alexander Rush, Thomas Wolf and Douwe Kiela* 09:00-10:30 (Atrium)
Evaluation is a key part of machine learning (ML), yet there is a lack of support and tooling to enable its informed and systematic practice. We introduce Evaluate and Evaluation on the Hub—a set of tools to facilitate the evaluation of models and datasets in ML. Evaluate is a library to support best practices for measurements, metrics, and comparisons of data and models. Its goal is to support reproducibility of evaluation, centralize and document the evaluation process, and broaden evaluation to cover more facets of model performance. It includes over 50 efficient canonical implementations for a variety of domains and scenarios, interactive documentation, and the ability to easily share implementations and outcomes. The library is available at https://github.com/huggingface/evaluate. In addition, we introduce Evaluation on the Hub, a platform that enables the large-scale evaluation of over 75,000 models and 11,000 datasets on the Hugging Face Hub, for free, at the click of a button. Evaluation on the Hub is available at https://huggingface.co/autoevaluate.

[DEMO] **KGxBoard: Explainable and Interactive Leaderboard for Evaluation of Knowledge Graph Completion Models**
*Haris Widjaja, Kiril Gashteovski, Wiem Ben Rim, Pengfei Liu, Christopher Malon, Daniel Ruffinelli, Carolin Lawrence and Graham Neubig*
09:00-10:30 (Atrium)
Knowledge Graphs (KGs) store information in the form of (head, predicate, tail)-triples. To augment KGs with new knowledge, researchers proposed models for KG Completion (KGC) tasks such as link prediction; i.e., answering (h; p; ?) or (?; p; t) queries. Such models are usually evaluated with averaged metrics on a held-out test set. While useful for tracking progress, averaged single-score metrics cannot reveal what exactly a model has learned or failed to learn. To address this issue, we propose KGxBoard: an interactive framework for performing fine-grained evaluation on meaningful subsets of the data, each of which tests individual and interpretable capabilities of a KGC model. In our experiments, we highlight the findings that we discovered with the use of KGxBoard, which would have been impossible to detect with standard averaged single-score metrics.

[DEMO] **SEAL: Interactive Tool for Systematic Error Analysis and Labeling**
*Nazneen Rajani, Weixin Liang, lingjiao chen, Margaret Mitchell and James Zou* 09:00-10:30 (Atrium)
With the advent of Transformers, large language models (LLMs) have saturated well-known NLP benchmarks and leaderboards with high aggregate performance. However, many times these models systematically fail on tail data or rare groups not obvious in aggregate evaluation. Identifying such problematic data groups is even more challenging when there are no explicit labels (e.g., ethnicity, gender, etc.) and further compounded for NLP datasets due to the lack of visual features to characterize failure modes (e.g., Asian males, animals indoors, waterbirds on land etc.). This paper introduces an interactive Systematic Error Analysis and Labeling (SEAL) tool that uses a two-step approach to first identify high-error slices of data and then, in the second step, introduce methods to give human-understandable semantics to those underperforming slices. We explore a variety of methods for coming up with coherent semantics for the error groups using language models for semantic labeling and a text-to-image model for generating visual features.SEAL is available at https://huggingface.co/spaces/nazneen/seal.

# Session 7 - 11:00-12:30

## Interpretability, Interactivity, and Analysis of Models for NLP 1

11:00-12:30 (Hall A, Room A)

**Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space**
*Mor Geva, Avi Caciularu, Kevin Wang and Yoav Goldberg* 11:00-11:15 (Hall A, Room A)
Transformer-based language models (LMs) are at the core of modern NLP, but their internal prediction construction process is opaque and largely not understood. In this work, we make a substantial step towards unveiling this underlying prediction process, by reverse-engineering the operation of the feed-forward network (FFN) layers, one of the building blocks of transformer models. We view the token representation as a changing distribution over the vocabulary, and the output from each FFN layer as an additive update to that distribution. Then, we analyze the FFN updates in the vocabulary space, showing that each update can be decomposed to sub-updates corresponding to single FFN parameter vectors, each promoting concepts that are often human-interpretable. We then leverage these findings for controlling LM predictions, where we reduce the toxicity of GPT2 by almost 50%, and for improving computation efficiency with a simple early exit rule, saving 20% of computation on average.

**Interpreting Language Models with Contrastive Explanations**
*Kayo Yin and Graham Neubig* 11:15-11:30 (Hall A, Room A)
Model interpretability methods are often used to explain NLP model decisions on tasks such as text classification, where the output space is relatively small. However, when applied to language generation, where the output space often consists of tens of thousands of tokens, these methods are unable to provide informative explanations. Language models must consider various features to predict a token, such as its part of speech, number, tense, or semantics. Existing explanation methods conflate evidence for all these features into a single explanation, which is less interpretable for human understanding.

To disentangle the different decisions in language modeling, we focus on explaining language models contrastively: we look for salient input tokens that explain why the model predicted one token instead of another. We demonstrate that contrastive explanations are quantifiably better than non-contrastive explanations in verifying major grammatical phenomena, and that they significantly improve contrastive model simulatability for human observers. We also identify groups of contrastive decisions where the model uses similar evidence, and we are able to characterize what input tokens models use during various language generation decisions.

**Balanced Adversarial Training: Balancing Tradeoffs between Fickleness and Obstinacy in NLP Models**
*Hannah Chen, Yangfeng Ji and David Evans* 11:30-11:45 (Hall A, Room A)
Traditional (fickle) adversarial examples involve finding a small perturbation that does not change an input's true label but confuses the classifier into outputting a different prediction. Conversely, obstinate adversarial examples occur when an adversary finds a small perturbation that preserves the classifier's prediction but changes the true label of an input. Adversarial training and certified robust training have shown some effectiveness in improving the robustness of machine learnt models to fickle adversarial examples. We show that standard adversarial training methods focused on reducing vulnerability to fickle adversarial examples may make a model more vulnerable to obstinate adversarial examples, with experiments for both natural language inference and paraphrase identification tasks. To counter this phenomenon, we introduce Balanced Adversarial Training, which incorporates contrastive learning to increase robustness against both fickle and obstinate adversarial examples.

**DropMix: A Textual Data Augmentation Combining Dropout with Mixup**
*Fanshuang Kong, Richong Zhang, Xiaohui Guo, Samuel Mensah and Yongyi Mao* 11:45-12:00 (Hall A, Room A)
Overfitting is a notorious problem when there is insufficient data to train deep neural networks in machine learning tasks. Data augmentation regularization methods such as Dropout, Mixup, and their enhanced variants are effective and prevalent, and achieve promising performance to overcome overfitting. However, in text learning, most of the existing regularization approaches merely adopt ideas from computer vision without considering the importance of dimensionality in natural language processing. In this paper, we argue that the property is essential to overcome overfitting in text learning. Accordingly, we present a saliency map informed textual data augmentation and regularization framework, which combines Dropout and Mixup, namely DropMix, to mitigate the overfitting problem in text learning. In addition, we design a procedure that drops and patches fine grained shapes of the saliency map under the DropMix framework to enhance regularization. Empirical studies confirm the effectiveness of the proposed approach on 12 text classification tasks.

**"Will You Find These Shortcuts?" A Protocol for Evaluating the Faithfulness of Input Salience Methods for Text Classification**
*Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm and Katja Filippova* 12:00-12:15 (Hall A, Room A)
Feature attribution a.k.a. input salience methods which assign an importance score to a feature are abundant but may produce surprisingly different results for the same model on the same input. While differences are expected if disparate definitions of importance are assumed, most methods claim to provide faithful attributions and point at the features most relevant for a model's prediction. Existing work on faithfulness evaluation is not conclusive and does not provide a clear answer as to how different methods are to be compared. Focusing on text classification and the model debugging scenario, our main contribution is a protocol for faithfulness evaluation that makes use of partially synthetic data to obtain ground truth for feature importance ranking. Following the protocol, we do an in-depth analysis of four standard salience method classes on a range of datasets and lexical shortcuts for BERT and LSTM models. We demonstrate that some of the most popular method configurations provide poor results even for simple shortcuts while a method judged to be too simplistic works remarkably well for BERT.

**On the Transformation of Latent Space in Fine-Tuned NLP Models**
*Nadir Durrani, Hassan Sajjad, Fahim Dalvi and Firoj Alam* 12:15-12:30 (Hall A, Room A)
We study the evolution of latent space in fine-tuned NLP models. Different from the commonly used probing-framework, we opt for an unsupervised method to analyze representations. More specifically, we discover latent concepts in the representational space using hierarchical clustering. We then use an alignment function to gauge the similarity between the latent space of a pre-trained model and its fine-tuned version. We use traditional linguistic concepts to facilitate our understanding and also study how the model space transforms towards task-specific information. We perform a thorough analysis, comparing pre-trained and fine-tuned models across three models and three downstream tasks. The notable findings of our work are: i) the latent space of the higher layers evolve towards task-specific concepts, ii) whereas the lower layers retain generic concepts acquired in the pre-trained model, iii) we discovered that some concepts in the higher layers acquire polarity towards the output class, and iv) that these concepts can be used for generating adversarial triggers.

## Machine Learning for NLP

11:00-12:30 (Hall A, Room B)

**Backdoor Attacks in Federated Learning by Rare Embeddings and Gradient Ensembling**
*Ki Yoon Yoo and Nojun Kwak* 11:00-11:15 (Hall A, Room B)
Recent advances in federated learning have demonstrated its promising capability to learn on decentralized datasets. However, a considerable amount of work has raised concerns due to the potential risks of adversaries participating in the framework to poison the global model for an adversarial purpose. This paper investigates the feasibility of model poisoning for backdoor attacks through rare word embeddings of NLP models. In text classification, less than 1% of adversary clients suffices to manipulate the model output without any drop in the performance of clean sentences. For a less complex dataset, a mere 0.1% of adversary clients is enough to poison the global model effectively. We also propose a technique specialized in the federated learning scheme called gradient ensemble, which enhances the backdoor performance in all experimental settings.

**When Can Transformers Ground and Compose: Insights from Compositional Generalization Benchmarks**
*Ankur Sikarwar, Arkil Patel and Navin Goyal* 11:15-11:30 (Hall A, Room B)
Humans can reason compositionally whilst grounding language utterances to the real world. Recent benchmarks like ReaSCAN (Wu et al., 2021) use navigation tasks grounded in a grid world to assess whether neural models exhibit similar capabilities. In this work, we present a simple transformer-based model that outperforms specialized architectures on ReaSCAN and a modified version (Qiu et al., 2021) of gSCAN (Ruis et al., 2020). On analyzing the task, we find that identifying the target location in the grid world is the main challenge for the models. Furthermore, we show that a particular split in ReaSCAN, which tests depth generalization, is unfair. On an amended version of this split, we show that transformers can generalize to deeper input structures. Finally, we design a simpler grounded compositional generalization task, RefEx, to investigate how transformers reason compositionally. We show that a single self-attention layer with a single head generalizes to novel combinations of object attributes. Moreover, we derive a precise mathematical construction of the transformer's computations from the learned network. Overall, we provide valuable insights about the grounded compositional generalization task and the behaviour of transformers on it, which would be useful for researchers working in this area.

**GammaE: Gamma Embeddings for Logical Queries on Knowledge Graphs**
*Dong Yang, Peijun Qing, Yang Li, Haonan Lu and xiaodong lin* 11:30-11:45 (Hall A, Room B)
Embedding knowledge graphs (KGs) for multi-hop logical reasoning is a challenging problem due to massive and complicated structures in many KGs. Recently, many promising works projected entities and queries into a geometric space to efficiently find answers. However, it remains challenging to model the negation and union operator. The negation operator has no strict boundaries, which generates overlapped embeddings and leads to obtaining ambiguous answers. An additional limitation is that the union operator is non-closure, which undermines the model to handle a series of union operators. To address these problems, we propose a novel probabilistic embedding model, namely Gamma Embeddings (GammaE), for encoding entities and queries to answer different types of FOL queries on KGs. We utilize the linear property and strong boundary support of the Gamma distribution to capture more features of entities and queries, which dramatically reduces model uncertainty. Furthermore, GammaE implements the Gamma mixture method to design the closed union operator. The performance of GammaE is validated on three large logical query datasets. Experimental results show that GammaE significantly outperforms state-of-the-art models on public benchmarks.

**Numerical Optimizations for Weighted Low-rank Estimation on Language Models**
*Ting Hua, Yen-Chang Hsu, Felicity Wang, Qian Lou, Yilin Shen and Hongxia Jin* 11:45-12:00 (Hall A, Room B)
Singular value decomposition (SVD) is one of the most popular compression methods that approximate a target matrix with smaller matrices. However, standard SVD treats the parameters within the matrix with equal importance, which is a simple but unrealistic assumption. The parameters of a trained neural network model may affect the task performance unevenly, which suggests non-equal importance among the parameters. Compared to SVD, the decomposition method aware of parameter importance is the more practical choice in real cases. Unlike standard SVD, weighed value decomposition is a non-convex optimization problem that lacks a closed-form solution. We systematically investigated multiple optimization strategies to tackle the problem and examined our method by compressing Transformer-based language models. Further, we designed a metric to predict when the SVD may introduce a significant performance drop, for which our method can be a rescue strategy. The extensive evaluations demonstrate that our method can perform better than current SOTA methods in compressing Transformer-based language models.

**Efficient Nearest Neighbor Search for Cross-Encoder Models using Matrix Factorization**
*Nishant Yadav, Nicholas Monath, Rico Angell, Manzil Zaheer and Andrew McCallum* 12:00-12:15 (Hall A, Room B)
Efficient k-nearest neighbor search is a fundamental task, foundational for many problems in NLP. When the similarity is measured by dot-product between dual-encoder vectors or L2-distance, there already exist many scalable and efficient search methods. But not so when similarity is measured by more accurate and expensive black-box neural similarity models, such as cross-encoders, which jointly encode the query and candidate neighbor. The cross-encoders' high computational cost typically limits their use to reranking candidates retrieved by a cheaper model, such as dual encoder or TF-IDF. However, the accuracy of such a two-stage approach is upper-bounded by the recall of the initial candidate set, and potentially requires additional training to align the auxiliary retrieval model with the cross-encoder model. In this paper, we present an approach that avoids the use of a dual-encoder for retrieval, relying solely on the cross-encoder. Retrieval is made efficient with CUR decomposition, a matrix decomposition approach that approximates all pairwise cross-encoder distances from a small subset of rows and columns of the distance matrix. Indexing items using our approach is computationally cheaper than training an auxiliary dual-encoder model through distillation. Empirically, for $k > 10$, our approach provides test-time recall-vs-computational cost trade-offs superior to the current widely-used methods that re-rank items retrieved using a dual-encoder or TF-IDF.

**Large language models are few-shot clinical information extractors**
*Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim and David Sontag* 12:15-12:30 (Hall A, Room B)
A long-running goal of the clinical NLP community is the extraction of important variables trapped in clinical notes. However, roadblocks have included dataset shift from the general domain and a lack of public clinical corpora and annotations. In this work, we show that large language models, such as InstructGPT (Ouyang et al., 2022), perform well at zero- and few-shot information extraction from clinical text despite not being trained specifically for the clinical domain. Whereas text classification and generation performance have already been studied extensively in such models, here we additionally demonstrate how to leverage them to tackle a diverse set of NLP tasks which require more structured outputs, including span identification, token-level sequence classification, and relation extraction. Further, due to the dearth of available data to evaluate these systems, we introduce new datasets for benchmarking few-shot clinical information extraction based on a manual re-annotation of the CASI dataset (Moon et al., 2014) for new tasks. On the clinical extraction tasks we studied, the GPT-3 systems significantly outperform existing zero- and few-shot baselines.

## Resources and Evaluation 2

11:00-12:30 (Hall A, Room C)

**StoryER: Automatic Story Evaluation via Ranking, Rating and Reasoning**
*Hong Chen, Duc Vo, Hiroya Takamura, Yusuke Miyao and Hideki Nakayama* 11:00-11:15 (Hall A, Room C)
Existing automatic story evaluation methods place a premium on story lexical level coherence, deviating from human preference. We go beyond this limitation by considering a novel Story Evaluation method that mimics human preference when judging a story, namely StoryER, which consists of three sub-tasks: Ranking, Rating and Reasoning. Given either a machine-generated or a human-written story, StoryER requires the machine to output 1) a preference score that corresponds to human preference, 2) specific ratings and their corresponding confidences and 3) comments for various aspects (e.g., opening, character-shaping). To support these tasks, we introduce a well-annotated dataset comprising (i) 100k ranked story pairs; and (ii) a set of 46k ratings and comments on various aspects of the story. We finetune Longformer-Encoder-Decoder (LED) on the collected dataset, with the encoder responsible for preference score and aspect prediction and the decoder for comment generation. Our comprehensive experiments result a competitive benchmark for each task, showing the high correlation to human preference. In addition, we have witnessed the joint learning of the preference scores, the aspect ratings, and the comments brings gain each single task. Our dataset and benchmarks are publicly available to advance the research of story evaluation tasks.

**Linguistic Corpus Annotation for Automatic Text Simplification Evaluation**
*Rémi Cardon, Adrien Bibal, Rodrigo Wilkens, David Alfter, Magali Norré, Adeline Müller, Watrin Patrick and Thomas François* 11:15-11:30 (Hall A, Room C)
Evaluating automatic text simplification (ATS) systems is a difficult task that is either performed by automatic metrics or user-based evaluations. However, from a linguistic point-of-view, it is not always clear on what bases these evaluations operate. In this paper, we propose annotations of the ASSET corpus that can be used to shed more light on ATS evaluation. In addition to contributing with this resource, we show how it can be used to analyze SARI's behavior and to re-evaluate existing ATS systems. We present our insights as a step to improve ATS evaluation protocols in the future.

**Near-Negative Distinction: Giving a Second Life to Human Evaluation Datasets**
*Philippe Laban, Chien-Sheng Wu, Wenhao Liu and Caiming Xiong* 11:30-11:45 (Hall A, Room C)
Precisely assessing the progress in natural language generation (NLG) tasks is challenging, and human evaluation to establish a preference in a model's output over another is often necessary. However, human evaluation is usually costly, difficult to reproduce, and non-reusable. In this paper, we propose a new and simple automatic evaluation method for NLG called Near-Negative Distinction (NND) that repurposes prior human annotations into NND tests. In an NND test, an NLG model must place a higher likelihood on a high-quality output candidate than on a near-negative candidate with a known error. Model performance is established by the number of NND tests a model passes, as well as the distribution over task-specific errors the model fails on. Through experiments on three NLG tasks (question generation, question answering, and summarization), we show that NND achieves a higher correlation with human judgments than standard NLG evaluation metrics. We then illustrate NND evaluation in four practical scenarios, for example performing fine-grain model analysis, or studying model training dynamics. Our findings suggest that NND can give a second life to human annotations and provide low-cost NLG evaluation.

**Stanceosaurus: Classifying Stance Towards Multicultural Misinformation**
*Jonathan Zheng, Ashutosh Baheti, Tarek Naous, Wei Xu and Alan Ritter* 11:45-12:00 (Hall A, Room C)
We present Stanceosaurus, a new corpus of 28,033 tweets in English, Hindi and Arabic annotated with stance towards 250 misinformation claims. As far as we are aware, it is the largest corpus annotated with stance towards misinformation claims. The claims in Stanceosaurus originate from 15 fact-checking sources that cover diverse geographical regions and cultures. Unlike existing stance datasets, we introduce a more fine-grained 5-class labeling strategy with additional subcategories to distinguish implicit stance. Pre-trained transformer-based stance classifiers that are fine-tuned on our corpus show good generalization on unseen claims and regional claims from countries outside the training data. Cross-lingual experiments demonstrate Stanceosaurus' capability of training multilingual models, achieving 53.1 F1 on Hindi and 50.4 F1 on Arabic without any target-language fine-tuning. Finally, we show how a domain adaptation method can be used to improve performance on Stanceosaurus using additional RumourEval-2019 data. We will make Stanceosaurus publicly available to the research community upon publication and hope it will encourage further work on misinformation identification across languages and cultures.

**When FLUE Meets FLANG: Benchmarks and Large Pretrained Language Model for Financial Domain**
*Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen and Diyi Yang* 12:00-12:15 (Hall A, Room C)
Pre-trained language models have shown impressive performance on a variety of tasks and domains. Previous research on financial language models usually employs a generic training scheme to train standard model architectures, without completely leveraging the richness of the financial data. We propose a novel domain specific Financial LANGuage model (FLANG) which uses financial keywords and phrases for better masking, together with span boundary objective and in-filing objective. Additionally, the evaluation benchmarks in the field have been limited. To this end, we contribute the Financial Language Understanding Evaluation (FLUE), an open-source comprehensive suite of benchmarks for the financial domain. These include new benchmarks across 5 NLP tasks in financial domain as well as common benchmarks used in the previous research. Experiments on these benchmarks suggest that our model outperforms those in prior literature on a variety of NLP tasks. Our models, code and benchmark data will be made publicly available on Github and Huggingface.

**Reproducibility in Computational Linguistics: Is Source Code Enough?**
*Mohammad Arvan, Luís Pina and Natalie Parde* 12:15-12:30 (Hall A, Room C)
The availability of source code has been put forward as one of the most critical factors for improving the reproducibility of scientific research. This work studies trends in source code availability at major computational linguistics conferences, namely, ACL, EMNLP, LREC, NAACL, and COLING. We observe positive trends, especially in conferences that actively promote reproducibility. We follow this by conducting a reproducibility study of eight papers published in EMNLP 2021, finding that source code releases leave much to be desired. Moving forward, we suggest all conferences require self-contained artifacts and provide a venue to evaluate such artifacts at the time of publication. Authors can include small-scale experiments and explicit scripts to generate each result to improve the reproducibility of their work.

## Theme Track & CL & Short Papers

11:00-12:30 (Hall A, Room D)

**Towards Climate Awareness in NLP Research**
*Daniel Hershcovich, Nicolas Webersinke, Mathias Kraus, Julia Bingler and Markus Leippold* 11:00-11:12 (Hall A, Room D)
The climate impact of AI, and NLP research in particular, has become a serious issue given the enormous amount of energy that is increasingly being used for training and running computational models. Consequently, increasing focus is placed on efficient NLP. However, this important initiative lacks simple guidelines that would allow for systematic climate reporting of NLP research. We argue that this deficiency is one of the reasons why very few publications in NLP report key figures that would allow a more thorough examination of environmental impact, and present a quantitative survey to demonstrate this. As a remedy, we propose a climate performance model card with the primary purpose of being practically usable with only limited information about experiments and the underlying computer hardware. We describe why this step is essential to increase awareness about the environmental impact of NLP research and, thereby, paving the way for more thorough discussions.

**Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection**
*Suchin Gururangan, Dallas Card, Sarah Dreier, Emily Gade, Leroy Wang, Zeyu Wang, Luke Zettlemoyer and Noah A. Smith* 11:12-11:24 (Hall A, Room D)
Language models increasingly rely on massive web crawls for diverse text data. However, these sources are rife with undesirable content. As such, resources like Wikipedia, books, and news often serve as anchors for automatically selecting web text most suitable for language modeling, a process typically referred to as quality filtering. Using a new dataset of U.S. high school newspaper articles—written by students from across the country—we investigate whose language is preferred by the quality filter used for GPT-3. We find that newspapers from larger schools, located in wealthier, educated, and urban zones (ZIP codes) are more likely to be classified as high quality. We also show that this quality measurement is unaligned with other sensible metrics, such as factuality or literary acclaim. We argue that privileging any corpus as high quality entails a language ideology, and more care is needed to construct training corpora for language models, with better transparency and justification for the inclusion or exclusion of various texts.

**Geographic Citation Gaps in NLP Research**
*Mukund Rungta, Janvijay Singh, Saif M. Mohammad and Diyi Yang* 11:24-11:36 (Hall A, Room D)
In a fair world, people have equitable opportunities to education, to conduct scientific research, to publish, and to get credit for their work, regardless of where they live. However, it is common knowledge among researchers that a vast number of papers accepted at top NLP venues come from a handful of western countries and (lately) China; whereas, very few papers from Africa and South America get published. Similar disparities are also believed to exist for paper citation counts. In the spirit of "what we do not measure, we cannot improve", this work asks a series of questions on the relationship between geographical location and publication success (acceptance in top NLP venues and citation impact). We first created a dataset of 70,000 papers from the ACL Anthology, extracted their meta-information, and generated their citation network. We then show that not only are there substantial geographical disparities in paper acceptance and citation but also that these disparities persist even when controlling for a number of variables such as venue of publication and sub-field of NLP. Further, despite some steps taken by the NLP community to improve geographical diversity, we show that the disparity in publication metrics across locations is still on an increasing trend since the early 2000s. We release our code and dataset here: https://github.com/iamjanvijay/acl-cite-net

[CL] **Information Theory-based Compositional Distributional Semantics**
*Enrique Amigó, Alejandro Ariza-Casabona, Víctor Fresno and M. Antònia Martí* 11:36-11:48 (Hall A, Room D)
In the context of text representation, Compositional Distributional Semantics models aim to fuse the Distributional Hypothesis and the Principle of Compositionality. Text embedding is based on coocurrence distributions and the representations are in turn combined by compositional functions taking into account the text structure. However, the theoretical basis of compositional functions is still an open issue. In this paper we define and study the notion of Information Theory-based Compositional Distributional Semantics (ICDS): i) We first establish formal properties for embedding, composition and similarity functions based on Shannon's Information Theory; ii) we analyse the existing approaches under this prism, checking whether or not they comply with the established desirable properties; iii) we propose two parameterisable composition and similarity functions that generalise traditional approaches while fulfilling the formal properties; and finally iv) we perform an empirical study on several textual similarity datasets that include sentences with a high and low lexical overlap, and on the similarity between words and their description. Our theoretical analysis and empirical results show that fulfilling formal properties affects positively the accuracy of text representation models in terms of correspondence (isometry) between the embedding and meaning spaces.

**Extracted BERT Model Leaks More Information than You Think!**
*Xuanli He, Lingjuan Lyu, Chen Chen and Qiongkai Xu* 11:48-12:00 (Hall A, Room D)
The collection and availability of big data, combined with advances in pre-trained models (e.g. BERT), have revolutionized the predictive performance of natural language processing tasks. This allows corporations to provide machine learning as a service (MLaaS) by encapsulating fine-tuned BERT-based models as APIs. Due to significant commercial interest, there has been a surge of attempts to steal remote services via model extraction. Although previous works have made progress in defending against model extraction attacks, there has been little discussion on their performance in preventing privacy leakage. This work bridges this gap by launching an attribute inference attack against the extracted BERT model. Our extensive experiments reveal that model extraction can cause severe privacy leakage even when victim models are facilitated with state-of-the-art defensive strategies.

**Exploiting domain-slot related keywords description for Few-Shot Cross-Domain Dialogue State Tracking**
*Gao Qixiang, Guanting Dong, Yutao Mou, Liwen Wang, Chen Zeng, Daichi Guo, Mingyang Sun and Weiran Xu* 12:00-12:12 (Hall A, Room D)
Collecting dialogue data with domain-slot-value labels for dialogue state tracking (DST) could be a costly process. In this paper, we propose a novel framework based on domain-slot related description to tackle the challenge of few-shot cross-domain DST. Specifically, we design an extraction module to extract domain-slot related verbs and nouns in the dialogue. Then, we integrates them into the description, which aims to prompt the model to identify the slot information. Furthermore, we introduce a random sampling strategy to improve the domain generalization ability of the model. We utilize a pre-trained model to encode contexts and description and generates answers with an auto-regressive manner. Experimental results show that our approaches substantially outperform the existing few-shot DST methods on MultiWOZ and gain strong improvements on the slot accuracy comparing to existing slot description methods.

**PRINCE: Prefix-Masked Decoding for Knowledge Enhanced Sequence-to-Sequence Pre-Training**
*Song Xu, Haoran Li, Peng Yuan, Youzheng Wu and Xiaodong He* 12:12-12:24 (Hall A, Room D)
Pre-trained Language Models (PLMs) have shown effectiveness in various Natural Language Processing (NLP) tasks. Denoising autoencoder is one of the most successful pre-training frameworks, learning to recompose the original text given a noise-corrupted one. The existing studies mainly focus on injecting noises into the input. This paper introduces a simple yet effective pre-training paradigm, equipped with a knowledge-enhanced decoder that predicts the next entity token with noises in the prefix, explicitly strengthening the representation learning of entities that span over multiple input tokens. Specifically, when predicting the next token within an entity, we feed masks into the prefix in place of some of the previous ground-truth tokens that constitute the entity. Our model achieves new state-of-the-art results on two knowledge-driven data-to-text generation tasks with up to 2% BLEU gains.

## Information Extraction 1

11:00-12:30 (Hall B)

**Transfer Learning from Semantic Role Labeling to Event Argument Extraction with Template-based Slot Querying**
*Zhisong Zhang, Emma Strubell and Eduard Hovy* 11:00-11:15 (Hall B)
In this work, we investigate transfer learning from semantic role labeling (SRL) to event argument extraction (EAE), considering their similar argument structures. We view the extraction task as a role querying problem, unifying various methods into a single framework. There are key discrepancies on role labels and distant arguments between semantic role and event argument annotations. To mitigate these discrepancies, we specify natural language-like queries to tackle the label mismatch problem and devise argument augmentation to recover distant arguments. We show that SRL annotations can serve as a valuable resource for EAE, and a template-based slot querying strategy is especially effective for facilitating the transfer. In extensive evaluations on two English EAE benchmarks, our proposed model obtains impressive zero-shot results by leveraging SRL annotations, reaching nearly 80% of the fullysupervised scores. It further provides benefits in low-resource cases, where few EAE annotations are available. Moreover, we show that our approach generalizes to cross-domain and multilingual scenarios.

**Generative Knowledge Graph Construction: A Review**
*Hongbin Ye, Ningyu Zhang, Hui Chen and Huajun Chen* 11:15-11:30 (Hall B)
Generative Knowledge Graph Construction (KGC) refers to those methods that leverage the sequence-to-sequence framework for building knowledge graphs, which is flexible and can be adapted to widespread tasks. In this study, we summarize the recent compelling progress in generative knowledge graph construction. We present the advantages and weaknesses of each paradigm in terms of different generation targets and provide theoretical insight and empirical analysis. Based on the review, we suggest promising research directions for the future. Our contributions are threefold: (1) We present a detailed, complete taxonomy for the generative KGC methods; (2) We provide a theoretical and empirical analysis of the generative KGC methods; (3) We propose several research directions that can be developed in the future.

**Graph-based Model Generation for Few-Shot Relation Extraction**
*Wanli Li and Tieyun Qian* 11:30-11:45 (Hall B)
Few-shot relation extraction (FSRE) has been a challenging problem since it only has a handful of training instances. Existing models follow a 'one-for-all' scheme where one general large model performs all individual N-way-K-shot tasks in FSRE, which prevents the model from achieving the optimal point on each task. In view of this, we propose a model generation framework that consists of one general model for all tasks and many tiny task-specific models for each individual task. The general model generates and passes the universal knowledge to the tiny models which will be further fine-tuned when performing specific tasks. In this way, we decouple the complexity of the entire task space from that of all individual tasks while absorbing the universal knowledge. Extensive experimental results on two public datasets demonstrate that our framework reaches a new state-of-the-art performance for FRSE tasks. Our code is available at: https://github.com/NLPWM-WHU/GM_GEN.

**A Good Neighbor, A Found Treasure: Mining Treasured Neighbors for Knowledge Graph Entity Typing**
*Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu and Jun Zhao* 11:45-12:00 (Hall B)
The task of knowledge graph entity typing (KGET) aims to infer the missing types for entities in knowledge graphs. Some pioneering work has proved that neighbor information is very important for the task. However, existing methods only leverage the one-hop neighbor information of the central entity, ignoring the multi-hop neighbor information that can provide valuable clues for inference. Besides, we also observe that there are co-occurrence relations between types, which is very helpful to alleviate false-negative problem. In this paper, we propose a novel method called Mining Treasured Neighbors (MiNer) to make use of these two characteristics. Firstly, we devise a Neighbor Information Aggregation module to aggregate the neighbor information. Then, we propose an Entity Type Inference module to mitigate the adverse impact of the irrelevant neighbor information. Finally, a Type Co-occurrence Regularization module is designed to prevent the model from overfitting the false negative examples caused by missing types. Experimental results on two widely used datasets indicate that our approach significantly outperforms previous state-of-the-art methods.

**ReSel: N-ary Relation Extraction from Scientific Text and Tables by Learning to Retrieve and Select**
*Yuchen Zhuang, Yinghao Li, Junyang Zhang, Yue Yu, Yingjun Mou, Xiang Chen, Le Song and Chao Zhang* 12:00-12:15 (Hall B)
We study the problem of extracting N-ary relation tuples from scientific articles. This task is challenging because the target knowledge tuples can reside in multiple parts and modalities of the document. Our proposed method ReSel decomposes this task into a two-stage procedure that first retrieves the most relevant paragraph/table and then selects the target entity from the retrieved component. For the high-level retrieval stage, ReSel designs a simple and effective feature set, which captures multi-level lexical and semantic similarities between the query and components. For the low-level selection stage, ReSel designs a cross-modal entity correlation graph along with a multi-view architecture, which models both semantic and document-structural relations between entities. Our experiments on three scientific information extraction datasets show that ReSel outperforms state-of-the-art baselines significantly.

**MAVEN-ERE: A Unified Large-scale Dataset for Event Coreference, Temporal, Causal, and Subevent Relation Extraction**
*Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li and Jie Zhou*
12:15-12:30 (Hall B)
The diverse relationships among real-world events, including coreference, temporal, causal, and subevent relations, are fundamental to understanding natural languages. However, two drawbacks of existing datasets limit event relation extraction (ERE) tasks: (1) Small scale. Due to the annotation complexity, the data scale of existing datasets is limited, which cannot well train and evaluate data-hungry models. (2) Absence of unified annotation. Different types of event relations naturally interact with each other, but existing datasets only cover limited relation types at once, which prevents models from taking full advantage of relation interactions. To address these issues, we construct a unified large-scale human-annotated ERE dataset MAVEN-ERE with improved annotation schemes. It contains 103,193 event coreference chains, 1,216,217 temporal relations, 57,992 causal relations, and 15,841 subevent relations, which is larger than existing datasets of all the ERE tasks by at least an order of magnitude. Experiments show that ERE on MAVEN-ERE is quite challenging, and considering relation interactions with joint learning can improve performances. The dataset and source codes can be obtained from https://github.com/THU-KEG/MAVEN-ERE.

## CL & TACL 2

11:00-12:30 (Collaboratorium)

[TACL] **Naturalistic Causal Probing for Morpho-Syntax**

*Afra Amini, Tiago Pimentel, Clara Meister and Ryan Cotterell* 11:00-11:15 (Collaboratorium)
Naturalistic Causal Probing for Morpho-Syntax Final paper abstract: Probing has become a go-to methodology for interpreting and analyzing deep neural models in natural language processing. However, there is still a lack of understanding of the limitations and weaknesses of various types of probes. In this work, we suggest a strategy for input-level intervention on naturalistic sentences. Using our approach, we intervene on the morpho-syntactic features of a sentence, while keeping the rest of the sentence unchanged. Such an intervention allows us to causally probe pre-trained models. We apply our naturalistic causal probing framework to analyze the effects of grammatical gender and number on contextualized representations extracted from three pre-trained models in Spanish, the multilingual versions of BERT, Roberta, and GPT-2. Our experiments suggest that naturalistic interventions lead to stable estimates of the causal effects of various linguistic properties. Moreover, our experiments demonstrate the importance of naturalistic causal probing when analyzing pre-trained models.

[CL] **Transformers and the representation of biomedical background knowledge**
*Oskar Wysocki, Zili Zhou, Paul O'Regan, Deborah Ferreira, Magdalena Wysocka, Dónal Landers and André Freitas* 11:15-11:30 (Collaboratorium)
Specialized transformers-based models (such as BioBERT and BioMegatron) are adapted for the biomedical domain based on publicly available biomedical corpora. As such, they have the potential to encode large-scale biological knowledge. We investigate the encoding and representation of biological knowledge in these models, and its potential utility to support inference in cancer precision medicine—namely, the interpretation of the clinical significance of genomic alterations. We compare the performance of different transformer baselines; we use probing to determine the consistency of encodings for distinct entities; and we use clustering methods to compare and contrast the internal properties of the embeddings for genes, variants, drugs, and diseases. We show that these models do indeed encode biological knowledge, although some of this is lost in fine-tuning for specific tasks. Finally, we analyze how the models behave with regard to biases and imbalances in the dataset.

[TACL] **Diff-Explainer: Differentiable Convex Optimization for Explainable Multi-hop Inference**
*Mokanarangan Thayaparan, Marco Valentino, Deborah Ferreira, Julia Rozanova and André Freitas* 11:30-11:45 (Collaboratorium)
This paper presents Diff-Explainer, the first hybrid framework for explainable multi-hop inference that integrates explicit constraints with neural architectures through differentiable convex optimization. Specifically, Diff-Explainer allows for the fine-tuning of neural representations within a constrained optimization framework to answer and explain multi-hop questions in natural language. To demonstrate the efficacy of the hybrid framework, we combine existing ILP-based solvers for multi-hop Question Answering (QA) with Transformer-based representations. An extensive empirical evaluation on scientific and commonsense QA tasks demonstrates that the integration of explicit constraints in a end-to-end differentiable framework can significantly improve the performance of non-differentiable ILP solvers (8.91% - 13.3%). Moreover, additional analysis reveals that Diff-Explainer is able to achieve strong performance when compared to standalone Transformers and previous multi-hop approaches while still providing structured explanations in support of its predictions.

[TACL] **Learning Fair Representations via Rate-Distortion Maximization**
*Somnath Basu Roy Chowdhury and Snigdha Chaturvedi* 11:45-12:00 (Collaboratorium)
Text representations learned by machine learning models often encode undesirable demographic information of the user. Predictive models based on these representations can rely on such information, resulting in biased decisions. We present a novel debiasing technique, Fairness-aware Rate Maximization (FaRM), that removes protected information by making representations of instances belonging to the same protected attribute class uncorrelated, using the rate-distortion function. FaRM is able to debias representations with or without a target task at hand. FaRM can also be adapted to remove information about multiple protected attributes simultaneously. Empirical evaluations show that FaRM achieves state-of-the-art performance on several datasets, and learned representations leak significantly less protected attribute information against an attack by a non-linear probing network.

[CL] **It Takes Two Flints to Make a Fire: Multitask Learning of Neural Relation and Explanation Classifiers**
*Zheng Tang and Mihai Surdeanu* 12:00-12:15 (Collaboratorium)
We propose an explainable approach for relation extraction that mitigates the tension between generalization and explainability by jointly training for the two goals. Our approach uses a multi-task learning architecture, which jointly trains a classifier for relation extraction, and a sequence model that labels words in the context of the relation that explain the decisions of the relation classifier. We also convert the model outputs to rules to bring global explanations to this approach. This sequence model is trained using a hybrid strategy: supervised, when supervision from pre-existing patterns is available, and semi-supervised otherwise. In the latter situation, we treat the sequence model's labels as latent variables, and learn the best assignment that maximizes the performance of the relation classifier. We evaluate the proposed approach on the two datasets and show that the sequence model provides labels that serve as accurate explanations for the relation classifier's decisions, and, importantly, that the joint training generally improves the performance of the relation classifier. We also evaluate the performance of the generated rules and show that the new rules are great add-on to the manual rules and bring the rule-based system much closer to the neural models.

[TACL] **Meta-Learning the Difference: Preparing Large Language Models for Efficient Adaptation**
*Zejiang Hou, Julian Salazar and George Polovets* 12:15-12:30 (Collaboratorium)
Large pretrained language models (PLMs) are often domain- or task-adapted via finetuning or prompting. Finetuning requires modifying all of the parameters and having enough data to avoid overfitting while prompting requires no training and few examples but limits performance. Instead, we prepare PLMs for data- and parameter-efficient adaptation by learning to learn the difference between general and adapted PLMs. This difference is expressed in terms of model weights and sublayer structure through our proposed dynamic low-rank reparameterization and learned architecture controller. Experiments on few-shot dialogue completion, low-resource abstractive summarization, and multi-domain language modeling show improvements in adaptation time and performance over direct finetuning or preparation via domain-adaptive pre-training. Ablations show our task-adaptive reparameterization (TARP) and model search (TAMS) components individually improve on other parameter-efficient transfer like adapters and structure-learning methods like learned sparsification.

## Poster Sessions 9 & 10

11:00-12:30 (Atrium)

**Retrieval Augmentation for Commonsense Reasoning: A Unified Approach**
*Wenhao Yu, Chenguang Zhu, Zhihan Zhang, Shuohang Wang, Zhuosheng Zhang, Yuwei Fang and Meng Jiang* 11:00-12:30 (Atrium)
A common thread of retrieval-augmented methods in the existing literature focuses on retrieving encyclopedic knowledge, such as Wikipedia, which facilitates well-defined entity and relation spaces that can be modeled. However, applying such methods to commonsense reasoning tasks faces two unique challenges, i.e., the lack of a general large-scale corpus for retrieval and a corresponding effective commonsense

retriever. In this paper, we systematically investigate how to leverage commonsense knowledge retrieval to improve commonsense reasoning tasks. We proposed a unified framework of retrieval-augmented commonsense reasoning (called RACo), including a newly constructed commonsense corpus with over 20 million documents and novel strategies for training a commonsense retriever. We conducted experiments on four different commonsense reasoning tasks. Extensive evaluation results showed that our proposed RACo can significantly outperform other knowledge-enhanced method counterparts, achieving new SoTA performance on the CommonGen and CREAK leaderboards.

**Rainier: Reinforced Knowledge Introspector for Commonsense Question Answering**
*Jiacheng Liu, Skyler Hallinan, Ximing Lu, Pengfei He, Sean Welleck, Hannaneh Hajishirzi and Yejin Choi* 11:00-12:30 (Atrium)
Knowledge underpins reasoning. Recent research demonstrates that when relevant knowledge is provided as additional context to commonsense question answering (QA), it can substantially enhance the performance even on top of state-of-the-art. The fundamental challenge is where and how to find such knowledge that is high quality and on point with respect to the question; knowledge retrieved from knowledge bases are incomplete and knowledge generated from language models are inconsistent.

We present Rainier, or Reinforced Knowledge Introspector, that learns to generate contextually relevant knowledge in response to given questions. Our approach starts by imitating knowledge generated by GPT-3, then learns to generate its own knowledge via reinforcement learning where rewards are shaped based on the increased performance on the resulting question answering. Rainier demonstrates substantial and consistent performance gains when tested over 9 different commonsense benchmarks: including 5 datasets that are seen during model training, as well as 4 datasets that are kept unseen. Our work is the first to report that knowledge generated by models that are orders of magnitude smaller than GPT-3, even without direct supervision on the knowledge itself, can exceed the quality of commonsense knowledge elicited from GPT-3.

**Empowering the Fact-checkers! Automatic Identification of Claim Spans on Twitter**
*Megha Sundriyal, Atharva Kulkarni, Vaibhav Pulastya, Md. Shad Akhtar and Tanmoy Chakraborty* 11:00-12:30 (Atrium)
The widespread diffusion of medical and political claims in the wake of COVID-19 has led to a voluminous rise in misinformation and fake news. The current vogue is to employ manual fact-checkers to efficiently classify and verify such data to combat this avalanche of claim-ridden misinformation. However, the rate of information dissemination is such that it vastly outpaces the fact-checkers' strength. Therefore, to aid manual fact-checkers in eliminating the superfluous content, it becomes imperative to automatically identify and extract the snippets of claim-worthy (mis)information present in a post. In this work, we introduce the novel task of Claim Span Identification (CSI). We propose CURT, a large-scale Twitter corpus with token-level claim spans on more than 7.5k tweets. Furthermore, along with the standard token classification baselines, we benchmark our dataset with DABERTa, an adapter-based variation of RoBERTa. The experimental results attest that DABERTa outperforms the baseline systems across several evaluation metrics, improving by about 1.5 points. We also report detailed error analysis to validate the model's performance along with the ablation studies. Lastly, we release our comprehensive span annotation guidelines for public use.

**Dealing with Abbreviations in the Slovenian Biographical Lexicon**
*Angel Daza, Antske Fokkens and Tomaž Erjavec* 11:00-12:30 (Atrium)
Abbreviations present a significant challenge for NLP systems because they cause tokenization and out-of-vocabulary errors. They can also make the text less readable, especially in reference printed books, where they are extensively used. Abbreviations are especially problematic in low-resource settings, where systems are less robust to begin with. In this paper, we propose a new method for addressing the problems caused by a high density of domain-specific abbreviations in a text. We apply this method to the case of a Slovenian biographical lexicon and evaluate it on a newly developed gold-standard dataset of 51 Slovenian biographies. Our abbreviation identification method performs significantly better than commonly used ad-hoc solutions, especially at identifying unseen abbreviations. We also propose and present the results of a method for expanding the identified abbreviations in context.

**Improving Multi-turn Emotional Support Dialogue Generation with Lookahead Strategy Planning**
*Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo WANG, Ruihui Zhao, Bang Liu, Xiaodan Liang and Yefeng Zheng* 11:00-12:30 (Atrium)
Providing Emotional Support (ES) to soothe people in emotional distress is an essential capability in social interactions. Most existing researches on building ES conversation systems only considered single-turn interactions with users, which was over-simplified. In comparison, multi-turn ES conversation systems can provide ES more effectively, but face several new technical challenges, including: (1) how to adopt appropriate support strategies to achieve the long-term dialogue goal of comforting the user's emotion; (2) how to dynamically model the user's state. In this paper, we propose a novel system MultiESC to address these issues. For strategy planning, drawing inspiration from the A* search algorithm, we propose lookahead heuristics to estimate the future user feedback after using particular strategies, which helps to select strategies that can lead to the best long-term effects. For user state modeling, MultiESC focuses on capturing users' subtle emotional expressions and understanding their emotion causes. Extensive experiments show that MultiESC significantly outperforms competitive baselines in both dialogue generation and strategy planning.

**Group is better than individual: Exploiting Label Topologies and Label Relations for Joint Multiple Intent Detection and Slot Filling**
*Bowen Xing and Ivor Tsang* 11:00-12:30 (Atrium)
Recent joint multiple intent detection and slot filling models employ label embeddings to achieve the semantics-label interactions. However, they treat all labels and label embeddings as uncorrelated individuals, ignoring the dependencies among them. Besides, they conduct the decoding for the two tasks independently, without leveraging the correlations between them. Therefore, in this paper, we first construct a Heterogeneous Label Graph (HLG) containing two kinds of topologies: (1) statistical dependencies based on labels' co-occurrence patterns and hierarchies in slot labels; (2) rich relations among the label nodes. Then we propose a novel model termed ReLa-Net. It can capture beneficial correlations among the labels from HLG. The label correlations are leveraged to enhance semantic-label interactions. Moreover, we also propose the label-aware inter-dependent decoding mechanism to further exploit the label correlations for decoding. Experiment results show that our ReLa-Net significantly outperforms previous models. Remarkably, ReLa-Net surpasses the previous best model by over 20

**Information-Theoretic Text Hallucination Reduction for Video-grounded Dialogue**
*Sunjae Yoon, Eunseop Yoon, Hee Suk Yoon, Junyeong Kim and Chang Yoo* 11:00-12:30 (Atrium)
Video-grounded Dialogue (VGD) aims to decode an answer sentence to a question regarding a given video and dialogue context. Despite the recent success of multi-modal reasoning to generate answer sentences, existing dialogue systems still suffer from a text hallucination problem, which denotes indiscriminate text-copying from input texts without an understanding of the question. This is due to learning spurious correlations from the fact that answer sentences in the dataset usually include the words of input texts, thus the VGD system excessively relies on copying words from input texts by hoping those words to overlap with ground-truth texts. Hence, we design Text Hallucination Mitigating (THAM) framework, which incorporates Text Hallucination Regularization (THR) loss derived from the proposed information-theoretic text hallucination measurement approach. Applying THAM with current dialogue systems validates the effectiveness on VGD benchmarks (i.e., AVSD@DSTC7 and AVSD@DSTC8) and shows enhanced interpretability.

**Dungeons and Dragons as a Dialog Challenge for Artificial Intelligence**
*Chris Callison-Burch, Gaurav Singh Tomar, Lara Martin, Daphne Ippolito, Suma Bailis and David Reitter* 11:00-12:30 (Atrium)

AI researchers have posited Dungeons and Dragons (D&D) as a challenge problem to test systems on various language-related capabilities. In this paper, we frame D&D specifically as a dialogue system challenge, where the tasks are to both generate the next conversational turn in the game and predict the state of the game given the dialogue history. We create a gameplay dataset consisting of nearly 900 games, with a total of 7,000 players, 800,000 dialogue turns, 500,000 dice rolls, and 58 million words. We automatically annotate the data with partial state information about the game play. We train a large language model (LM) to generate the next game turn, conditioning it on different information. The LM can respond as a particular character or as the player who runs the game—i.e., the Dungeon Master (DM). It is trained to produce dialogue that is either in-character (roleplaying in the fictional world) or out-of-character (discussing rules or strategy). We perform a human evaluation to determine what factors make the generated output plausible and interesting. We further perform an automatic evaluation to determine how well the model can predict the game state given the history and examine how well tracking the game state improves its ability to produce plausible conversational output.

**An Efficient Memory-Augmented Transformer for Knowledge-Intensive NLP Tasks**
*Yuxiang Wu, Yu Zhao, Baotian Hu, Pasquale Minervini, Pontus Stenetorp and Sebastian Riedel* 11:00-12:30 (Atrium)
Access to external knowledge is essential for many natural language processing tasks, such as question answering and dialogue. Existing methods often rely on a parametric model that stores knowledge in its parameters, or use a retrieval-augmented model that has access to an external knowledge source. Parametric and retrieval-augmented models have complementary strengths in terms of computational efficiency and predictive accuracy. To combine the strength of both approaches, we propose the Efficient Memory-Augmented Transformer (EMAT) – it encodes external knowledge into a key-value memory and exploits the fast maximum inner product search for memory querying. We also introduce pre-training tasks that allow EMAT to encode informative key-value representations, and to learn an implicit strategy to integrate multiple memory slots into the transformer. Experiments on various knowledge-intensive tasks such as question answering and dialogue datasets show that, simply augmenting parametric models (T5-base) using our method produces more accurate results (e.g., 25.8 → 44.3 EM on NQ) while retaining a high throughput (e.g., 1000 queries/s on NQ). Compared to retrieval-augmented models, EMAT runs substantially faster across the board and produces more accurate results on WoW and ELI5.

**Leveraging QA Datasets to Improve Generative Data Augmentation**
*Dheeraj Mekala, Tu Vu, Timo Schick and Jingbo Shang* 11:00-12:30 (Atrium)
The ability of generative language models (GLMs) to generate text has improved considerably in the last few years, enabling their use for generative data augmentation. In this work, we propose CONDA, an approach to further improve GLM's ability to generate synthetic data by reformulating data generation as context generation for a given question-answer (QA) pair and leveraging QA datasets for training context generators. Then, we cast downstream tasks into the same question answering format and adapt the fine-tuned context generators to the target task domain. Finally, we use the fine-tuned GLM to generate relevant contexts, which are in turn used as synthetic training data for their corresponding tasks. We perform extensive experiments on multiple classification datasets and demonstrate substantial improvements in performance for both few- and zero-shot settings. Our analysis reveals that QA datasets that require high-level reasoning abilities (e.g., abstractive and common-sense QA datasets) tend to give the best boost in performance in both few-shot and zero-shot settings.

**An Empirical Study on the Transferability of Transformer Modules in Parameter-efficient Fine-tuning**
*Mohammad AkbarTajari, Sara Rajaee and Mohammad Taher Pilehvar* 11:00-12:30 (Atrium)
Parameter-efficient fine-tuning has garnered lots of attention in recent studies. On this subject, we investigate the capability of different transformer modules in transferring knowledge from a pre-trained model to a downstream task. Our empirical results suggest that every transformer module is a winning ticket such that fine-tuning the specific module while the rest of the network is frozen achieves a comparable performance to the full fine-tuning case. Among different modules in LMs, LayerNorms exhibit a significant capacity for transfer learning to the extent that with only 0.003% updateable parameters in the layer-wise analysis, they can show acceptable performance on various target tasks. We argue that the performance of LayerNorms could be attributed to their high-magnitude weights compared to other components in a pre-trained model.

**Ethics consideration sections in natural language processing papers**
*Luciana Benotti and Patrick Blackburn* 11:00-12:30 (Atrium)
In this paper, we present the results of a manual classification of all ethical consideration sections for ACL 2021. We also compare how many papers had an ethics consideration section per track and per world region in ACL 2021. We classified papers according to the ethical issues covered (research benefits, potential harms, and vulnerable groups affected) and whether the paper was marked as requiring ethics review by at least one reviewer. Moreover, we discuss recurring obstacles we have observed (highlighting some interesting texts we found along the way) and conclude with three suggestions. We think that this paper may be useful for anyone who needs to write — or review — an ethics section and would like to get an overview of what others have done.

**An Empirical Study on Finding Spans**
*Weiwei Gu, Boyuan Zheng, Yunmo Chen, Tongfei Chen and Benjamin Van Durme* 11:00-12:30 (Atrium)
We present an empirical study on methods for span finding, the selection of consecutive tokens in text for some downstream tasks. We focus on approaches that can be employed in training end-to-end information extraction systems, and find there is no definitive solution without considering task properties, and provide our observations to help with future design choices: 1) a tagging approach often yields higher precision while span enumeration and boundary prediction provide higher recall; 2) span type information can benefit a boundary prediction approach; 3) additional contextualization does not help span finding in most cases.

**Simple Questions Generate Named Entity Recognition Datasets**
*Hyunjae Kim, jaehyo yoo, Seunghyun Yoon, Jinhyuk Lee and Jaewoo Kang* 11:00-12:30 (Atrium)
Recent named entity recognition (NER) models often rely on human-annotated datasets requiring the vast engagement of professional knowledge on the target domain and entities. This work introduces an ask-to-generate approach, which automatically generates NER datasets by asking simple natural language questions to an open-domain question answering system (e.g., "Which disease?"). Despite using fewer training resources, our models solely trained on the generated datasets largely outperform strong low-resource models by 19.5 F1 score across six popular NER benchmarks. Our models also show competitive performance with rich-resource models that additionally leverage in-domain dictionaries provided by domain experts. In few-shot NER, we outperform the previous best model by 5.2 F1 score on three benchmarks and achieve new state-of-the-art performance.

**Exploring Dual Encoder Architectures for Question Answering**
*Zhe Dong, Jianmo Ni, Dan Bikel, Enrique Alfonseca, Yuan Wang, Chen Qu and Imed Zitouni* 11:00-12:30 (Atrium)
Dual encoders have been used for question-answering (QA) and information retrieval (IR) tasks with good results. There are two major types of dual encoders, Siamese Dual Encoders (SDE), with parameters shared across two encoders, and Asymmetric Dual Encoder (ADE), with two distinctly parameterized encoders. In this work, we explore the dual encoder architectures for QA retrieval tasks. By evaluating on MS MARCO, open domain NQ, and the MultiReQA benchmarks, we show that SDE performs significantly better than ADE. We further propose three different improved versions of ADEs. Based on the evaluation of QA retrieval tasks and direct analysis of the embeddings, we

demonstrate that sharing parameters in projection layers would enable ADEs to perform competitively with SDEs.

**Let the CAT out of the bag: Contrastive Attributed explanations for Text**
*Saneem Chemmengath, Amar Prakash Azad, Ronny Luss and Amit Dhurandhar* 11:00-12:30 (Atrium)
Contrastive explanations for understanding the behavior of black box models has gained a lot of attention recently as they provide potential for recourse. In this paper, we propose a method Contrastive Attributed explanations for Text (CAT) which provides contrastive explanations for natural language text data with a novel twist as we build and exploit attribute classifiers leading to more semantically meaningful explanations. To ensure that our contrastive generated text has the fewest possible edits with respect to the original text, while also being fluent and close to a human generated contrastive, we resort to a minimal perturbation approach regularized using a BERT language model and attribute classifiers trained on available attributes. We show through qualitative examples and a user study that our method not only conveys more insight because of these attributes, but also leads to better quality (contrastive) text. Quantitatively, we show that our method outperforms other state-of-the-art methods across four data sets on four benchmark metrics.

**Attentional Probe: Estimating a Module's Functional Potential**
*Tiago Pimentel, Josef Valvoda, Niklas Stoehr and Ryan Cotterell* 11:00-12:30 (Atrium)

**Predicting Fine-Tuning Performance with Probing**
*Zining Zhu, Soroosh Shahtalebi and Frank Rudzicz* 11:00-12:30 (Atrium)
Large NLP models have recently shown impressive performance in language understanding tasks, typically evaluated by their fine-tuned performance. Alternatively, probing has received increasing attention as being a lightweight method for interpreting the intrinsic mechanisms of large NLP models. In probing, post-hoc classifiers are trained on "out-of-domain" datasets that diagnose specific abilities. While probing the language models has led to insightful findings, they appear disjointed from the development of models. This paper explores the utility of probing deep NLP models to extract a proxy signal widely used in model development – the fine-tuning performance. We find that it is possible to use the accuracies of only three probing tests to predict the fine-tuning performance with errors 40% - 80% smaller than baselines. We further discuss possible avenues where probing can empower the development of deep NLP models.

**BioReader: a Retrieval-Enhanced Text-to-Text Transformer for Biomedical Literature**
*Giacomo Frisoni, Miki Mizutani, Gianluca Moro and Lorenzo Valgimigli* 11:00-12:30 (Atrium)
The latest batch of research has equipped language models with the ability to attend over relevant and factual information from non-parametric external sources, drawing a complementary path to architectural scaling. Besides mastering language, exploiting and contextualizing the latent world knowledge is crucial in complex domains like biomedicine. However, most works in the field rely on general-purpose models supported by databases like Wikipedia and Books. We introduce BioReader, the first retrieval-enhanced text-to-text model for biomedical natural language processing. Our domain-specific T5-based solution augments the input prompt by fetching and assembling relevant scientific literature chunks from a neural database with $\approx$ million tokens centered on PubMed. We fine-tune and evaluate BioReader on a broad array of downstream tasks, significantly outperforming several state-of-the-art methods despite using up to 3x fewer parameters. In tandem with extensive ablation studies, we show that domain knowledge can be easily altered or supplemented to make the model generate correct predictions bypassing the retraining step and thus addressing the literature overload issue.

**Bernice: A Multilingual Pre-trained Encoder for Twitter**
*Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik and Mark Dredze* 11:00-12:30 (Atrium)
The language of Twitter differs significantly from that of other domains commonly included in large language model training. While tweets are typically multilingual and contain informal language, including emoji and hashtags, most pre-trained language models for Twitter are either monolingual, adapted from other domains rather than trained exclusively on Twitter, or are trained on a limited amount of in-domain Twitter data. We introduce Bernice, the first multilingual RoBERTa language model trained from scratch on 2.5 billion tweets with a custom tweet-focused tokenizer. We evaluate on a variety of monolingual and multilingual Twitter benchmarks, finding that our model consistently exceeds or matches the performance of a variety of models adapted to social media data as well as strong multilingual baselines, despite being trained on less data overall. We posit that it is more efficient compute- and data-wise to train completely on in-domain data with a specialized domain-specific tokenizer.

**ATTEMPT: Parameter-Efficient Multi-task Tuning via Attentional Mixtures of Soft Prompts**
*Akari Asai, Mohammadreza Salehi, Matthew Peters and Hannaneh Hajishirzi* 11:00-12:30 (Atrium)
This work introduces a new multi-task, parameter-efficient language model (LM) tuning method that learns to transfer knowledge across different tasks via a mixture of soft prompts—small prefix embedding vectors pre-trained for different tasks. Our method, called ATTEMPT (ATTEntional Mixtures of Prompt Tuning), obtains source prompts as encodings of large-scale source tasks into a small number of parameters and trains an attention module to interpolate the source prompts and a newly initialized target prompt for every instance in the target task. During training, only the target task prompt and the attention weights, which are shared between tasks in multi-task training, are updated, while the original LM and source prompts are intact. ATTEMPT is highly parameter-efficient (e.g., updates 2,300 times fewer parameters than full fine-tuning), while it overcomes instability of prompt tuning and achieves high task performance using learned knowledge from high-resource tasks. Moreover, it is modular using pre-trained soft prompts, and can flexibly add or remove source prompts for effective knowledge transfer. Our experimental results across 21 diverse NLP datasets show that ATTEMPT significantly outperforms prompt tuning and outperforms or matches fully fine-tuned or other parameter-efficient tuning approaches that use 10 times more parameters. Finally, ATTEMPT outperforms previous work in few-shot learning settings.

**SocioProbe: What, When, and Where Language Models Learn about Sociodemographics**
*Anne Lauscher, Federico Bianchi, Samuel R. Bowman and Dirk Hovy* 11:00-12:30 (Atrium)
Pre-trained language models (PLMs) have outperformed other NLP models on a wide range of tasks. Opting for a more thorough understanding of their capabilities and inner workings, researchers have established the extend to which they capture lower-level knowledge like grammaticality, and mid-level semantic knowledge like factual understanding. However, there is still little understanding of their knowledge of higher-level aspects of language. In particular, despite the importance of sociodemographic aspects in shaping our language, the questions of whether, where, and how PLMs encode these aspects, e.g., gender or age, is still unexplored. We address this research gap by probing the sociodemographic knowledge of different single-GPU PLMs on multiple English data sets via traditional classifier probing and information-theoretic minimum description length probing. Our results show that PLMs do encode these sociodemographics, and that this knowledge is sometimes spread across the layers of some of the tested PLMs. We further conduct a multilingual analysis and investigate the effect of supplementary training to further explore to what extent, where, and with what amount of pre-training data the knowledge is encoded. Our overall results indicate that sociodemographic knowledge is still a major challenge for NLP. PLMs require large amounts of pre-training data to acquire the knowledge and models that excel in general language understanding do not seem to own more knowledge about these aspects.

**ZeroGen: Efficient Zero-shot Learning via Dataset Generation**
*Jiacheng Ye, Jiahui Gao, Qintong Li, Hang XU, Jiangtao Feng, Zhiyong Wu, Tao Yu and Lingpeng Kong* 11:00-12:30 (Atrium)

There is a growing interest in dataset generation recently due to the superior generative capacity of large pre-trained language models (PLMs). In this paper, we study a flexible and efficient zero-short learning method, ZeroGen. Given a zero-shot task, we first generate a dataset from scratch using PLMs in an unsupervised manner. Then, we train a tiny task model (e.g., LSTM) under the supervision of the synthesized dataset. This approach allows highly efficient inference as the final task model only has orders of magnitude fewer parameters comparing to PLMs (e.g., GPT2-XL). Apart from being annotation-free and efficient, we argue that ZeroGen can also provide useful insights from the perspective of data-free model-agnostic knowledge distillation, and unreferenced text generation evaluation. Experiments and analysis on different NLP tasks, namely, text classification, question answering, and natural language inference, show the effectiveness of ZeroGen.

**Entropy- and Distance-Based Predictors From GPT-2 Attention Patterns Predict Reading Times Over and Above GPT-2 Surprisal**
*Byung-Doh Oh and William Schuler* 11:00-12:30 (Atrium)
Transformer-based large language models are trained to make predictions about the next word by aggregating representations of previous tokens through their self-attention mechanism. In the field of cognitive modeling, such attention patterns have recently been interpreted as embodying the process of cue-based retrieval, in which attention over multiple targets is taken to generate interference and latency during retrieval. Under this framework, this work first defines an entropy-based predictor that quantifies the diffuseness of self-attention, as well as distance-based predictors that capture the incremental change in attention patterns across timesteps. Moreover, following recent studies that question the informativeness of attention weights, we also experiment with alternative methods for incorporating vector norms into attention weights. Regression experiments using predictors calculated from the GPT-2 language model show that these predictors deliver a substantially better fit to held-out self-paced reading and eye-tracking data over a rigorous baseline including GPT-2 surprisal.

**Discourse Context Predictability Effects in Hindi Word Order**
*Sidharth Ranjan, Marten van Schijndel, Sumeet Agarwal and Rajakrishnan Rajkumar* 11:00-12:30 (Atrium)
We test the hypothesis that discourse predictability influences Hindi syntactic choice. While prior work has shown that a number of factors (e.g., information status, dependency length, and syntactic surprisal) influence Hindi word order preferences, the role of discourse predictability is underexplored in the literature. Inspired by prior work on syntactic priming, we investigate how the words and syntactic structures in a sentence influence the word order of the following sentences. Specifically, we extract sentences from the Hindi-Urdu Treebank corpus (HUTB), permute the preverbal constituents of those sentences, and build a classifier to predict which sentences actually occurred in the corpus against artificially generated distractors. The classifier uses a number of discourse-based features and cognitive features to make its predictions, including dependency length, surprisal, and information status. We find that information status and LSTM-based discourse predictability influence word order choices, especially for non-canonical object-fronted orders. We conclude by situating our results within the broader syntactic priming literature.

**Exploration of the Usage of Color Terms by Color-blind Participants in Online Discussion Platforms**
*Ella Rabinovich and Boaz Carmeli* 11:00-12:30 (Atrium)
Prominent questions about the role of sensory vs. linguistic input in the way we acquire and use language have been extensively studied in the psycholinguistic literature. However, the relative effect of various factors in a person's overall experience on their linguistic system remains unclear. We study this question by making a step forward towards a better understanding of the conceptual perception of colors by color-blind individuals, as reflected in their spontaneous linguistic productions. Using a novel and carefully curated dataset, we show that red-green color-blind speakers use the "red" and "green" color terms in less predictable contexts, and in linguistic environments evoking mental image to a lower extent, when compared to their normal-sighted counterparts. These findings shed some new and interesting light on the role of sensory experience on our linguistic system.

[TACL] **Assessing the capacity of transformer to abstract syntactic representations: a contrastive analysis based on long-distance agreement**
*Bingzhi Li, Guillaume Wisniewski and Benoit Crabbé* 11:00-12:30 (Atrium)
Many works have shown that transformers are able to predict subject-verb agreement, demonstrating their ability to uncover an abstract representation of the sentence in an unsupervised way. Thanks to this representation, transformers are able to capture syntactic dependencies between words and escape their linear order. Recently, li et al. (2021) found that transformers were also able to predict the object-past participle agreement in French. This kind of agreement involves a sequence of words similar to that of the subject-verb agreement but its modeling in formal grammar, which relies on a movement and an anaphora resolution, is fundamentally different from that of subject-verb agreement. To better understand the internal working of transformers, we propose, in this work, to contrast how they handle these two kinds of agreement: we aim at testing whether they encode the same abstract structure in their internal representation or, on the contrary, if this abstract structure is consistent with the distinction formal grammar is making in the modeling of these two agreements. Using probing and a new counterfactual analysis method, our experiments on French agreements show that i) the agreement task suffers from several confounders which partially question the conclusions drawn so far and ii) transformers handle subject-verb and object-past participle agreements in a way that is consistent with their modeling in theoretical linguistics.

[INDUSTRY] **Zero-Shot Dynamic Quantization for Transformer Inference**
*Yousef El-Kurdi, Jerry Quinn and Avi Sil* 11:00-12:30 (Atrium)
We introduce a novel run-time method for significantly reducing the accuracy loss associated with quantizing BERT-like models to 8-bit integers. Existing methods for quantizing models either modify the training procedure, or they require an additional calibration step to adjust parameters that also requires a selected held-out dataset. Our method permits taking advantage of quantization without the need for these adjustments. We present results on several NLP tasks demonstrating the usefulness of this technique.

[INDUSTRY] **Prototype-Representations for Training Data Filtering in Weakly-Supervised Information Extraction**
*Nasser Zalmout and Xian Li* 11:00-12:30 (Atrium)
The availability of high quality training data is still a bottleneck for the practical utilization of information extraction models, despite the breakthroughs in zero and few-shot learning techniques. This is further exacerbated for industry applications, where new tasks, domains, and specific use cases keep arising, which makes it impractical to depend on manually annotated data. Therefore, weak and distant supervision emerged as popular approaches to bootstrap training, utilizing labeling functions to guide the annotation process. Weakly-supervised annotation of training data is fast and efficient, however, it results in many irrelevant and out-of-context matches. This is a challenging problem that can degrade the performance in downstream models, or require a manual data cleaning step that can incur significant overhead. In this paper we present a prototype-based filtering approach, that can be utilized to denoise weakly supervised training data. The system is very simple, unsupervised, scalable, and requires little manual intervention, yet results in significant precision gains. We apply the technique in the task of attribute value extraction in e-commerce websites, and achieve up to 9% gain in precision for the downstream models, with a minimal drop in recall.

[INDUSTRY] **Entity-level Sentiment Analysis in Contact Center Telephone Conversations**
*Xue-Yong Fu, Cheng Chen, Md Tahmid Rahman Laskar, Shayna Gardiner, Pooja Hiranandani and Shashi Bhushan TN*11:00-12:30 (Atrium)
Entity-level sentiment analysis predicts the sentiment about entities mentioned in a given text. It is very useful in a business context to under-

stand user emotions towards certain entities, such as products or companies. In this paper, we demonstrate how we developed an entity-level sentiment analysis system that analyzes English telephone conversation transcripts in contact centers to provide business insight. We present two approaches, one entirely based on the transformer-based DistilBERT model, and another that uses a neural network supplemented with some heuristic rules.

[INDUSTRY] **QUILL: Query Intent with Large Language Models using Retrieval Augmentation and Multi-stage Distillation**
*Krishna Srinivasan, Karthik Raman, Anupam Samanta, Lingrui Liao, Luca Bertelli and Michael Bendersky* 11:00-12:30 (Atrium)
Large Language Models (LLMs) have shown impressive results on a variety of text understanding tasks. Search queries though pose a unique challenge, given their short-length and lack of nuance or context. Complicated feature engineering efforts do not always lead to downstream improvements as their performance benefits may be offset by increased complexity of knowledge distillation. Thus, in this paper we make the following contributions: (1) We demonstrate that Retrieval Augmentation of queries provides LLMs with valuable additional context enabling improved understanding. While Retrieval Augmentation typically increases latency of LMs (thus hurting distillation efficacy), (2) we provide a practical and effective way of distilling Retrieval Augmentation LLMs. Specifically, we use a novel two-stage distillation approach that allows us to carry over the gains of retrieval augmentation, without suffering the increased compute typically associated with it. (3) We demonstrate the benefits of the proposed approach (QUILL) on a billion-scale, real-world query understanding system resulting in huge gains. Via extensive experiments, including on public benchmarks, we believe this work offers a recipe for practical use of retrieval-augmented query understanding.

**PATS: Sensitivity-aware Noisy Learning for Pretrained Language Models**
*Yupeng Zhang, Hongzhi Zhang, Sirui Wang, Wei Wu and Zhoujun Li* 11:00-12:30 (Atrium)
A wide range of NLP tasks benefit from the fine-tuning of pretrained language models (PLMs). However, a number of redundant parameters which contribute less to the downstream task are observed in a directly fine-tuned model. We consider the gap between pretraining and downstream tasks hinders the training of these redundant parameters, and results in a suboptimal performance of the overall model. In this paper, we present PATS (Perturbation According To Sensitivity), a noisy training mechanism which considers each parameter's importance in the downstream task to help fine-tune PLMs. The main idea of PATS is to add bigger noise to parameters with lower sensitivity and vice versa, in order to activate more parameters' contributions to downstream tasks without affecting the sensitive ones much. Extensive experiments conducted on different tasks of the GLUE benchmark show PATS can consistently empower the fine-tuning of different sizes of PLMs, and the parameters in the well-performing models always have more concentrated distributions of sensitivities, which experimentally proves the effectiveness of our method.

**Complex Hyperbolic Knowledge Graph Embeddings with Fast Fourier Transform**
*Huiru Xiao, Xin Liu, Yangqiu Song, Ginny Wong and Simon See* 11:00-12:30 (Atrium)
The choice of geometric space for knowledge graph (KG) embeddings can have significant effects on the performance of KG completion tasks. The hyperbolic geometry has been shown to capture the hierarchical patterns due to its tree-like metrics, which addressed the limitations of the Euclidean embedding models. Recent explorations of the complex hyperbolic geometry further improved the hyperbolic embeddings for capturing a variety of hierarchical structures. However, the performance of the hyperbolic KG embedding models for non-transitive relations is still unpromising, while the complex hyperbolic embeddings do not deal with multi-relations. This paper aims to utilize the representation capacity of the complex hyperbolic geometry in multi-relational KG embeddings. To apply the geometric transformations which account for different relations and the attention mechanism in the complex hyperbolic space, we propose to use the fast Fourier transform (FFT) as the conversion between the real and complex hyperbolic space. Constructing the attention-based transformations in the complex space is very challenging, while the proposed Fourier transform-based complex hyperbolic approaches provide a simple and effective solution. Experimental results show that our methods outperform the baselines, including the Euclidean and the real hyperbolic embedding models.

**CTL++: Evaluating Generalization on Never-Seen Compositional Patterns of Known Functions, and Compatibility of Neural Representations**
*Róbert Csordás, Kazuki Irie and Juergen Schmidhuber* 11:00-12:30 (Atrium)
Well-designed diagnostic tasks have played a key role in studying the failure of neural nets (NNs) to generalize systematically. Famous examples include SCAN and Compositional Table Lookup (CTL). Here we introduce CTL++, a new diagnostic dataset based on compositions of unary symbolic functions. While the original CTL is used to test length generalization or productivity, CTL++ is designed to test systematicity of NNs, that is, their capability to generalize to unseen compositions of known functions. CTL++ splits functions into groups and tests performance on group elements composed in a way not seen during training. We show that recent CTL-solving Transformer variants fail on CTL++. The simplicity of the task design allows for fine-grained control of task difficulty, as well as many insightful analyses. For example, we measure how much overlap between groups is needed by tested NNs for learning to compose. We also visualize how learned symbol representations in outputs of functions from different groups are compatible in case of success but not in case of failure. These results provide insights into failure cases reported on more complex compositions in the natural language domain. Our code is public.

**Adaptive Label Smoothing with Self-Knowledge in Natural Language Generation**
*Dongkyu Lee, Ka Chun Cheung and Nevin Zhang* 11:00-12:30 (Atrium)
Overconfidence has been shown to impair generalization and calibration of a neural network. Previous studies remedy this issue by adding a regularization term to a loss function, preventing a model from making a peaked distribution. Label smoothing smoothes target labels with a pre-defined prior label distribution; as a result, a model is learned to maximize the likelihood of predicting the soft label. Nonetheless, the amount of smoothing is the same in all samples and remains fixed in training. In other words, label smoothing does not reflect the change in probability distribution mapped by a model over the course of training. To address this issue, we propose a regularization scheme that brings dynamic nature into the smoothing parameter by taking model probability distribution into account, thereby varying the parameter per instance. A model in training self-regulates the extent of smoothing on the fly during forward propagation. Furthermore, inspired by recent work in bridging label smoothing and knowledge distillation, our work utilizes self-knowledge as a prior label distribution in softening target labels, and presents theoretical support for the regularization effect by knowledge distillation and the dynamic smoothing parameter. Our regularizer is validated comprehensively, and the result illustrates marked improvements in model generalization and calibration, enhancing robustness and trustworthiness of a model.

**MM-Align: Learning Optimal Transport-based Alignment Dynamics for Fast and Accurate Inference on Missing Modality Sequences**
*Wei Han, Hui Chen, Min-Yen Kan and Soujanya Poria* 11:00-12:30 (Atrium)
Existing multimodal tasks mostly target at the complete input modality setting, i.e., each modality is either complete or completely missing in both training and test sets. However, the randomly missing situations have still been underexplored. In this paper, we present a novel approach named MM-Align to address the missing-modality inference problem. Concretely, we propose 1) an alignment dynamics learning module based on the theory of optimal transport (OT) for missing data imputation; 2) a denoising training algorithm to enhance the quality of imputation as well as the accuracy of model predictions. Compared with previous generative methods which devote to restoring the missing inputs, MM-Align learns to capture and imitate the alignment dynamics between modality sequences. Results of comprehensive experiments on two multimodal tasks empirically demonstrate that our method can perform more accurate and faster inference and alleviate the overfitting

issue under different missing conditions.

**Diverse Parallel Data Synthesis for Cross-Database Adaptation of Text-to-SQL Parsers**
*Abhijeet Awasthi, Ashutosh Sathe and Sunita Sarawagi* 11:00-12:30 (Atrium)
Text-to-SQL parsers typically struggle with databases unseen during the train time. Adapting Text-to-SQL parsers to new database schemas is a challenging problem owing to a vast diversity of schemas and zero availability of natural language queries in new schemas. We present ReFill, a framework for synthesizing high-quality and textually diverse parallel datasets for adapting Text-to-SQL parsers. Unlike prior methods that utilize SQL-to-Text generation, ReFill learns to retrieve-and-edit text queries in existing schemas and transfer them to the new schema. ReFill utilizes a simple method for retrieving diverse existing text, masking their schema-specific tokens, and refilling with tokens relevant to the new schema. We show that this process leads to significantly more diverse text queries than achievable by standard SQL-to-Text generation models. Through experiments on several databases, we show that adapting a parser by finetuning it on datasets synthesized by ReFill consistently outperforms prior data-augmentation methods.

**Normalizing Mutual Information for Robust Adaptive Training for Translation**
*Youngwon Lee, Changmin Lee, Hojin Lee and Seung-won Hwang* 11:00-12:30 (Atrium)
Despite the success of neural machine translation models, tensions between fluency of optimizing target language modeling and source-faithfulness remain as challenges. Previously, Conditional Bilingual Mutual Information (CBMI), a scoring metric for the importance of target sentences and tokens, was proposed to encourage fluent and faithful translations. The score is obtained by combining the probability from the translation model and the target language model, which is then used to assign different weights to losses from sentences and tokens. Meanwhile, we argue this metric is not properly normalized, for which we propose Normalized Pointwise Mutual Information (NPMI). NPMI utilizes an additional language model on source language to approximate the joint likelihood of source-target pair and the likelihood of the source, which is then used for normalizing the score. We showed that NPMI better captures the dependence between source-target and that NPMI-based token-level adaptive training brings improvements over baselines with empirical results from En-De, De-En, and En-Ro translation tasks.

**Bilingual Synchronization: Restoring Translational Relationships with Editing Operations**
*Jitao Xu, Josep Crego and François Yvon* 11:00-12:30 (Atrium)
Machine Translation (MT) is usually viewed as a one-shot process that generates the target language equivalent of some source text from scratch. We consider here a more general setting which assumes an initial target sequence, that must be transformed into a valid translation of the source, thereby restoring parallelism between source and target. For this bilingual synchronization task, we consider several architectures (both autoregressive and non-autoregressive) and training regimes, and experiment with multiple practical settings such as simulated interactive MT, translating with Translation Memory (TM) and TM cleaning. Our results suggest that one single generic edit-based system, once fine-tuned, can compare with, or even outperform, dedicated systems specifically trained for these tasks.

**Does Joint Training Really Help Cascaded Speech Translation?**
*Viet Anh Khoa Tran, David Thulke, Yingbo Gao, Christian Herold and Hermann Ney* 11:00-12:30 (Atrium)
Currently, in speech translation, the straightforward approach - cascading a recognition system with a translation system - delivers state-of-the-art results. However, fundamental challenges such as error propagation from the automatic speech recognition system still remain. To mitigate these problems, recently, people turn their attention to direct data and propose various joint training methods. In this work, we seek to answer the question of whether joint training really helps cascaded speech translation. We review recent papers on the topic and also investigate a joint training criterion by marginalizing the transcription posterior probabilities. Our findings show that a strong cascaded baseline can diminish any improvements obtained using joint training, and we suggest alternatives to joint training. We hope this work can serve as a refresher of the current speech translation landscape, and motivate research in finding more efficient and creative ways to utilize the direct data for speech translation.

**Discovering Language-neutral Sub-networks in Multilingual Language Models**
*Negar Foroutan, Mohammadreza Banaei, Rémi Lebret, Antoine Bosselut and Karl Aberer* 11:00-12:30 (Atrium)
Multilingual pre-trained language models transfer remarkably well on cross-lingual downstream tasks. However, the extent to which they learn language-neutral representations (i.e., shared representations that encode similar phenomena across languages), and the effect of such representations on cross-lingual transfer performance, remain open questions.

In this work, we conceptualize language neutrality of multilingual models as a function of the overlap between language-encoding sub-networks of these models. We employ the lottery ticket hypothesis to discover sub-networks that are individually optimized for various languages and tasks. Our evaluation across three distinct tasks and eleven typologically-diverse languages demonstrates that sub-networks for different languages are topologically similar (i.e., language-neutral), making them effective initializations for cross-lingual transfer with limited performance degradation.

**Don't Stop Fine-Tuning: On Training Regimes for Few-Shot Cross-Lingual Transfer with Multilingual Language Models**
*Fabian David Schmidt, Ivan Vulić and Goran Glavaš* 11:00-12:30 (Atrium)
A large body of recent work highlights the fallacies of zero-shot cross-lingual transfer (ZS-XLT) with large multilingual language models. Namely, their performance varies substantially for different target languages and is the weakest where needed the most: for low-resource languages distant to the source language. One remedy is few-shot transfer (FS-XLT), where leveraging only a few task-annotated instances in the target language(s) may yield sizable performance gains. However, FS-XLT also succumbs to large variation, as models easily overfit to the small datasets. In this work, we present a systematic study focused on a spectrum of FS-XLT fine-tuning regimes, analyzing key properties such as effectiveness, (in)stability, and modularity. We conduct extensive experiments on both higher-level (NLI, paraphrasing) and lower-level tasks (NER, POS), presenting new FS-XLT strategies that yield both improved and more stable FS-XLT across the board. Our findings challenge established FS-XLT methods: e.g., we propose to replace sequential fine-tuning with joint fine-tuning on source and target language instances, offering consistent gains with different number of shots (including resource-rich scenarios). We also show that further gains can be achieved with multi-stage FS-XLT training in which joint multilingual fine-tuning precedes the bilingual source-target specialization.

**Improving Low-Resource Languages in Pre-Trained Multilingual Language Models**
*Viktor Hangya, Hossain Shaikh Saadi and Alexander Fraser* 11:00-12:30 (Atrium)
Pre-trained multilingual language models are the foundation of many NLP approaches, including cross-lingual transfer solutions. However, languages with small available monolingual corpora are often not well-supported by these models leading to poor performance. We propose an unsupervised approach to improve the cross-lingual representations of low-resource languages by bootstrapping word translation pairs from monolingual corpora and using them to improve language alignment in pre-trained language models. We perform experiments on nine languages, using contextual word retrieval and zero-shot named entity recognition to measure both intrinsic cross-lingual word representation quality and downstream task performance, showing improvements on both tasks. Our results show that it is possible to improve pre-trained multilingual language models by relying only on non-parallel resources.

**RED-ACE: Robust Error Detection for ASR using Confidence Embeddings**
*Zorik Gekhman, Dina Zverinski, Jonathan Mallinson and Genady Beryozkin* 11:00-12:30 (Atrium)
ASR Error Detection (AED) models aim to post-process the output of Automatic Speech Recognition (ASR) systems, in order to detect transcription errors. Modern approaches usually use text-based input, comprised solely of the ASR transcription hypothesis, disregarding additional signals from the ASR model. Instead, we utilize the ASR system's word-level confidence scores for improving AED performance. Specifically, we add an ASR Confidence Embedding (ACE) layer to the AED model's encoder, allowing us to jointly encode the confidence scores and the transcribed text into a contextualized representation. Our experiments show the benefits of ASR confidence scores for AED, their complementary effect over the textual signal, as well as the effectiveness and robustness of ACE for combining these signals. To foster further research, we publish a novel AED dataset consisting of ASR outputs on the LibriSpeech corpus with annotated transcription errors.

**Towards Compositional Generalization in Code Search**
*Hojae Han, Seung-won Hwang, Shuai Lu, Nan Duan and Seungtaek Choi* 11:00-12:30 (Atrium)
We study compositional generalization, which aims to generalize on unseen combinations of seen structural elements, for code search. Unlike existing approaches of partially pursuing this goal, we study how to extract structural elements, which we name a template that directly targets compositional generalization. Thus we propose CTBERT, or Code Template BERT, representing codes using automatically extracted templates as building blocks. We empirically validate CTBERT on two public code search benchmarks, AdvTest and CSN. Further, we show that templates are complementary to data flow graphs in GraphCodeBERT, by enhancing structural context around variables.

**Conditional set generation using Seq2seq models**
*Aman Madaan, Dheeraj Rajagopal, Niket Tandon, Yiming Yang and Antoine Bosselut* 11:00-12:30 (Atrium)
Conditional set generation learns a mapping from an input sequence of tokens to a set. Several NLP tasks, such as entity typing and dialogue emotion tagging, are instances of set generation. Seq2Seq models are a popular choice to model set generation but they treat a set as a sequence and do not fully leverage its key properties, namely order-invariance and cardinality. We propose a novel algorithm for effectively sampling informative orders over the combinatorial space of label orders. Further, we jointly model the set cardinality and output by listing the set size as the first element and taking advantage of the autoregressive factorization used by Seq2Seq models. Our method is a model-independent data augmentation approach that endows any Seq2Seq model with the signals of order-invariance and cardinality. Training a Seq2Seq model on this new augmented data (without any additional annotations), gets an average relative improvement of 20% for four benchmarks datasets across models spanning from BART-base, T5-11B, and GPT-3. We will release all code and data upon acceptance.

**Controlled Text Reduction**
*Aviv Slobodkin, Paul Roit, Eran Hirsch, Ori Ernst and Ido Dagan* 11:00-12:30 (Atrium)
Producing a reduced version of a source text, as in generic or focused summarization, inherently involves two distinct subtasks: deciding on targeted content and generating a coherent text conveying it. While some popular approaches address summarization as a single end-to-end task, prominent works support decomposed modeling for individual subtasks. Further, semi-automated text reduction is also very appealing, where users may identify targeted content while models would generate a corresponding coherent summary. In this paper, we focus on the second subtask, of generating coherent text given pre-selected content. Concretely, we formalize *Controlled Text Reduction* as a standalone task, whose input is a source text with marked spans of targeted content ("highlighting"). A model then needs to generate a coherent text that includes all and only the target information. We advocate the potential of such models, both for modular fully-automatic summarization, as well as for semi-automated human-in-the-loop use cases. Facilitating proper research, we crowdsource high-quality dev and test datasets for the task. Further, we automatically generate a larger "silver" training dataset from available summarization benchmarks, leveraging a pretrained summary-source alignment model. Finally, employing these datasets, we present a supervised baseline model, showing promising results and insightful analyses.

**Break it Down into BTS: Basic, Tiniest Subword Units for Korean**
*Nayeon Kim, Jun-Hyung Park, Joon-Young Choi, Eojin Jeon, Youjin Kang and SangKeun Lee* 11:00-12:30 (Atrium)
We introduce Basic, Tiniest Subword (BTS) units for the Korean language, which are inspired by the invention principle of Hangeul, the Korean writing system. Instead of relying on 51 Korean consonant and vowel letters, we form the letters from BTS units by adding strokes or combining them. To examine the impact of BTS units on Korean language processing, we develop a novel BTS-based word embedding framework that is readily applicable to various models. Our experiments reveal that BTS units significantly improve the performance of Korean word embedding on all intrinsic and extrinsic tasks in our evaluation. In particular, BTS-based word embedding outperforms the state-of-theart Korean word embedding by 11.8% in word analogy. We further investigate the unique advantages provided by BTS units through indepth analysis.

**Improving Passage Retrieval with Zero-Shot Question Generation**
*Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau and Luke Zettlemoyer* 11:00-12:30 (Atrium)
We propose a simple and effective re-ranking method for improving passage retrieval in open question answering. The re-ranker re-scores retrieved passages with a zero-shot question generation model, which uses a pre-trained language model to compute the probability of the input question conditioned on a retrieved passage. This approach can be applied on top of any retrieval method (e.g. neural or keyword-based), does not require any domain- or task-specific training (and therefore is expected to generalize better to data distribution shifts), and provides rich cross-attention between query and passage (i.e. it must explain every token in the question). When evaluated on a number of open-domain retrieval datasets, our re-ranker improves strong unsupervised retrieval models by 6%-18% absolute and strong supervised models by up to 12% in terms of top-20 passage retrieval accuracy. We also obtain new state-of-the-art results on full open-domain question answering by simply adding the new re-ranker to existing models with no further changes.

**Analogical Math Word Problems Solving with Enhanced Problem-Solution Association**
*Zhenwen Liang, Jipeng Zhang and Xiangliang Zhang* 11:00-12:30 (Atrium)
Math word problem (MWP) solving is an important task in question answering which requires human-like reasoning ability. Analogical reasoning has long been used in mathematical education, as it enables students to apply common relational structures of mathematical situations to solve new problems. In this paper, we propose to build a novel MWP solver by leveraging analogical MWPs, which advance the solver's generalization ability across different kinds of MWPs. The key idea, named analogy identification, is to associate the analogical MWP pairs in a latent space, i.e., encoding an MWP close to another analogical MWP, while leaving away from the non-analogical ones. Moreover, a solution discriminator is integrated into the MWP solver to enhance the association between an MWP and its true solution. The evaluation results verify that our proposed analogical learning strategy promotes the performance of MWP-BERT on Math23k over the state-of-the-art model Generate2Rank, with 5 times fewer parameters in the encoder. We also find that our model has a stronger generalization ability in solving difficult MWPs due to the analogical learning from easy MWPs.

**Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks**
*Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan*

*Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, rushang karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit and Xudong Shen* 11:00-12:30 (Atrium)
How well can NLP models generalize to a variety of unseen tasks when provided with task instructions? To address this question, we first introduce Super-NaturalInstructions, a benchmark of 1,616 diverse NLP tasks and their expert-written instructions. Our collection covers 76 distinct task types, including but not limited to classification, extraction, infilling, sequence tagging, text rewriting, and text composition. This large and diverse collection of tasks enables rigorous benchmarking of cross-task generalization under instructions—training models to follow instructions on a subset of tasks and evaluating them on the remaining unseen ones. Furthermore, we build Tk-Instruct, a transformer model trained to follow a variety of in-context instructions (plain language task definitions or k-shot examples). Our experiments show that Tk-Instruct outperforms existing instruction-following models such as InstructGPT by over 9% on our benchmark despite being an order of magnitude smaller. We further analyze generalization as a function of various scaling parameters, such as the number of observed tasks, the number of instances per task, and model sizes. We hope our dataset and model facilitate future progress towards more general-purpose NLP models.

**DivEMT: Neural Machine Translation Post-Editing Effort Across Typologically Diverse Languages**
*Gabriele Sarti, Arianna Bisazza, Ana Guerberof-Arenas and Antonio Toral* 11:00-12:30 (Atrium)
We introduce DivEMT, the first publicly available post-editing study of Neural Machine Translation (NMT) over a typologically diverse set of target languages. Using a strictly controlled setup, 18 professional translators were instructed to translate or post-edit the same set of English documents into Arabic, Dutch, Italian, Turkish, Ukrainian, and Vietnamese. During the process, their edits, keystrokes, editing times and pauses were recorded, enabling an in-depth, cross-lingual evaluation of NMT quality and post-editing effectiveness. Using this new dataset, we assess the impact of two state-of-the-art NMT systems, Google Translate and the multilingual mBART-50 model, on translation productivity. We find that post-editing is consistently faster than translation from scratch. However, the magnitude of productivity gains varies widely across systems and languages, highlighting major disparities in post-editing effectiveness for languages at different degrees of typological relatedness to English, even when controlling for system architecture and training data size. We publicly release the complete dataset including all collected behavioral data, to foster new research on the translation capabilities of NMT systems for typologically diverse languages.

**Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering**
*Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu and Marco Guerini* 11:00-12:30 (Atrium)
Fighting online hate speech is a challenge that is usually addressed using Natural Language Processing via automatic detection and removal of hate content. Besides this approach, counter narratives have emerged as an effective tool employed by NGOs to respond to online hate on social media platforms. For this reason, Natural Language Generation is currently being studied as a way to automatize counter narrative writing. However, the existing resources necessary to train NLG models are limited to 2-turn interactions (a hate speech and a counter narrative as response), while in real life, interactions can consist of multiple turns. In this paper, we present a hybrid approach for dialogical data collection, which combines the intervention of human expert annotators over machine generated dialogues obtained using 19 different configurations. The result of this work is DIALOCONAN, the first dataset comprising over 3000 fictitious multi-turn dialogues between a hater and an NGO operator, covering 6 targets of hate.

**Bloom Library: Multimodal Datasets in 300+ Languages for a Variety of Downstream Tasks**
*Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni and Daniel Whitenack* 11:00-12:30 (Atrium)
We present Bloom Library, a linguistically diverse set of multimodal and multilingual datasets for language modeling, image captioning, visual storytelling, and speech synthesis/recognition. These datasets represent either the most, or among the most, multilingual datasets for each of the included downstream tasks. In total, the initial release of the Bloom Library datasets covers 363 languages across 32 language families. We train downstream task models for various languages represented in the data, showing the viability of the data for future work in low-resource, multimodal NLP and establishing the first known baselines for these downstream tasks in certain languages (e.g., Bisu [bzi], with an estimated population of 700 users). Some of these first-of-their-kind baselines are comparable to state-of-the-art performance for higher-resourced languages. The Bloom Library datasets are released under Creative Commons licenses on the Hugging Face datasets hub to catalyze more linguistically diverse research in the included downstream tasks.

**DiscoSense: Commonsense Reasoning with Discourse Connectives**
*Prajjwal Bhargava and Vincent Ng* 11:00-12:30 (Atrium)
We present DiscoSense, a benchmark for commonsense reasoning via understanding a wide variety of discourse connectives. We generate compelling distractors in DiscoSense using Conditional Adversarial Filtering, an extension of Adversarial Filtering that employs conditional generation. We show that state-of-the-art pre-trained language models struggle to perform well on DiscoSense, which makes this dataset ideal for evaluating next-generation commonsense reasoning systems.

**GraphQ IR: Unifying the Semantic Parsing of Graph Query Languages with One Intermediate Representation**
*Lunyiu Nie, Shulin Cao, Jiaxin Shi, Jiuding Sun, Qi Tian, Lei Hou, Juanzi Li and Jidong Zhai* 11:00-12:30 (Atrium)
Subject to the huge semantic gap between natural and formal languages, neural semantic parsing is typically bottlenecked by its complexity of dealing with both input semantics and output syntax. Recent works have proposed several forms of supplementary supervision but none is generalized across multiple formal languages. This paper proposes a unified intermediate representation for graph query languages, named GraphQ IR. It has a natural-language-like expression that bridges the semantic gap and formally defined syntax that maintains the graph structure. Therefore, a neural semantic parser can more precisely convert user queries into GraphQ IR, which can be later losslessly compiled into various downstream graph query languages. Extensive experiments on several benchmarks including KQA Pro, Overnight, GrailQA, and MetaQA-Cypher under the standard i.i.d., out-of-distribution, and low-resource settings validate GraphQ IR's superiority over the previous state-of-the-arts with a maximum 11% accuracy improvement.

**QASem Parsing: Text-to-text Modeling of QA-based Semantics**
*Ayal Klein, Eran Hirsch, Ron Eliav, Valentina Pyatkin, Avi Caciularu and Ido Dagan* 11:00-12:30 (Atrium)
Various works suggest the appeals of incorporating explicit semantic representations when addressing challenging realistic NLP scenarios. Common approaches offer either comprehensive linguistically-based formalisms, like AMR, or alternatively Open-IE, which provides a shallow and partial representation. More recently, an appealing trend introduces semi-structured natural-language structures as an intermediate meaning-capturing representation, often in the form of questions and answers.

In this work, we further promote this line of research by considering three prior QA-based semantic representations. These cover verbal, nominalized and discourse-based predications, regarded as jointly providing a comprehensive representation of textual information — termed QASem. To facilitate this perspective, we investigate how to best utilize pre-trained sequence-to-sequence language models, which seem particularly promising for generating representations that consist of natural language expressions (questions and answers). In particular, we examine and analyze input and output linearization strategies, as well as data augmentation and multitask learning for a scarce training data

setup. Consequently, we release the first unified QASem parsing tool, easily applicable for downstream tasks that can benefit from an explicit semi-structured account of information units in text.

**Correcting Diverse Factual Errors in Abstractive Summarization via Post-Editing and Language Model Infilling**
*Vidhisha Balachandran, Hannaneh Hajishirzi, William Cohen and Yulia Tsvetkov* 11:00-12:30 (Atrium)
Abstractive summarization models often generate inconsistent summaries containing factual errors or hallucinated content. Recent works focus on correcting factual errors in generated summaries via post-editing. Such correction models are trained using adversarial non-factual summaries constructed using heuristic rules for injecting errors. However, generating non-factual summaries using heuristics often does not generalize well to actual model errors. In this work, we propose to generate hard, representative synthetic examples of non-factual summaries through infilling language models. With this data, we train a more robust fact-correction model to post-edit the summaries to improve factual consistency. Through quantitative and qualitative experiments on two popular summarization datasets— CNN/DM and XSum—we show that our approach vastly outperforms prior methods in correcting erroneous summaries. Our model—FactEdit—improves factuality scores by over 11 points on CNN/DM and over 31 points on XSum on average across multiple summarization models, producing more factual summaries while maintaining competitive summarization quality.

**On Parsing as Tagging**
*Afra Amini and Ryan Cotterell* 11:00-12:30 (Atrium)
There are many proposals to reduce constituency parsing to tagging. To figure out what these approaches have in common, we offer a unifying pipeline, which consists of three steps: linearization, learning, and decoding. We prove that classic shift–reduce parsing can be reduced to tetratagging—the state-of-the-art constituency tagger—under two assumptions: right-corner transformation in the linearization step and factored scoring in the learning step. We ask what is the most critical factor that makes parsing-as-tagging methods accurate while being efficient. To answer this question, we empirically evaluate a taxonomy of tagging pipelines with different choices of linearizers, learners, and decoders. Based on the results in English as well as a set of 8 typologically diverse languages, we conclude that the linearization of the derivation tree and its alignment with the input sequence is the most critical factor in achieving accurate parsers as taggers.

**Structural generalization is hard for sequence-to-sequence models**
*Yuekun Yao and Alexander Koller* 11:00-12:30 (Atrium)
Sequence-to-sequence (seq2seq) models have been successful across many NLP tasks, including ones that require predicting linguistic structure. However, recent work on compositional generalization has shown that seq2seq models achieve very low accuracy in generalizing to linguistic structures that were not seen in training. We present new evidence that this is a general limitation of seq2seq models that is present not just in semantic parsing, but also in syntactic parsing and in text-to-text tasks, and that this limitation can often be overcome by neurosymbolic models that have linguistic knowledge built in. We further report on some experiments that give initial answers on the reasons for these limitations.

**CONDAQA: A Contrastive Reading Comprehension Dataset for Reasoning about Negation**
*Abhilasha Ravichander, Matt Gardner and Ana Marasovic* 11:00-12:30 (Atrium)
The full power of human language-based communication cannot be realized without negation. All human languages have some form of negation. Despite this, negation remains a challenging phenomenon for current natural language understanding systems. To facilitate the future development of models that can process negation effectively, we present CONDAQA, the first English reading comprehension dataset which requires reasoning about the implications of negated statements in paragraphs. We collect paragraphs with diverse negation cues, then have crowdworkers ask questions about the implications of the negated statement in the passage. We also have workers make three kinds of edits to the passage—paraphrasing the negated statement, changing the scope of the negation, and reversing the negation—resulting in clusters of question-answer pairs that are difficult for models to answer with spurious shortcuts. CONDAQA features 14,182 question-answer pairs with over 200 unique negation cues and is challenging for current state-of-the-art models. The best performing model on CONDAQA (UnifiedQA-v2-3b) achieves only 42% on our consistency metric, well below human performance which is 81%. We release our dataset, along with fully-finetuned, few-shot, and zero-shot evaluations, to facilitate the development of future NLP methods that work on negated language.

**Rethinking Style Transformer with Energy-based Interpretation: Adversarial Unsupervised Style Transfer using a Pretrained Model**
*Hojun Cho, Dohee Kim, Seungwoo Ryu, ChaeHun Park, Hyungjong Noh, Jeong-in Hwang, Minseok Choi, Edward Choi and Jaegul Choo*
11:00-12:30 (Atrium)
Style control, content preservation, and fluency determine the quality of text style transfer models. To train on a nonparallel corpus, several existing approaches aim to deceive the style discriminator with an adversarial loss. However, adversarial training significantly degrades fluency compared to the other two metrics. In this work, we explain this phenomenon using energy-based interpretation, and leverage a pretrained language model to improve fluency. Specifically, we propose a novel approach which applies the pretrained language model to the text style transfer framework by restructuring the discriminator and the model itself, allowing the generator and the discriminator to also take advantage of the power of the pretrained model. We evaluated our model on three public benchmarks GYAFC, Amazon, and Yelp and achieved state-of-the-art performance on the overall metrics.

## Demo Session 4

11:00-12:30 (Atrium)

[DEMO] **DeepGen: Diverse Search Ad Generation and Real-Time Customization**
*Konstantin Golobokov, Junyi Chai, Victor Ye Dong, Mandy Gu, Bingyu Chi, Jie Cao, Yulan Yan and Yi Liu* 11:00-12:30 (Atrium)
We present DeepGen, a system deployed at web scale for automatically creating sponsored search advertisements (ads) for BingAds customers. We leverage state-of-the-art natural language generation (NLG) models to generate fluent ads from advertiser's web pages in an abstractive fashion and solve practical issues such as factuality and inference speed. In addition, our system creates a customized ad in real-time in response to the user's search query, therefore highlighting different aspects of the same product based on what the user is looking for. To achieve this, our system generates a diverse choice of smaller pieces of the ad ahead of time and, at query time, selects the most relevant ones to be stitched into a complete ad. We improve generation diversity by training a controllable NLG model to generate multiple ads for the same web page highlighting different selling points. Our system design further improves diversity horizontally by first running an ensemble of generation models trained with different objectives and then using a diversity sampling algorithm to pick a diverse subset of generation results for online selection. Experimental results show the effectiveness of our proposed system design. Our system is currently deployed in production, serving about 4

[DEMO] **ACCoRD: A Multi-Document Approach to Generating Diverse Descriptions of Scientific Concepts**
*Sonia Krishna Murthy, Kyle Lo, Daniel King, Chandra Bhagavatula, Bailey Kuehl, Sophie Johnson, Jonathan Borchardt, Daniel Weld, Tom Hope and Doug Downey* 11:00-12:30 (Atrium)
Systems that automatically define unfamiliar terms hold the promise of improving the accessibility of scientific texts, especially for readers who may lack prerequisite background knowledge. However, current systems assume a single "best" description per concept, which fails to account for the many ways a concept can be described. We present ACCoRD, an end-to-end system tackling the novel task of generating sets of descriptions of scientific concepts. Our system takes advantage of the myriad ways a concept is mentioned across the scientific literature to produce distinct, diverse descriptions of target concepts in terms of different reference concepts. In a user study, we find that users prefer (1) descriptions produced by our end-to-end system, and (2) multiple descriptions to a single "best description." We release the ACCoRD corpus which includes 1,275 labeled contexts and 1,787 expert-authored concept descriptions to support research on our task.

[DEMO] **SUMMARY WORKBENCH: Unifying Application and Evaluation of Text Summarization Models**
*Shahbaz Syed, Dominik Schwabe and Martin Potthast* 11:00-12:30 (Atrium)
This paper presents Summary Workbench, a new tool for developing and evaluating text summarization models. New models and evaluation measures can be easily integrated as Docker-based plugins, allowing to examine the quality of their summaries against any input and to evaluate them using various evaluation measures. Visual analyses combining multiple measures provide insights into the models' strengths and weaknesses. The tool is hosted at https://tldr.demo.webis.de and also supports local deployment for private resources.

[DEMO] **GEMv2: Multilingual NLG Benchmarking in a Single Line of Code**
*Sebastian Gehrmann, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papangelis, Aman Madaan, Angelina McMillan-Major, Anna Shvets, Ashish Upadhyay, Bernd Bohnet, Bingsheng Yao, Bryan Wilie, Chandra Bhagavatula, Craig Thomson, Cristina Garbacea, Dakuo Wang, Daniel Deutsch, Deyi Xiong, Di Jin, Dimitra Gkatzia, Dragomir Radev, Elizabeth Clark, Esin Durmus, Faisal Ladhak, Filip Ginter, Genta Indra Winata, Hendrik Strobelt, Hiroaki Hayashi, Jekaterina Novikova, Jenna Kanerva, Jenny Chim, Jiawei Zhou, João Sedoc, Jordan Clive, Joshua Maynez, Juraj Juraska, Kaustubh Dhole, Khyathi Raghavi Chandu, Laura Perez-Beltrachini, Leonardo F. R. Ribeiro, Lewis Tunstall, Li Zhang, Mahima Pushkarna, Mathias Creutz, Michael White, Mihir Kale, Moussa Kamal Eddine, Nico Daheim, Nishant Subramani, Ondřej Dušek, Paul Pu Liang, Pawan Sasanka Ammanamanchi, Qi Zhu, Ratish Puduppully, Reno Kriz, Rifat Shahriyar, Ronald Ahmed Cardenas, Saad Mahamood, Salomey Osei, Samuel Cahyawijaya, Sanja Stajner, Sebastien Montella, Shailza Jolly, Simon Mille, Tahmid Hasan, Tianhao Shen, Tosin Adewumi, Vikas Raunak, Vipul Raheja, Vitaly Nikolaev, Vivian Tsai, Yacine Jernite, Ying Xu, Yisi Sang, Yixin Liu and Yufang Hou* 11:00-12:30 (Atrium)
Evaluations in machine learning rarely use the latest metrics, datasets, or human evaluation in favor of remaining compatible with prior work. The compatibility, often facilitated through leaderboards, thus leads to outdated but standardized evaluation practices. We pose that the standardization is taking place in the wrong spot. Evaluation infrastructure should enable researchers to use the latest methods and what should be standardized instead is how to incorporate these new evaluation advances. We introduce GEMv2, the new version of the Generation, Evaluation, and Metrics Benchmark which uses a modular infrastructure for dataset, model, and metric developers to benefit from each other's work. GEMv2 supports 40 documented datasets in 51 languages, ongoing online evaluation for all datasets, and our interactive tools make it easier to add new datasets to the living benchmark.

# Session 9 - 15:30-17:00

## Virtual Portal 7

15:30-17:00 (Hall A, Room A)

**Back to the Future: Bidirectional Information Decoupling Network for Multi-turn Dialogue Modeling**
*Yiyang Li, Hai Zhao and Zhuosheng Zhang* 15:30-17:00 (Hall A, Room A)
Multi-turn dialogue modeling as a challenging branch of natural language understanding (NLU), aims to build representations for machines to understand human dialogues, which provides a solid foundation for multiple downstream tasks. Recent studies of dialogue modeling commonly employ pre-trained language models (PrLMs) to encode the dialogue history as successive tokens, which is insufficient in capturing the temporal characteristics of dialogues. Therefore, we propose Bidirectional Information Decoupling Network (BiDeN) as a universal dialogue encoder, which explicitly incorporates both the past and future contexts and can be generalized to a wide range of dialogue-related tasks. Experimental results on datasets of different downstream tasks demonstrate the universality and effectiveness of our BiDeN.

**Neural-based Mixture Probabilistic Query Embedding for Answering FOL queries on Knowledge Graphs**
*xiao long, Liansheng Zhuang, Li Aodi, Shafei Wang and Houqiang Li* 15:30-17:00 (Hall A, Room A)
Query embedding (QE)—which aims to embed entities and first-order logical (FOL) queries in a vector space, has shown great power in answering FOL queries on knowledge graphs (KGs). Existing QE methods divide a complex query into a sequence of mini-queries according to its computation graph and perform logical operations on the answer sets of mini-queries to get answers. However, most of them assume that answer sets satisfy an individual distribution (e.g., Uniform, Beta, or Gaussian), which is often violated in real applications and limit their performance. In this paper, we propose a Neural-based Mixture Probabilistic Query Embedding Model (NMP-QEM) that encodes the answer set of each mini-query as a mixed Gaussian distribution with multiple means and covariance parameters, which can approximate any random distribution arbitrarily well in real KGs. Additionally, to overcome the difficulty in defining the closed solution of negation operation, we introduce neural-based logical operators of projection, intersection and negation for a mixed Gaussian distribution to answer all the FOL queries. Extensive experiments demonstrate that NMP-QEM significantly outperforms existing state-of-the-art methods on benchmark datasets. In NELL995, NMP-QEM achieves a 31

**Knowledge Prompting in Pre-trained Language Model for Natural Language Understanding**
*Jianing Wang, Wenkang Huang, Minghui Qiu, Qiuhui Shi, Hongbin Wang, Xiang Li and Ming Gao* 15:30-17:00 (Hall A, Room A)
Knowledge-enhanced Pre-trained Language Model (PLM) has recently received significant attention, which aims to incorporate factual knowledge into PLMs. However, most existing methods modify the internal structures of fixed types of PLMs by stacking complicated modules, and introduce redundant and irrelevant factual knowledge from knowledge bases (KBs). In this paper, to address these problems, we introduce a seminal knowledge prompting paradigm and further propose a knowledge-prompting-based PLM framework KP-PLM. This framework can be flexibly combined with existing mainstream PLMs. Specifically, we first construct a knowledge sub-graph from KBs for each context. Then we design multiple continuous prompts rules and transform the knowledge sub-graph into natural language prompts. To further leverage the factual knowledge from these prompts, we propose two novel knowledge-aware self-supervised tasks including prompt relevance inspection and masked prompt modeling. Extensive experiments on multiple natural language understanding (NLU) tasks show the superiority of

KP-PLM over other state-of-the-art methods in both full-resource and low-resource settings. Our source codes will be released upon the acceptance of the paper.

**Red Teaming Language Models with Language Models**
*Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese and Geoffrey Irving*
15:30-17:00 (Hall A, Room A)
Language Models (LMs) often cannot be deployed because of their potential to harm users in hard-to-predict ways. Prior work identifies harmful behaviors before deployment by using human annotators to hand-write test cases. However, human annotation is expensive, limiting the number and diversity of test cases. In this work, we automatically find cases where a target LM behaves in a harmful way, by generating test cases ("red teaming") using another LM. We evaluate the target LM's replies to generated test questions using a classifier trained to detect offensive content, uncovering tens of thousands of offensive replies in a 280B parameter LM chatbot. We explore several methods, from zero-shot generation to reinforcement learning, for generating test cases with varying levels of diversity and difficulty. Furthermore, we use prompt engineering to control LM-generated test cases to uncover a variety of other harms, automatically finding groups of people that the chatbot discusses in offensive ways, personal and hospital phone numbers generated as the chatbot's own contact info, leakage of private training data in generated text, and harms that occur over the course of a conversation. Overall, LM-based red teaming is one promising tool (among many needed) for finding and fixing diverse, undesirable LM behaviors before impacting users.

**CGoDial: A Large-Scale Benchmark for Chinese Goal-oriented Dialog Evaluation**
*Yinpei Dai, Wanwei He, Bowen Li, Yuchuan Wu, Zheng Cao, Zhongqi An, Jian Sun and Yongbin Li* 15:30-17:00 (Hall A, Room A)
Practical dialog systems need to deal with various knowledge sources, noisy user expressions, and the shortage of annotated data. To better solve the above problems, we propose CGoDial, a new challenging and comprehensive Chinese benchmark for multi-domain Goal-oriented Dialog evaluation. It contains 96,763 dialog sessions, and 574,949 dialog turns totally, covering three datasets with different knowledge sources: 1) a slot-based dialog (SBD) dataset with table-formed knowledge, 2) a flow-based dialog (FBD) dataset with tree-formed knowledge, and a retrieval-based dialog (RBD) dataset with candidate-formed knowledge. To bridge the gap between academic benchmarks and spoken dialog scenarios, we either collect data from real conversations or add spoken features to existing datasets via crowd-sourcing. The proposed experimental settings include the combinations of training with either the entire training set or a few-shot training set, and testing with either the standard test set or a hard test subset, which can assess model capabilities in terms of general prediction, fast adaptability and reliable robustness.

**Supervised Prototypical Contrastive Learning for Emotion Recognition in Conversation**
*Xiaohui Song, Longtao Huang, Hui Xue and Songlin Hu* 15:30-17:00 (Hall A, Room A)
Capturing emotions within a conversation plays an essential role in modern dialogue systems. However, the weak correlation between emotions and semantics brings many challenges to emotion recognition in conversation (ERC). Even semantically similar utterances, the emotion may vary drastically depending on contexts or speakers. In this paper, we propose a Supervised Prototypical Contrastive Learning (SPCL) loss for the ERC task. Leveraging the Prototypical Network, the SPCL targets at solving the imbalanced classification problem through contrastive learning and does not require a large batch size. Meanwhile, we design a difficulty measure function based on the distance between classes and introduce curriculum learning to alleviate the impact of extreme samples. We achieve state-of-the-art results on three widely used benchmarks. Further, we conduct analytical experiments to demonstrate the effectiveness of our proposed SPCL and curriculum learning strategy.

**Invariant Language Modeling**
*Maxime Peyrard, Sarvjeet Ghotra, Martin Josifoski, Vidhan Agarwal, Barun Patra, Dean Carignan, Emre Kiciman, Saurabh Tiwary and Robert West* 15:30-17:00 (Hall A, Room A)
Modern pretrained language models are critical components of NLP pipelines. Yet, they suffer from spurious correlations, poor out-of-domain generalization, and biases. Inspired by recent progress in causal machine learning, in particular the invariant risk minimization (IRM) paradigm, we propose invariant language modeling, a framework for learning invariant representations that generalize better across multiple environments. In particular, we adapt a game-theoretic implementation of IRM (IRM-games) to language models, where the invariance emerges from a specific training schedule in which all the environments compete to optimize their own environment-specific loss by updating subsets of the model in a round-robin fashion. We focused on controlled experiments to precisely demonstrate the ability of our method to (i) remove structured noise, (ii) ignore specific spurious correlations without affecting global performance, and (iii) achieve better out-of-domain generalization. These benefits come with a negligible computational overhead compared to standard training, do not require changing the local loss, and can be applied to any language model. We believe this framework is promising to help mitigate spurious correlations and biases in language models.

**Calibrating Student Models for Emotion-related Tasks**
*Mahshid Hosseini and Cornelia Caragea* 15:30-17:00 (Hall A, Room A)
Knowledge Distillation (KD) is an effective method to transfer knowledge from one network (a.k.a. teacher) to another (a.k.a. student). In this paper, we study KD on the emotion-related tasks from a new perspective: calibration. We further explore the impact of the mixup data augmentation technique on the distillation objective and propose to use a simple yet effective mixup method informed by training dynamics for calibrating the student models. Underpinned by the regularization impact of the mixup process by providing better training signals to the student models using training dynamics, our proposed mixup strategy gradually enhances the student model's calibration while effectively improving its performance. We evaluate the calibration of pre-trained language models through knowledge distillation over three tasks of emotion detection, sentiment analysis, and empathy detection. By conducting extensive experiments on different datasets, with both in-domain and out-of-domain test sets, we demonstrate that student models distilled from teacher models trained using our proposed mixup method obtained the lowest Expected Calibration Errors (ECEs) and best performance on both in-domain and out-of-domain test sets.

**Automatic Document Selection for Efficient Encoder Pretraining**
*Yukun Feng, Patrick Xia, Benjamin Van Durme and João Sedoc* 15:30-17:00 (Hall A, Room A)
Building pretrained language models is considered expensive and data-intensive, but must we increase dataset size to achieve better performance? We propose an alternative to larger training sets by automatically identifying smaller yet domain-representative subsets. We extend Cynical Data Selection, a statistical sentence scoring method that conditions on a representative target domain corpus. As an example, we treat the OntoNotes corpus as a target domain and pretrain a RoBERTa-like encoder from a cynically selected subset of the Pile. On both perplexity and across several downstream tasks in the target domain, it consistently outperforms random selection with 20x less data, 3x fewer training iterations, and 2x less estimated cloud compute cost, validating the recipe of automatic document selection for LM pretraining.

**Towards Efficient Dialogue Pre-training with Transferable and Interpretable Latent Structure**
*Xueliang Zhao, Lemao Liu, Tingchen Fu, Shuming Shi, Dongyan Zhao and Rui Yan* 15:30-17:00 (Hall A, Room A)
With the availability of massive general-domain dialogue data, pre-trained dialogue generation appears to be super appealing to transfer knowledge from the general domain to downstream applications. In most existing work, such transferable ability is mainly obtained by fitting

a large model with hundreds of millions of parameters on massive data in an exhaustive way, leading to inefficient running and poor interpretability. This paper proposes a novel dialogue generation model with a latent structure that is easily transferable from the general domain to downstream tasks in a lightweight and transparent way. Experiments on two benchmarks validate the effectiveness of the proposed model. Thanks to the transferable latent structure, our model is able to yield better dialogue responses than four strong baselines in terms of both automatic and human evaluations, and our model with about 22% parameters particularly delivers a 5x speedup in running time compared with the strongest baseline. Moreover, the proposed model is explainable by interpreting the discrete latent variables.

**FETA: A Benchmark for Few-Sample Task Transfer in Open-Domain Dialogue**
*Alon Albalak, Yi-Lin Tuan, Pegah Jandaghi, Connor Pryor, Luke Yoffe, Deepak Ramachandran, Lise Getoor, Jay Pujara and William Yang Wang* 15:30-17:00 (Hall A, Room A)
Task transfer, transferring knowledge contained in related tasks, holds the promise of reducing the quantity of labeled data required to fine-tune language models. Dialogue understanding encompasses many diverse tasks, yet task transfer has not been thoroughly studied in conversational AI. This work explores conversational task transfer by introducing FETA: a benchmark for FEw-sample TAsk transfer in open-domain dialogue. FETA contains two underlying sets of conversations upon which there are 10 and 7 tasks annotated, enabling the study of intra-dataset task transfer; task transfer without domain adaptation. We utilize three popular language models and three learning algorithms to analyze the transferability between 132 source-target task pairs and create a baseline for future work. We run experiments in the single- and multi-source settings and report valuable findings, e.g., most performance trends are model-specific, and span extraction and multiple-choice tasks benefit the most from task transfer. In addition to task transfer, FETA can be a valuable resource for future research into the efficiency and generalizability of pre-training datasets and model architectures, as well as for learning settings such as continual and multitask learning.

**$IM^2$: an Interpretable and Multi-category Integrated Metric Framework for Automatic Dialogue Evaluation**
*Zhihua Jiang, Guanghui Ye, Dongning Rao, Di Wang and Xin Miao* 15:30-17:00 (Hall A, Room A)
Evaluation metrics shine the light on the best models and thus strongly influence the research directions, such as the recently developed dialogue metrics USR, FED, and GRADE. However, most current metrics evaluate the dialogue data as isolated and static because they only focus on a single quality or several qualities. To mitigate the problem, this paper proposes an interpretable, multi-faceted, and controllable framework $IM^2$ (Interpretable and Multi-category Integrated Metric) to combine a large number of metrics which are good at measuring different qualities. The $IM^2$ framework first divides current popular dialogue qualities into different categories and then applies or proposes dialogue metrics to measure the qualities within each category and finally generates an overall $IM^2$ score. An initial version of $IM^2$ was submitted to the AAAI 2022 Track5.1@DSTC10 challenge and took the $2^{nd}$ place on both of the development and test leaderboard. After the competition, we develop more metrics and improve the performance of our model. We compare $IM^2$ with other 13 current dialogue metrics and experimental results show that $IM^2$ correlates more strongly with human judgments than any of them on each evaluated dataset.

**Finding Skill Neurons in Pre-trained Transformer-based Language Models**
*Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu and Juanzi Li* 15:30-17:00 (Hall A, Room A)
Transformer-based pre-trained language models have demonstrated superior performance on various natural language processing tasks. However, it remains unclear how the skills required to handle these tasks distribute among model parameters. In this paper, we find that after prompt tuning for specific tasks, the activations of some neurons within pre-trained Transformers are highly predictive of the task labels. We dub these neurons skill neurons and confirm they encode task-specific skills by finding that: (1) Skill neurons are crucial for handling tasks. Performances of pre-trained Transformers on a task significantly drop when corresponding skill neurons are perturbed. (2) Skill neurons are task-specific. Similar tasks tend to have similar distributions of skill neurons. Furthermore, we demonstrate the skill neurons are most likely generated in pre-training rather than fine-tuning by showing that the skill neurons found with prompt tuning are also crucial for other fine-tuning methods freezing neuron weights, such as the adapter-based tuning and BitFit. We also explore the applications of skill neurons, including accelerating Transformers with network pruning and building better transferability indicators. These findings may promote further research on understanding Transformers. The source code can be obtained from https://github.com/THU-KEG/Skill-Neuron.

**COLD: A Benchmark for Chinese Offensive Language Detection**
*Jiawen Deng, Jingyan ZHOU, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng and Minlie Huang* 15:30-17:00 (Hall A, Room A)
Offensive language detection is increasingly crucial for maintaining a civilized social media platform and deploying pre-trained language models. However, this task in Chinese is still under exploration due to the scarcity of reliable datasets. To this end, we propose a benchmark –COLD for Chinese offensive language analysis, including a Chinese Offensive Language Dataset –COLDATASET and a baseline detector –COLDETECTOR which is trained on the dataset. We show that the COLD benchmark contributes to Chinese offensive language detection which is challenging for existing resources. We then deploy the COLDETECTOR and conduct detailed analyses on popular Chinese pre-trained language models. We first analyze the offensiveness of existing generative models and show that these models inevitably expose varying degrees of offensive issues. Furthermore, we investigate the factors that influence the offensive generations, and we find that anti-bias contents and keywords referring to certain groups or revealing negative attitudes trigger offensive outputs easier.

**Model Criticism for Long-Form Text Generation**
*Yuntian Deng, Volodymyr Kuleshov and Alexander Rush* 15:30-17:00 (Hall A, Room A)
Language models have demonstrated the ability to generate highly fluent text; however, it remains unclear whether their output retains coherent high-level structure (e.g., story progression). Here, we propose to apply a statistical tool, model criticism in latent space, to evaluate the high-level structure of the generated text. Model criticism compares the distributions between real and generated data in a latent space obtained according to an assumptive generative process. Different generative processes identify specific failure modes of the underlying model. We perform experiments on three representative aspects of high-level discourse—coherence, coreference, and topicality—and find that transformer-based language models are able to capture topical structures but have a harder time maintaining structural coherence or modeling coreference.

[INDUSTRY] **DynaMaR: Dynamic Prompt with Mask Token Representation**
*Xiaodi Sun, Sunny Rajagopalan, Priyanka Nigam, Weiyi Lu, Yi Xu, Iman Keivanloo, Belinda Zeng and Trishul Chilimbi* 15:30-17:00 (Hall A, Room A)
Recent research has shown that large language models pretrained using unsupervised approaches can achieve significant performance improvement on many downstream tasks. Typically when adapting these language models to downstream tasks, like a classification or regression task, we employ a fine-tuning paradigm in which the sentence representation from the language model is input to a task-specific head; the model is then fine-tuned end-to-end. However, with the emergence of models like GPT-3, prompt-based fine-tuning has been proven to be a successful approach for few-shot tasks. Inspired by this work, we study discrete prompt technologies in practice. There are two issues that arise with the standard prompt approach. First, it can overfit on the prompt template. Second, it requires manual effort to formulate the downstream task as a language model problem. In this paper, we propose an improvement to prompt-based fine-tuning that addresses these two issues. We refer to our approach as DynaMaR – Dynamic Prompt with Mask Token Representation. Results show that DynaMaR can achieve an

average improvement of 10% in few-shot settings and improvement of 3.7% in data-rich settings over the standard fine-tuning approach on four e-commerce applications.

[INDUSTRY] **PENTATRON: PErsonalized coNText-Aware Transformer for Retrieval-based cOnversational uNderstanding**
*Niranjan Uma Naresh, Ziyan Jiang, Ankit Ankit, Sungjin Lee, Jie Hao, Xing Fan and Chenlei Guo* 15:30-17:00 (Hall A, Room A)
Conversational understanding is an integral part of modern intelligent devices. In a large fraction of the global traffic from customers using smart digital assistants, frictions in dialogues may be attributed to incorrect understanding of the entities in a customer's query due to factors including ambiguous mentions, mispronunciation, background noise and faulty on-device signal processing. Such errors are compounded by two common deficiencies from intelligent devices namely, (1) the device not being tailored to individual customers, and (2) the device responses being unaware of the context in the conversation session. Viewing this problem via the lens of retrieval-based search engines, we build and evaluate a scalable entity correction system, PENTATRON. The system leverages a parametric transformer-based language model to learn patterns from in-session customer-device interactions coupled with a non-parametric personalized entity index to compute the correct query, which aids downstream components in reasoning about the best response. In addition to establishing baselines and demonstrating the value of personalized and context-aware systems, we use multitasking to learn the domain of the correct entity. We also investigate the utility of language model prompts. Through extensive experiments, we show a significant upward movement of the key metric (Exact Match) by up to 500.97% (relative to the baseline).

[INDUSTRY] **Ask-and-Verify: Span Candidate Generation and Verification for Attribute Value Extraction**
*Yifan Ding, Yan Liang, Nasser Zalmout, Xian Li, Christan Grant and Tim Weninger* 15:30-17:00 (Hall A, Room A)
The product attribute value extraction (AVE) task aims to capture key factual information from product profiles, and is useful for several downstream applications in e-Commerce platforms. Previous contributions usually formulate this task using sequence labeling or reading comprehension architectures. However, sequence labeling models tend to be conservative in their predictions resulting in a high false negative rate. Existing reading comprehension formulations, on the other hand, can over-generate attribute values which hinders precision. In the present work we address these limitations with a new end-to-end pipeline framework called Ask-and-Verify. Given a product and an attribute query, the Ask step detects the top-K span candidates (i.e. possible attribute values) from the product profiles, then the Verify step filters out false positive candidates. We evaluate Ask-and-Verify model on Amazon's product pages and AliExpress public dataset, and present a comparative analysis as well as a detailed ablation study. Despite its simplicity, we show that Ask-and-Verify outperforms recent state-of-the-art models by up to 3.1% F1 absolute improvement points, while also scaling to thousands of attributes.

[INDUSTRY] **Deploying a Retrieval based Response Model for Task Oriented Dialogues**
*Lahari Poddar, György Szarvas, Cheng Wang, Jorge Balazs, Pavel Danchenko and Patrick Ernst* 15:30-17:00 (Hall A, Room A)
Task-oriented dialogue systems in industry settings need to have high conversational capability, be easily adaptable to changing situations and conform to business constraints. This paper describes a 3-step procedure to develop a conversational model that satisfies these criteria and can efficiently scale to rank a large set of response candidates. First, we provide a simple algorithm to semi-automatically create a high-coverage template set from historic conversations without any annotation. Second, we propose a neural architecture that encodes the dialogue context and applicable business constraints as profile features for ranking the next turn. Third, we describe a two-stage learning strategy with self-supervised training, followed by supervised fine-tuning on limited data collected through a human-in-the-loop platform. Finally, we describe offline experiments and present results of deploying our model with human-in-the-loop to converse with live customers online.

[INDUSTRY] **SLATE: A Sequence Labeling Approach for Task Extraction from Free-form Inked Content**
*Apurva Gandhi, Ryan Serrao, Biyi Fang, Gilbert Antonius, Jenna Hong, Tra My Nguyen, Sheng Yi, Ehi Nosakhare, Irene Shaffer, Soundararajan Srinivasan and Vivek Gupta* 15:30-17:00 (Hall A, Room A)
We present SLATE, a sequence labeling approach for extracting tasks from free-form content such as digitally handwritten (or "inked") notes on a virtual whiteboard. Our approach allows us to create a single, low-latency model to simultaneously perform sentence segmentation and classification of these sentences into task/non-task sentences. SLATE greatly outperforms a baseline two-model (sentence segmentation followed by classification model) approach, achieving a task F1 score of 84.4%, a sentence segmentation (boundary similarity) score of 88.4% and three times lower latency compared to the baseline. Furthermore, we provide insights into tackling challenges of performing NLP on the inking domain. We release both our code and dataset for this novel task.

[INDUSTRY] **Meta-learning Pathologies from Radiology Reports using Variance Aware Prototypical Networks**
*Arijit Sehanobish, Kawshik Kannan, Nabila Abraham, Anasuya Das and Benjamin Odry* 15:30-17:00 (Hall A, Room A)
Large pretrained Transformer-based language models like BERT and GPT have changed the landscape of Natural Language Processing (NLP). However, fine tuning such models still requires a large number of training examples for each target task, thus annotating multiple datasets and training these models on various downstream tasks becomes time consuming and expensive. In this work, we propose a simple extension of the Prototypical Networks for few-shot text classification. Our main idea is to replace the class prototypes by Gaussians and introduce a regularization term that encourages the examples to be clustered near the appropriate class centroids. Experimental results show that our method outperforms various strong baselines on 13 public and 4 internal datasets. Furthermore, we use the class distributions as a tool for detecting potential out-of-distribution (OOD) data points during deployment.

[INDUSTRY] **Bringing the State-of-the-Art to Customers: A Neural Agent Assistant Framework for Customer Service Support**
*Stephen Chinedu Obadinma, Faiza Khan Khattak, Shirley Wang, Tania Sidhorn, Elaine Lau, Sean Robertson, Jingcheng Niu, Winnie Au, Alif Munim, Karthik Raja Kalaiselvi Bhaskar, Bencheng Wei, Iris Ren, Muhammad Waqar, Erin Li, Bukola Ishola, Michael Wang, Griffin Tanner, Yu-Jia Shiah, Sean X. Zhang, Kwesi Apponsah, Kanishk Patel, Jaswinder Narain, Pandya Deval, Xiaodan Zhu, Frank Rudzicz and Elham Dolatabadi* 15:30-17:00 (Hall A, Room A)
Building Agent Assistants that can help improve customer service support requires inputs from industry users and their customers, as well as knowledge about state-of-the-art Natural Language Processing (NLP) technology. We combine expertise from academia and industry to bridge the gap and build task/domain-specific Neural Agent Assistants (NAA) with three high-level components for: (1) Intent Identification, (2) Context Retrieval, and (3) Response Generation. In this paper, we outline the pipeline of the NAA's core system and also present three case studies in which three industry partners successfully adapt the framework to find solutions to their unique challenges. Our findings suggest that a collaborative process is instrumental in spurring the development of emerging NLP models for Conversational AI tasks in industry. The full reference implementation code and results are available at https://github.com/VectorInstitute/NAA.

[DEMO] **EasyNLP: A Comprehensive and Easy-to-use Toolkit for Natural Language Processing**
*Chengyu Wang, Minghui Qiu, Taolin Zhang, Tingting Liu, Lei Li, Jianing Wang, Ming Wang, jun huang and Wei Lin* 15:30-17:00 (Hall A, Room A)
Pre-Trained Models (PTMs) have reshaped the development of Natural Language Processing (NLP) and achieved significant improvement in various benchmarks. Yet, it is not easy for industrial practitioners to obtain high-performing PTM-based models without a large amount of labeled training data and deploy them online with fast inference speed. To bridge this gap, EasyNLP is designed to make it easy to build NLP applications, which supports a comprehensive suite of NLP algorithms. It further features knowledge-enhanced pre-training, knowledge

distillation and few-shot learning functionalities, and provides a unified framework of model training, inference and deployment for real-world applications. EasyNLP has powered over ten business units within Alibaba Group and is seamlessly integrated to the Platform of AI (PAI) products on Alibaba Cloud. The source code of EasyNLP is released at GitHub (https://github.com/alibaba/EasyNLP).

## Virtual Portal 8

15:30-17:00 (Hall A, Room B)

**ConvTrans: Transforming Web Search Sessions for Conversational Dense Retrieval**
*Kelong Mao, Zhicheng Dou, Hongjin Qian, Fengran Mo, Xiaohua Cheng and Zhao Cao* 15:30-17:00 (Hall A, Room B)
Conversational search provides users with a natural and convenient new search experience. Recently, conversational dense retrieval has shown to be a promising technique for realizing conversational search. However, as conversational search systems have not been widely deployed, it is hard to get large-scale real conversational search sessions and relevance labels to support the training of conversational dense retrieval. To tackle this data scarcity problem, previous methods focus on developing better few-shot learning approaches or generating pseudo relevance labels, but the data they use for training still heavily rely on manual generation.
In this paper, we present ConvTrans, a data augmentation method that can automatically transform easily-accessible web search sessions into conversational search sessions to fundamentally alleviate the data scarcity problem for conversational dense retrieval. ConvTrans eliminates the gaps between these two types of sessions in terms of session quality and query form to achieve effective session transformation. Extensive evaluations on two widely used conversational search benchmarks, i.e., CAsT-19 and CAsT-20, demonstrate that the same model trained on the data generated by ConvTrans can achieve comparable retrieval performance as it trained on high-quality but expensive artificial conversational search data.

**Improving Multi-task Stance Detection with Multi-task Interaction Network**
*Heyan Chai, Siyu Tang, Jinhao Cui, Ye Ding, Binxing Fang and Qing Liao* 15:30-17:00 (Hall A, Room B)
Stance detection aims to identify people's standpoints expressed in the text towards a target, which can provide powerful information for various downstream tasks. Recent studies have proposed multi-task learning models that introduce sentiment information to boost stance detection. However, they neglect to explore capturing the fine-grained task-specific interaction between stance detection and sentiment tasks, thus degrading performance. To address this issue, this paper proposes a novel multi-task interaction network (MTIN) for improving the performance of stance detection and sentiment analysis tasks simultaneously. Specifically, we construct heterogeneous task-related graphs to automatically identify and adapt the roles that a word plays with respect to a specific task. Also, a multi-task interaction module is designed to capture the word-level interaction between tasks, so as to obtain richer task representations. Extensive experiments on two real-world datasets show that our proposed approach outperforms state-of-the-art methods in both stance detection and sentiment analysis tasks.

**Generative Entity Typing with Curriculum Learning**
*Siyu Yuan, Deqing Yang, Jiaqing Liang, Zhixu Li, Jinxi Liu, Jingyue Huang and Yanghua Xiao* 15:30-17:00 (Hall A, Room B)
Entity typing aims to assign types to the entity mentions in given texts. The traditional classification-based entity typing paradigm has two unignorable drawbacks: 1) it fails to assign an entity to the types beyond the predefined type set, and 2) it can hardly handle few-shot and zero-shot situations where many long-tail types only have few or even no training instances. To overcome these drawbacks, we propose a novel generative entity typing (GET) paradigm: given a text with an entity mention, the multiple types for the role that the entity plays in the text are generated with a pre-trained language model (PLM). However, PLMs tend to generate coarse-grained types after fine-tuning upon the entity typing dataset. In addition, only the heterogeneous training data consisting of a small portion of human-annotated data and a large portion of auto-generated but low-quality data are provided for model training. To tackle these problems, we employ curriculum learning (CL) to train our GET model on heterogeneous data, where the curriculum could be self-adjusted with the self-paced learning according to its comprehension of the type granularity and data heterogeneity. Our extensive experiments upon the datasets of different languages and downstream tasks justify the superiority of our GET model over the state-of-the-art entity typing models. The code has been released on https://github.com/siyuyuan/GET.

**A Unified Positive-Unlabeled Learning Framework for Document-Level Relation Extraction with Different Levels of Labeling**
*Ye Wang, Xinxin Liu, Wenxin Hu and Tao Zhang* 15:30-17:00 (Hall A, Room B)
Document-level relation extraction (RE) aims to identify relations between entities across multiple sentences. Most previous methods focused on document-level RE under full supervision. However, in real-world scenario, it is expensive and difficult to completely label all relations in a document because the number of entity pairs in document-level RE grows quadratically with the number of entities. To solve the common incomplete labeling problem, we propose a unified positive-unlabeled learning framework - shift and squared ranking loss positive-unlabeled (SSR-PU) learning. We use positive-unlabeled (PU) learning on document-level RE for the first time. Considering that labeled data of a dataset may lead to prior shift of unlabeled data, we introduce a PU learning under prior shift of training data. Also, using none-class score as an adaptive threshold, we propose squared ranking loss and prove its Bayesian consistency with multi-label ranking metrics. Extensive experiments demonstrate that our method achieves an improvement of about 14 F1 points relative to the previous baseline with incomplete labeling. In addition, it outperforms previous state-of-the-art results under both fully supervised and extremely unlabeled settings as well.

**Retrieval-Augmented Generative Question Answering for Event Argument Extraction**
*Xinya Du and Heng Ji* 15:30-17:00 (Hall A, Room B)
Event argument extraction has long been studied as a sequential prediction problem with extractive-based methods, tackling each argument in isolation. Although recent work proposes generation-based methods to capture cross-argument dependency, they require generating and post-processing a complicated target sequence (template). Motivated by these observations and recent pretrained language models' capabilities of learning from demonstrations. We propose a retrieval-augmented generative QA model (R-GQA) for event argument extraction. It retrieves the most similar QA pair and augments it as prompt to the current example's context, then decodes the arguments as answers. Our approach outperforms substantially prior methods across various settings (i.e. fully supervised, domain transfer, and fewshot learning). Finally, we propose a clustering-based sampling strategy (JointEnc) and conduct a thorough analysis of how different strategies influence the few-shot learning performances.

**Explicit Query Rewriting for Conversational Dense Retrieval**
*Hongjin Qian and Zhicheng Dou* 15:30-17:00 (Hall A, Room B)
In a conversational search scenario, a query might be context-dependent because some words are referred to previous expressions or omitted. Previous works tackle the issue by either reformulating the query into a self-contained query (query rewriting) or learning a contextualized query embedding from the query context (context modelling). In this paper, we propose a model CRDR that can perform query rewriting and context modelling in a unified framework in which the query rewriting's supervision signals further enhance the context modelling. In-

stead of generating a new query, CRDR only performs necessary modifications on the original query, which improves both accuracy and efficiency of query rewriting. In the meantime, the query rewriting benefits the context modelling by explicitly highlighting relevant terms in the query context, which improves the quality of the learned contextualized query embedding. To verify the effectiveness of CRDR, we perform comprehensive experiments on TREC CAsT-19 and TREC CAsT-20 datasets, and the results show that our method outperforms all baseline models in terms of both quality of query rewriting and quality of context-aware ranking.

**Syntactically Rich Discriminative Training: An Effective Method for Open Information Extraction**
*Frank Mtumbuka and Thomas Lukasiewicz* 15:30-17:00 (Hall A, Room B)
Open information extraction (OIE) is the task of extracting facts "(Subject, Relation, Object)" from natural language text. We propose several new methods for training neural OIE models in this paper. First, we propose a novel method for computing syntactically rich text embeddings using the structure of dependency trees. Second, we propose a new discriminative training approach to OIE in which tokens in the generated fact are classified as "real" or "fake", i.e., those tokens that are in both the generated and gold tuples, and those that are only in the generated tuple but not in the gold tuple. We also address the issue of repetitive tokens in generated facts and improve the models' ability to generate implicit facts. Our approach reduces repetitive tokens by a factor of 23%. Finally, we present paraphrased versions of the CaRB, OIE2016, and LSOIE datasets, and show that the models' performance substantially improves when trained on augmented datasets. Our best model beats the SOTA of IMoJIE on the recent CaRB dataset, with an improvement of 39.63% in F1 score.

**Reduce Catastrophic Forgetting of Dense Retrieval Training with Teleportation Negatives**
*Si Sun, Chenyan Xiong, Yue Yu, Arnold Overwijk, Zhiyuan Liu and Jie Bao* 15:30-17:00 (Hall A, Room B)
In this paper, we investigate the instability in the standard dense retrieval training, which iterates between model training and hard negative selection using the being-trained model. We show the catastrophic forgetting phenomena behind the training instability, where models learn and forget different negative groups during training iterations. We then propose ANCE-Tele, which accumulates momentum negatives from past iterations and approximates future iterations using lookahead negatives, as "teleportations" along the time axis to smooth the learning process. On web search and OpenQA, ANCE-Tele outperforms previous state-of-the-art systems of similar size, eliminates the dependency on sparse retrieval negatives, and is competitive among systems using significantly more (50x) parameters. Our analysis demonstrates that teleportation negatives reduce catastrophic forgetting and improve convergence speed for dense retrieval training. The source code of this paper is available at https://github.com/OpenMatch/ANCE-Tele.

**Improving Event Coreference Resolution Using Document-level and Topic-level Information**
*Sheng Xu, Peifeng Li and Qiaoming Zhu* 15:30-17:00 (Hall A, Room B)
Event coreference resolution (ECR) aims to cluster event mentions that refer to the same real-world events. Deep learning methods have achieved SOTA results on the ECR task. However, due to the encoding length limitation, previous methods either adopt classical pairwise models based on sentence-level context or split each document into multiple chunks and encode them separately. They failed to capture the interactions and contextual cues among those long-distance event mentions. Besides, high-level information, such as event topics, is rarely considered to enhance representation learning for ECR. To address the above two issues, we first apply a Longformer-based encoder to obtain the document-level embeddings and an encoder with a trigger-mask mechanism to learn sentence-level embeddings based on local context. In addition, we propose an event topic generator to infer the latent topic-level representations. Finally, using the above event embeddings, we employ a multiple tensor matching method to capture their interactions at the document, sentence, and topic levels. Experimental results on the KBP 2017 dataset show that our model outperforms the SOTA baselines.

**Modeling Label Correlations for Ultra-Fine Entity Typing with Neural Pairwise Conditional Random Field**
*Chengyue Jiang, Yong Jiang, Weiqi Wu, Pengjun Xie and Kewei Tu* 15:30-17:00 (Hall A, Room B)
Ultra-fine entity typing (UFET) aims to predict a wide range of type phrases that correctly describe the categories of a given entity mention in a sentence. Most recent works infer each entity type independently, ignoring the correlations between types, e.g., when an entity is inferred as a *president*, it should also be a *politician* and a *leader*. To this end, we use an undirected graphical model called pairwise conditional random field (PCRF) to formulate the UFET problem, in which the type variables are not only unarily influenced by the input but also pairwisely relate to all the other type variables. We use various modern backbones for entity typing to compute unary potentials, and derive pairwise potentials from type phrase representations that both capture prior semantic information and facilitate accelerated inference. We use mean-field variational inference for efficient type inference on very large type sets and unfold it as a neural network module to enable end-to-end training. Experiments on UFET show that the Neural-PCRF consistently outperforms its backbones with little cost and results in a competitive performance against cross-encoder based SOTA while being *thousands of times* faster. We also find Neural-PCRF effective on a widely used fine-grained entity typing dataset with a smaller type set. We pack Neural-PCRF as a network module that can be plugged onto multi-label type classifiers with ease and release it in code.

**An Adaptive Logical Rule Embedding Model for Inductive Reasoning over Temporal Knowledge Graphs**
*Xin Mei, Libin Yang, Xiaoyan Cai and Zuowei Jiang* 15:30-17:00 (Hall A, Room B)
Temporal knowledge graphs (TKGs) extrapolation reasoning predicts future events based on historical information, which has great research significance and broad application value. Existing methods can be divided into embedding-based methods and logical rule-based methods. Embedding-based methods rely on learned entity and relation embeddings to make predictions and thus lack interpretability. Logical rule-based methods bring scalability problems due to being limited by the learned logical rules. We combine the two methods to capture deep causal logic by learning rule embeddings, and propose an interpretable model for temporal knowledge graph reasoning called adaptive logical rule embedding model for inductive reasoning (ALRE-IR). ALRE-IR can adaptively extract and assess reasons contained in historical events, and make predictions based on causal logic. Furthermore, we propose a one-class augmented matching loss for optimization. When evaluated on the ICEWS14, ICEWS0515 and ICEWS18 datasets, the performance of ALRE-IR outperforms other state-of-the-art baselines. The results also demonstrate that ALRE-IR still shows outstanding performance when transferred to related dataset with common relation vocabulary, indicating our proposed model has good zero-shot reasoning ability.

**Query-based Instance Discrimination Network for Relational Triple Extraction**
*Zeqi Tan, Yongliang Shen, Xuming Hu, Wenqi Zhang, Xiaoxia Cheng, Weiming Lu and Yueting Zhuang* 15:30-17:00 (Hall A, Room B)
Joint entity and relation extraction has been a core task in the field of information extraction. Recent approaches usually consider the extraction of relational triples from a stereoscopic perspective, either learning a relation-specific tagger or separate classifiers for each relation type. However, they still suffer from error propagation, relation redundancy and lack of high-level connections between triples. To address these issues, we propose a novel query-based approach to construct instance-level representations for relational triples. By metric-based comparison between query embeddings and token embeddings, we can extract all types of triples in one step, thus eliminating the error propagation problem. In addition, we learn the instance-level representation of relational triples via contrastive learning. In this way, relational triples can not only enclose rich class-level semantics but also access to high-order global connections. Experimental results show that our proposed method achieves the state of the art on five widely used benchmarks.

**Towards Better Document-level Relation Extraction via Iterative Inference**

*Liang Zhang, Jinsong Su, Yidong Chen, Zhongjian Miao, Min Zijun, Qingguo Hu and xiaodong shi* 15:30-17:00 (Hall A, Room B)
Document-level relation extraction (RE) aims to extract the relations between entities from the input document that usually containing many difficultly-predicted entity pairs whose relations can only be predicted through relational inference. Existing methods usually directly predict the relations of all entity pairs of input document in a one-pass manner, ignoring the fact that predictions of some entity pairs heavily depend on the predicted results of other pairs. To deal with this issue, in this paper, we propose a novel document-level RE model with iterative inference. Our model is mainly composed of two modules: 1) a base module expected to provide preliminary relation predictions on entity pairs; 2) an inference module introduced to refine these preliminary predictions by iteratively dealing with difficultly-predicted entity pairs depending on other pairs in an easy-to-hard manner. Unlike previous methods which only consider feature information of entity pairs, our inference module is equipped with two Extended Cross Attention units, allowing it to exploit both feature information and previous predictions of entity pairs during relational inference. Furthermore, we adopt a two-stage strategy to train our model. At the first stage, we only train our base module. During the second stage, we train the whole model, where contrastive learning is introduced to enhance the training of inference module. Experimental results on three commonly-used datasets show that our model consistently outperforms other competitive baselines.

**Learning Cross-Task Dependencies for Joint Extraction of Entities, Events, Event Arguments, and Relations**
*Minh Van Nguyen, Bonan Min, Franck Dernoncourt and Thien Nguyen* 15:30-17:00 (Hall A, Room B)
Extracting entities, events, event arguments, and relations (i.e., task instances) from text represents four main challenging tasks in information extraction (IE), which have been solved jointly (JointIE) to boost the overall performance for IE. As such, previous work often leverages two types of dependencies between the tasks, i.e., cross-instance and cross-type dependencies representing relatedness between task instances and correlations between information types of the tasks. However, the cross-task dependencies in prior work are not optimal as they are only designed manually according to some task heuristics. To address this issue, we propose a novel model for JointIE that aims to learn cross-task dependencies from data. In particular, we treat each task instance as a node in a dependency graph where edges between the instances are inferred through information from different layers of a pretrained language model (e.g., BERT). Furthermore, we utilize the Chow-Liu algorithm to learn a dependency tree between information types for JointIE by seeking to approximate the joint distribution of the types from data. Finally, the Chow-Liu dependency tree is used to generate cross-type patterns, serving as anchor knowledge to guide the learning of representations and dependencies between instances for JointIE. Experimental results show that our proposed model significantly outperforms strong JointIE baselines over four datasets with different languages.

**Entity-centered Cross-document Relation Extraction**
*Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji and Bo Cai* 15:30-17:00 (Hall A, Room B)
Relation Extraction (RE) is a fundamental task of information extraction, which has attracted a large amount of research attention. Previous studies focus on extracting the relations within a sentence or document, while currently researchers begin to explore cross-document RE. However, current cross-document RE methods directly utilize text snippets surrounding target entities in multiple given documents, which brings considerable noisy and non-relevant sentences. Moreover, they utilize all the text paths in a document bag in a coarse-grained way, without considering the connections between these text paths.In this paper, we aim to address both of these shortages and push the state-of-the-art for cross-document RE. First, we focus on input construction for our RE model and propose an entity-based document-context filter to retain useful information in the given documents by using the bridge entities in the text paths. Second, we propose a cross-document RE model based on cross-path entity relation attention, which allow the entity relations across text paths to interact with each other. We compare our cross-document RE method with the state-of-the-art methods in the dataset CodRED. Our method outperforms them by at least 10% in F1, thus demonstrating its effectiveness.

[DEMO] **DeepKE: A Deep Learning Based Knowledge Extraction Toolkit for Knowledge Base Population**
*Ningyu Zhang, Xin Xu, Liankuan Tao, Haiyang Yu, Hongbin Ye, Shuofei Qiao, Xin Xie, Xiang Chen, Zhoubo Li, Lei Li, Xiaozhuan Liang, Yunzhi Yao, Shumin Deng, Peng Wang, Wen Zhang, Zhenru Zhang, Chuanqi Tan, Qiang Chen, Feiyu Xiong, Fei Huang, Guozhou Zheng and Huajun Chen* 15:30-17:00 (Hall A, Room B)
We present an open-source and extensible knowledge extraction toolkit DeepKE, supporting complicated low-resource, document-level and multimodal scenarios in the knowledge base population. DeepKE implements various information extraction tasks, including named entity recognition, relation extraction and attribute extraction. With a unified framework, DeepKE allows developers and researchers to customize datasets and models to extract information from unstructured data according to their requirements. Specifically, DeepKE not only provides various functional modules and model implementation for different tasks and scenarios but also organizes all components by consistent frameworks to maintain sufficient modularity and extensibility. We release the source code at GitHub in https://github.com/zjunlp/DeepKE with Google Colab tutorials and comprehensive documents for beginners. Besides, we present an online system in `http://deepke.openkg.cn/EN/re_doc_show.html` for real-time extraction of various tasks, and a demo video.

## Virtual Portal 9

15:30-17:00 (Hall A, Room C)

**Conformal Predictor for Improving Zero-Shot Text Classification Efficiency**
*Prafulla Kumar Choubey, Yu Bai, Chien-Sheng Wu, Wenhao Liu and Nazneen Rajani* 15:30-17:00 (Hall A, Room C)
Pre-trained language models (PLMs) have been shown effective for zero-shot (0shot) text classification. 0shot models based on natural language inference (NLI) and next sentence prediction (NSP) employ cross-encoder architecture and infer by making a forward pass through the model for each label-text pair separately. This increases the computational cost to make inferences linearly in the number of labels. In this work, we improve the efficiency of such cross-encoder-based 0shot models by restricting the number of likely labels using another fast base classifier-based conformal predictor (CP) calibrated on samples labeled by the 0shot model. Since a CP generates prediction sets with coverage guarantees, it reduces the number of target labels without excluding the most probable label based on the 0shot model. We experiment with three intent and two topic classification datasets. With a suitable CP for each dataset, we reduce the average inference time for NLI- and NSP-based models by 25.6% and 22.2% respectively, without dropping performance below the predefined error rate of 1%.

**Multi-level Distillation of Semantic Knowledge for Pre-training Multilingual Language Model**
*Mingqi Li, Fei Ding, Dan Zhang, Long Cheng, Hongxin Hu and feng luo* 15:30-17:00 (Hall A, Room C)
Pre-trained multilingual language models play an important role in cross-lingual natural language understanding tasks. However, existing methods did not focus on learning the semantic structure of representation, and thus could not optimize their performance. In this paper, we propose Multi-level Multilingual Knowledge Distillation (MMKD), a novel method for improving multilingual language models. Specifically, we employ a teacher-student framework to adopt rich semantic representation knowledge in English BERT. We propose token-, word-, sentence-, and structure-level alignment objectives to encourage multiple levels of consistency between source-target pairs and correlation

similarity between teacher and student models. We conduct experiments on cross-lingual evaluation benchmarks including XNLI, PAWS-X, and XQuAD. Experimental results show that MMKD outperforms other baseline models of similar size on XNLI and XQuAD and obtains comparable performance on PAWS-X. Especially, MMKD obtains significant performance gains on low-resource languages.

**Improving Stability of Fine-Tuning Pretrained Language Models via Component-Wise Gradient Norm Clipping**
*Chenghao Yang and Xuezhe Ma* 15:30-17:00 (Hall A, Room C)
Fine-tuning over large pretrained language models (PLMs) has established many state-of-the-art results. Despite its superior performance, such fine-tuning can be unstable, resulting in significant variance in performance and potential risks for practical applications. Previous works have attributed such instability to the catastrophic forgetting problem in the top layers of PLMs, which indicates iteratively fine-tuning layers in a top-down manner is a promising solution. In this paper, we first point out that this method does not always work out due to the different convergence speeds of different layers/modules. Inspired by this observation, we propose a simple component-wise gradient norm clipping method to adjust the convergence speed for different components. Experiment results demonstrate that our method achieves consistent improvements in terms of generalization performance, convergence speed, and training stability. The codebase can be found at https://github.com/yangalan123/FineTuningStability.

**Norm-based Noisy Corpora Filtering and Refurbishing in Neural Machine Translation**
*Yu Lu and Jiajun Zhang* 15:30-17:00 (Hall A, Room C)
Recent advances in neural machine translation depend on massive parallel corpora, which are collected from any open source without much guarantee of quality. It stresses the need for noisy corpora filtering, but existing methods are insufficient to solve this issue. They spend much time ensembling multiple scorers trained on clean bitexts, unavailable for low-resource languages in practice. In this paper, we propose a norm-based noisy corpora filtering and refurbishing method with no external data and costly scorers. The noisy and clean samples are separated based on how much information from the source and target sides the model requires to fit the given translation. For the unparallel sentence, the target-side history translation is much more important than the source context, contrary to the parallel ones. The amount of these two information flows can be measured by norms of source-/target-side context vectors. Moreover, we propose to reuse the discovered noisy data by generating pseudo labels via online knowledge distillation. Extensive experiments show that our proposed filtering method performs comparably with state-of-the-art noisy corpora filtering techniques but is more efficient and easier to operate. Noisy sample refurbishing further enhances the performance by making the most of the given data.

**Helping the Weak Makes You Strong: Simple Multi-Task Learning Improves Non-Autoregressive Translators**
*Xinyou Wang, Zaixiang Zheng and Shujian Huang* 15:30-17:00 (Hall A, Room C)
Recently, non-autoregressive (NAR) neural machine translation models have received increasing attention due to their efficient parallel decoding. However, the probabilistic framework of NAR models necessitates conditional independence assumption on target sequences, falling short of characterizing human language data. This drawback results in less informative learning signals for NAR models under conventional MLE training, thereby yielding unsatisfactory accuracy compared to their autoregressive (AR) counterparts. In this paper, we propose a simple and model-agnostic multi-task learning framework to provide more informative learning signals. During training stage, we introduce a set of sufficiently weak AR decoders that solely rely on the information provided by NAR decoder to make prediction, forcing the NAR decoder to become stronger or else it will be unable to support its weak AR partners. Experiments on WMT and IWSLT datasets show that our approach can consistently improve accuracy of multiple NAR baselines without adding any additional decoding overhead.

**Modeling Consistency Preference via Lexical Chains for Document-level Neural Machine Translation**
*Xinglin Lyu, Junhui Li, shimin tao, Hao Yang, Ying Qin and Min Zhang* 15:30-17:00 (Hall A, Room C)
In this paper we aim to relieve the issue of lexical translation inconsistency for document-level neural machine translation (NMT) by modeling consistency preference for lexical chains, which consist of repeated words in a source-side document and provide a representation of the lexical consistency structure of the document. Specifically, we first propose lexical-consistency attention to capture consistency context among words in the same lexical chains. Then for each lexical chain we define and learn a consistency-tailored latent variable, which will guide the translation of corresponding sentences to enhance lexical translation consistency. Experimental results on Chinese→English and French→English document-level translation tasks show that our approach not only significantly improves translation performance in BLEU, but also substantially alleviates the problem of the lexical translation inconsistency.

**Breaking the Representation Bottleneck of Chinese Characters: Neural Machine Translation with Stroke Sequence Modeling**
*Zhijun Wang, Xuebo Liu and Min Zhang* 15:30-17:00 (Hall A, Room C)
Existing research generally treats Chinese character as a minimum unit for representation. However, such Chinese character representation will suffer two bottlenecks: 1) Learning bottleneck, the learning cannot benefit from its rich internal features (e.g., radicals and strokes); and 2) Parameter bottleneck, each individual character has to be represented by a unique vector. In this paper, we introduce a novel representation method for Chinese characters to break the bottlenecks, namely StrokeNet, which represents a Chinese character by a Latinized stroke sequence. Specifically, StrokeNet maps each stroke to a specific Latin character, thus allowing similar Chinese characters to have similar Latin representations. With the introduction of StrokeNet to neural machine translation (NMT), many powerful but not applicable techniques to non-Latin languages (e.g., shared subword vocabulary learning and ciphertext-based data augmentation) can now be perfectly implemented. Experiments on the widely-used NIST Chinese-English, WMT17 Chinese-English and IWSLT17 Japanese-English NMT tasks show that StrokeNet can provide a significant performance boost over the strong baselines with fewer model parameters, achieving 26.5 BLEU on the WMT17 Chinese-English task which is better than any previously reported results without using monolingual data. Code and scripts are freely available at https://github.com/zjwang21/StrokeNet.

**Increasing Visual Awareness in Multimodal Neural Machine Translation from an Information Theoretic Perspective**
*Baijun Ji, Tong Zhang, Yicheng Zou, Bojie Hu and si shen* 15:30-17:00 (Hall A, Room C)
Multimodal machine translation (MMT) aims to improve translation quality by equipping the source sentence with its corresponding image. Despite the promising performance, MMT models still suffer the problem of input degradation: models focus more on textual information while visual information is generally overlooked. In this paper, we endeavor to improve MMT performance by increasing visual awareness from an information theoretic perspective. In detail, we decompose the informative visual signals into two parts: source-specific information and target-specific information. We use mutual information to quantify them and propose two methods for objective optimization to better leverage visual signals. Experiments on two datasets demonstrate that our approach can effectively enhance the visual awareness of MMT model and achieve superior results against strong baselines.

**Learning Inter-Entity-Interaction for Few-Shot Knowledge Graph Completion**
*Yuling Li, Kui Yu, Xiaoling Huang and Yuhong Zhang* 15:30-17:00 (Hall A, Room C)
Few-shot knowledge graph completion (FKGC) aims to infer unknown fact triples of a relation using its few-shot reference entity pairs. Recent FKGC studies focus on learning semantic representations of entity pairs by separately encoding the neighborhoods of head and tail entities. Such practice, however, ignores the inter-entity interaction, resulting in low-discrimination representations for entity pairs, especially when these entity pairs are associated with 1-to-N, N-to-1, and N-to-N relations. To address this issue, this paper proposes a novel FKGC

model, named Cross-Interaction Attention Network (CIAN) to investigate the inter-entity interaction between head and tail entities. Specifically, we first explore the interactions within entities by computing the attention between the task relation and each entity neighbor, and then model the interactions between head and tail entities by letting an entity to attend to the neighborhood of its paired entity. In this way, CIAN can figure out the relevant semantics between head and tail entities, thereby generating more discriminative representations for entity pairs. Extensive experiments on two public datasets show that CIAN outperforms several state-of-the-art methods. The source code is available at https://github.com/cjlyl/FKGC-CIAN.

**Is the Brain Mechanism for Hierarchical Structure Building Universal Across Languages? An fMRI Study of Chinese and English**
*Xiaohan Zhang, Shaonan Wang, Nan Lin and Chengqing Zong* 15:30-17:00 (Hall A, Room C)
Evidence from psycholinguistic studies suggests that the human brain builds a hierarchical syntactic structure during language comprehension. However, it is still unknown whether the neural basis of such structures is universal across languages. In this paper, we first analyze the differences in language structure between two diverse languages: Chinese and English. By computing the working memory requirements when applying parsing strategies to different language structures, we find that top-down parsing generates less memory load for the right-branching English and bottom-up parsing is less memory-demanding for Chinese. Then we use functional magnetic resonance imaging (fMRI) to investigate whether the brain has different syntactic adaptation strategies in processing Chinese and English. Specifically, for both Chinese and English, we extract predictors from the implementations of different parsing strategies, i.e., bottom-up and top-down. Then, these predictors are separately associated with fMRI signals. Results show that for Chinese and English, the brain utilizes bottom-up and top-down parsing strategies separately. These results reveal that the brain adopts parsing strategies with less memory processing load according to different language structures.

**Efficient Adversarial Training with Robust Early-Bird Tickets**
*zhiheng xi, Rui Zheng, Tao Gui, Qi Zhang and Xuanjing Huang* 15:30-17:00 (Hall A, Room C)
Adversarial training is one of the most powerful methods to improve the robustness of pre-trained language models (PLMs). However, this approach is typically more expensive than traditional fine-tuning because of the necessity to generate adversarial examples via gradient descent. Delving into the optimization process of adversarial training, we find that robust connectivity patterns emerge in the early training phase (typically $0.15 \sim 0.3$ epochs), far before parameters converge. Inspired by this finding, we dig out robust early-bird tickets (i.e., subnetworks) to develop an efficient adversarial training method: (1) searching for robust tickets with structured sparsity in the early stage; (2) fine-tuning robust tickets in the remaining time. To extract the robust tickets as early as possible, we design a ticket convergence metric to automatically terminate the searching process. Experiments show that the proposed efficient adversarial training method can achieve up to $7\times \sim 13\times$ training speedups while maintaining comparable or even better robustness compared to the most competitive state-of-the-art adversarial training methods.

**Label-aware Multi-level Contrastive Learning for Cross-lingual Spoken Language Understanding**
*Shining Liang, Linjun Shou, Jian Pei, Ming Gong, Wanli Zuo, Xianglin Zuo and Daxin Jiang* 15:30-17:00 (Hall A, Room C)
Despite the great success of spoken language understanding (SLU) in high-resource languages, it remains challenging in low-resource languages mainly due to the lack of labeled training data. The recent multilingual code-switching approach achieves better alignments of model representations across languages by constructing a mixed-language context in zero-shot cross-lingual SLU. However, current code-switching methods are limited to implicit alignment and disregard the inherent semantic structure in SLU, i.e., the hierarchical inclusion of utterances, slots and words. In this paper, we propose to model the utterance-slot-word structure by a multi-level contrastive learning framework at the utterance, slot and word levels to facilitate explicit alignment. Novel code-switching schemes are introduced to generate hard negative examples for our contrastive learning framework. Furthermore, we develop a label-aware joint model leveraging label semantics to enhance the implicit alignment and feed to contrastive learning. Our experimental results show that our proposed methods significantly improve the performance compared with the strong baselines on two zero-shot cross-lingual SLU benchmark datasets.

**Adaptive Token-level Cross-lingual Feature Mixing for Multilingual Neural Machine Translation**
*Junpeng Liu, Kaiyu Huang, Jiuyi Li, Huan Liu, Jinsong Su and Degen Huang* 15:30-17:00 (Hall A, Room C)
Multilingual neural machine translation aims to translate multiple language pairs in a single model and has shown great success thanks to the knowledge transfer across languages with the shared parameters. Despite promising, this share-all paradigm suffers from insufficient ability to capture language-specific features. Currently, the common practice is to insert or search language-specific networks to balance the shared and specific features. However, those two types of features are not sufficient enough to model the complex commonality and divergence across languages, such as the locally shared features among similar languages, which leads to sub-optimal transfer, especially in massively multilingual translation. In this paper, we propose a novel token-level feature mixing method that enables the model to capture different features and dynamically determine the feature sharing across languages. Based on the observation that the tokens in the multilingual model are usually shared by different languages, we we insert a feature mixing layer into each Transformer sublayer and model each token representation as a mix of different features, with a proportion indicating its feature preference. In this way, we can perform fine-grained feature sharing and achieve better multilingual transfer. Experimental results on multilingual datasets show that our method outperforms various strong baselines and can be extended to zero-shot translation. Further analyses reveal that our method can capture different linguistic features and bridge the representation gap across languages.

**Low-resource Neural Machine Translation with Cross-modal Alignment**
*Zhe Yang, Qingkai Fang and Yang Feng* 15:30-17:00 (Hall A, Room C)
How to achieve neural machine translation with limited parallel data? Existing techniques often rely on large-scale monolingual corpus, which is impractical for some low-resource languages. In this paper, we turn to connect several low-resource languages to a particular high-resource one by additional visual modality. Specifically, we propose a cross-modal contrastive learning method to learn a shared space for all languages, where both a coarse-grained sentence-level objective and a fine-grained token-level one are introduced. Experimental results and further analysis show that our method can effectively learn the cross-modal and cross-lingual alignment with a small amount of image-text pairs, and achieves significant improvements over the text-only baseline under both zero-shot and few-shot scenarios.

**Simplified Graph Learning for Inductive Short Text Classification**
*Kaixin Zheng, Yaqing Wang, Quanming Yao and Dejing Dou* 15:30-17:00 (Hall A, Room C)
Short text classification (STC) is hard as short texts lack context information and labeled data is not enough. Graph neural networks obtain the state-of-the-art on STC since they can merge various auxiliary information via the message passing framework. However, existing works conduct transductive learning, which requires retraining to accommodate new samples and takes large memory. In this paper, we present SimpleSTC which handles inductive STC problem but only leverages words. We construct word graph from an external large corpus to compensate for the lack of semantic information, and learn text graph to handle the lack of labeled data. Results show that SimpleSTC obtains state-of-the-art performance with lower memory consumption and faster inference speed.

**Interventional Training for Out-Of-Distribution Natural Language Understanding**

*Sicheng Yu, Jing Jiang, Hao Zhang, Yulei Niu, Qianru Sun and Lidong Bing* 15:30-17:00 (Hall A, Room C)
Out-of-distribution (OOD) settings are used to measure a model's performance when the distribution of the test data is different from that of the training data. NLU models are known to suffer in OOD. We study this issue from the perspective of causality, which sees confounding bias as the reason for models to learn spurious correlations. While a common solution is to perform intervention, existing methods handle only known and single confounder, but in many NLU tasks the confounders can be both unknown and multifactorial. In this paper, we propose a novel interventional training method called Bottom-up Automatic Intervention (BAI) that performs multi-granular intervention with identified multifactorial confounders. Our experiments on three NLU tasks, namely, natural language inference, fact verification and paraphrase identification, show the effectiveness of BAI for tackling OOD settings.

[DEMO] **BMCook: A Task-agnostic Compression Toolkit for Big Models**
*Zhengyan Zhang, Baitao Gong, Yingfa Chen, Xu Han, Guoyang Zeng, Weilin Zhao, Yanxu Chen, Zhiyuan Liu and Maosong Sun* 15:30-17:00 (Hall A, Room C)
Recently, pre-trained language models (PLMs) have achieved great success on various NLP tasks and have shown a trend of exponential growth in model size. To alleviate the unaffordable computational costs brought by the size growth, model compression has been widely explored. Existing efforts have achieved promising results in compressing medium-sized models for specific tasks, while task-agnostic compression for big models with over billions of parameters is rarely studied. Task-agnostic compression can provide an efficient and versatile big model for both prompting and delta tuning, leading to a more general impact than task-specific compression. Hence, we introduce a task-agnostic compression toolkit BMCook for big models. In BMCook, we implement four representative compression methods, including quantization, pruning, distillation, and MoEfication. Developers can easily combine these methods towards better efficiency. To evaluate BMCook, we apply it to compress T5-3B (a PLM with 3 billion parameters). We achieve nearly 12x efficiency improvement while maintaining over 97

## Virtual Portal 10

15:30-17:00 (Hall A, Room D)

**CapOnImage: Context-driven Dense-Captioning on Image**
*Yiqi Gao, Xinglin Hou, Yuanmeng Zhang, Tiezheng Ge, Yuning Jiang and peng wang* 15:30-17:00 (Hall A, Room D)
Existing image captioning systems are dedicated to generating narrative captions for images, which are spatially detached from the image in presentation. However, texts can also be used as decorations on the image to highlight the key points and increase the attractiveness of images. In this work, we introduce a new task called captioning on image (CapOnImage), which aims to generate dense captions at different locations of the image based on contextual information. To fully exploit the surrounding visual context to generate the most suitable caption for each location, we propose a multi-modal pre-training model with multi-level pre-training tasks that progressively learn the correspondence between texts and image locations from easy to difficult. Since the model may generate redundant captions for nearby locations, we further enhance the location embedding with neighbor locations as context. For this new task, we also introduce a large-scale benchmark called CapOnImage2M, which contains 2.1 million product images, each with an average of 4.8 spatially localized captions. Compared with other image captioning model variants, our model achieves the best results in both captioning accuracy and diversity aspects.

**Distilling Causal Effect from Miscellaneous Other-Class for Continual Named Entity Recognition**
*Junhao Zheng, Zhanxian Liang, Haibin Chen and Qianli Ma* 15:30-17:00 (Hall A, Room D)
Continual Learning for Named Entity Recognition (CL-NER) aims to learn a growing number of entity types over time from a stream of data. However, simply learning Other-Class in the same way as new entity types amplifies the catastrophic forgetting and leads to a substantial performance drop. The main cause behind this is that Other-Class samples usually contain old entity types, and the old knowledge in these Other-Class samples is not preserved properly. Thanks to the causal inference, we identify that the forgetting is caused by the missing causal effect from the old data.To this end, we propose a unified causal framework to retrieve the causality from both new entity types and Other-Class.Furthermore, we apply curriculum learning to mitigate the impact of label noise and introduce a self-adaptive weight for balancing the causal effects between new entity types and Other-Class. Experimental results on three benchmark datasets show that our method outperforms the state-of-the-art method by a large margin. Moreover, our method can be combined with the existing state-of-the-art methods to improve the performance in CL-NER.

**MedCLIP: Contrastive Learning from Unpaired Medical Images and Text**
*Zifeng Wang, Zhenbang Wu, Dinesh Agarwal and Jimeng Sun* 15:30-17:00 (Hall A, Room D)
Existing vision-text contrastive learning like CLIP aims to match the paired image and caption embeddings while pushing others apart, which improves representation transferability and supports zero-shot prediction. However, medical image-text datasets are orders of magnitude below the general images and captions from the internet. Moreover, previous methods encounter many false negatives, i.e., images and reports from separate patients probably carry the same semantics but are wrongly treated as negatives. In this paper, we decouple images and texts for multimodal contrastive learning, thus scaling the usable training data in a combinatorial magnitude with low cost. We also propose to replace the InfoNCE loss with semantic matching loss based on medical knowledge to eliminate false negatives in contrastive learning. We prove that MedCLIP is a simple yet effective framework: it outperforms state-of-the-art methods on zero-shot prediction, supervised classification, and image-text retrieval. Surprisingly, we observe that with only 20K pre-training data, MedCLIP wins over the state-of-the-art method (using 200K data). The code is available at https://github.com/RyanWangZf/MedCLIP.

**DSM: Question Generation over Knowledge Base via Modeling Diverse Subgraphs with Meta-learner**
*Shasha Guo, Jing Zhang, Yanling Wang, Qianyi Zhang, Cuiping Li and Hong Chen* 15:30-17:00 (Hall A, Room D)
Existing methods on knowledge base question generation (KBQG) learn a one-size-fits-all model by training together all subgraphs without distinguishing the diverse semantics of subgraphs. In this work, we show that making use of the past experience on semantically similar subgraphs can reduce the learning difficulty and promote the performance of KBQG models. To achieve this, we propose a novel approach to model diverse subgraphs with meta-learner (DSM). Specifically, we devise a graph contrastive learning-based retriever to identify semantically similar subgraphs, so that we can construct the semantics-aware learning tasks for the meta-learner to learn semantics-specific and semantics-agnostic knowledge on and across these tasks. Extensive experiments on two widely-adopted benchmarks for KBQG show that DSM derives new state-of-the-art performance and benefits the question answering tasks as a means of data augmentation.

**Improving Chinese Spelling Check by Character Pronunciation Prediction: The Effects of Adaptivity and Granularity**
*Jiahao Li, Quan Wang, Zhendong Mao, Junbo Guo, Yanyan Yang and Yongdong Zhang* 15:30-17:00 (Hall A, Room D)
Chinese spelling check (CSC) is a fundamental NLP task that detects and corrects spelling errors in Chinese texts. As most of these spelling errors are caused by phonetic similarity, effectively modeling the pronunciation of Chinese characters is a key factor for CSC. In this paper,

we consider introducing an auxiliary task of Chinese pronunciation prediction (CPP) to improve CSC, and, for the first time, systematically discuss the adaptivity and granularity of this auxiliary task. We propose SCOPE which builds upon a shared encoder two parallel decoders, one for the primary CSC task and the other for a fine-grained auxiliary CPP task, with a novel adaptive weighting scheme to balance the two tasks. In addition, we design a delicate iterative correction strategy for further improvements during inference. Empirical evaluation shows that SCOPE achieves new state-of-the-art on three CSC benchmarks, demonstrating the effectiveness and superiority of the auxiliary CPP task. Comprehensive ablation studies further verify the positive effects of adaptivity and granularity of the task.

**A Speaker-Aware Co-Attention Framework for Medical Dialogue Information Extraction**
*Yuan Xia, Zhenhui Shi, Jingbo Zhou, Jiayu Xu, Chao Lu, Yehui Yang, Lei Wang, Haifeng Huang, Xia Zhang and Junwei Liu* 15:30-17:00 (Hall A, Room D)
With the development of medical digitization, the extraction and structuring of Electronic Medical Records (EMRs) have become challenging but fundamental tasks. How to accurately and automatically extract structured information from medical dialogues is especially difficult because the information needs to be inferred from complex interactions between the doctor and the patient. To this end, in this paper, we propose a speaker-aware co-attention framework for medical dialogue information extraction. To better utilize the pre-trained language representation model to perceive the semantics of the utterance and the candidate item, we develop a speaker-aware dialogue encoder with multi-task learning, which considers the speaker's identity into account. To deal with complex interactions between different utterances and the correlations between utterances and candidate items, we propose a co-attention fusion network to aggregate the utterance information. We evaluate our framework on the public medical dialogue extraction datasets to demonstrate the superiority of our method, which can outperform the state-of-the-art methods by a large margin. Codes will be publicly available upon acceptance.

**Contrastive Learning enhanced Author-Style Headline Generation**
*Hui Liu, Weidong Guo, Yige Chen and Xiangyang Li* 15:30-17:00 (Hall A, Room D)
Headline generation is a task of generating an appropriate headline for a given article, which can be further used for machine-aided writing or enhancing the click-through ratio. Current works only use the article itself in the generation, but have not taken the writing style of headlines into consideration. In this paper, we propose a novel Seq2Seq model called CLH3G (Contrastive Learning enhanced Historical Headlines based Headline Generation) which can use the historical headlines of the articles that the author wrote in the past to improve the headline generation of current articles. By taking historical headlines into account, we can integrate the stylistic features of the author into our model, and generate a headline not only appropriate for the article, but also consistent with the author's style. In order to efficiently learn the stylistic features of the author, we further introduce a contrastive learning based auxiliary task for the encoder of our model. Besides, we propose two methods to use the learned stylistic features to guide both the pointer and the decoder during the generation. Experimental results show that historical headlines of the same user can improve the headline generation significantly, and both the contrastive learning module and the two style features fusion methods can further boost the performance.

**Affective Knowledge Enhanced Multiple-Graph Fusion Networks for Aspect-based Sentiment Analysis**
*Siyu Tang, Heyan Chai, Ziyi Yao, Ye Ding, Cuiyun Gao, Binxing Fang and Qing Liao* 15:30-17:00 (Hall A, Room D)
Aspect-based sentiment analysis aims to identify sentiment polarity of social media users toward different aspects. Most recent methods adopt the aspect-centric latent tree to connect aspects and their corresponding opinion words, thinking that would facilitate establishing the relationship between aspects and opinion words. However, these methods ignore the roles of syntax dependency relation labels and affective semantic information in determining the sentiment polarity, resulting in the wrong prediction. In this paper, we propose a novel multi-graph fusion network (MGFN) based on latent graph to leverage the richer syntax dependency relation label information and affective semantic information of words. Specifically, we construct a novel syntax-aware latent graph (SaLG) to fully leverage the syntax dependency relation label information to facilitate the learning of sentiment representations. Subsequently, a multi-graph fusion module is proposed to fuse semantic information of surrounding contexts of aspects adaptively. Furthermore, we design an affective refinement strategy to guide the MGFN to capture significant affective clues. Extensive experiments on three datasets demonstrate that our MGFN model outperforms all state-of-the-art methods and verify the effectiveness of our model.

**Investigating the Robustness of Natural Language Generation from Logical Forms via Counterfactual Samples**
*Chengyuan Liu, Leilei Gan, Kun Kuang and Fei Wu* 15:30-17:00 (Hall A, Room D)
The aim of Logic2Text is to generate controllable and faithful texts conditioned on tables and logical forms, which not only requires a deep understanding of the tables and logical forms, but also warrants symbolic reasoning over the tables according to the logical forms. State-of-the-art methods based on pre-trained models have achieved remarkable performance on the standard test dataset. However, we question whether these methods really learn how to perform logical reasoning, rather than just relying on the spurious correlations between the headers of the tables and operators of the logical form. To verify this hypothesis, we manually construct a set of counterfactual samples, which modify the original logical forms to generate counterfactual logical forms with rare co-occurred headers and operators and corresponding counterfactual references. SOTA methods give much worse results on these counterfactual samples compared with the results on the original test dataset, which verifies our hypothesis. To deal with this problem, we firstly analyze this bias from a causal perspective, based on which we propose two approaches to reduce the model's reliance on the shortcut. The first one incorporates the hierarchical structure of the logical forms into the model. The second one exploits automatically generated counterfactual data for training. Automatic and manual experimental results on the original test dataset and counterfactual dataset show that our method is effective to alleviate the spurious correlation. Our work points out the weakness of current methods and takes a further step toward developing Logic2Text models with real logical reasoning ability.

**R2D2: Robust Data-to-Text with Replacement Detection**
*Linyong Nan, Lorenzo Jaime Flores, Yilun Zhao, Yixin Liu, Luke Benson, Weijin Zou and Dragomir Radev* 15:30-17:00 (Hall A, Room D)
Unfaithful text generation is a common problem for text generation systems. In the case of Data-to-Text (D2T) systems, the factuality of the generated text is particularly crucial for any real-world applications. We introduce R2D2, a training framework that addresses unfaithful Data-to-Text generation by training a system both as a generator and a faithfulness discriminator with additional replacement detection and unlikelihood learning tasks. To facilitate such training, we propose two methods for sampling unfaithful sentences. We argue that the poor entity retrieval capability of D2T systems is one of the primary sources of unfaithfulness, so in addition to the existing metrics, we further propose named entity based metrics to evaluate the fidelity of D2T generations. Our experimental results show that R2D2 systems could effectively mitigate the unfaithful text generation, and they achieve new state-of-theart results on FeTaQA, LogicNLG, and ToTTo, all with significant improvements.

**Precisely the Point: Adversarial Augmentations for Faithful and Informative Text Generation**
*Wenhao Wu, Wei Li, Jiachen Liu, Xinyan Xiao, Sujian Li and Yajuan Lyu* 15:30-17:00 (Hall A, Room D)
Though model robustness has been extensively studied in language understanding, the robustness of Seq2Seq generation remains understudied. In this paper, we conduct the first quantitative analysis on the robustness of pre-trained Seq2Seq models. We find that even current SOTA pre-trained Seq2Seq model (BART) is still vulnerable, which leads to significant degeneration in faithfulness and informativeness for text generation tasks. This motivated us to further propose a novel adversarial augmentation framework, namely AdvSeq, for generally improving faithfulness and informativeness of Seq2Seq models via enhancing their robustness. AdvSeq automatically constructs two types of

adversarial augmentations during training, including implicit adversarial samples by perturbing word representations and explicit adversarial samples by word swapping, both of which effectively improve Seq2Seq robustness. Extensive experiments on three popular text generation tasks demonstrate that AdvSeq significantly improves both the faithfulness and informativeness of Seq2Seq generation under both automatic and human evaluation settings.

**FormLM: Recommending Creation Ideas for Online Forms by Modelling Semantic and Structural Information**
*Yijia Shao, Mengyu Zhou, Yifan Zhong, Tao Wu, Hongwei Han, Shi Han, Gideon Huang and Dongmei Zhang* 15:30-17:00 (Hall A, Room D)
Online forms are widely used to collect data from human and have a multi-billion market. Many software products provide online services for creating semi-structured forms where questions and descriptions are organized by predefined structures. However, the design and creation process of forms is still tedious and requires expert knowledge. To assist form designers, in this work we present FormLM to model online forms (by enhancing pre-trained language model with form structural information) and recommend form creation ideas (including question / options recommendations and block type suggestion). For model training and evaluation, we collect the first public online form dataset with 62K online forms. Experiment results show that FormLM significantly outperforms general-purpose language models on all tasks, with an improvement by 4.71 on Question Recommendation and 10.6 on Block Type Suggestion in terms of ROUGE-1 and Macro-F1, respectively.

**Towards Inter-character Relationship-driven Story Generation**
*Anvesh Rao Vijjini, Faeze Brahman and Snigdha Chaturvedi* 15:30-17:00 (Hall A, Room D)
In this paper, we introduce the task of modeling interpersonal relationships for story generation. For addressing this task, we propose Relationships as Latent Variables for Story Generation, (ReLiSt). ReLiSt generates stories sentence by sentence and has two major components - a relationship selector and a story continuer. The relationship selector specifies a latent variable to pick the relationship to exhibit in the next sentence and the story continuer generates the next sentence while expressing the selected relationship in a coherent way. Our automatic and human evaluations demonstrate that ReLiSt is able to generate stories with relationships that are more faithful to desired relationships while maintaining the content quality. The relationship assignments to sentences during inference brings interpretability to ReLiSt.

**Mask the Correct Tokens: An Embarrassingly Simple Approach for Error Correction**
*Kai Shen, Yichong Leng, Xu Tan, Siliang Tang, Yuan Zhang, Wenjie Liu and Edward Lin* 15:30-17:00 (Hall A, Room D)
Text error correction aims to correct the errors in text sequences such as those typed by humans or generated by speech recognition models. Previous error correction methods usually take the source (incorrect) sentence as encoder input and generate the target (correct) sentence through the decoder. Since the error rate of the incorrect sentence is usually low (e.g., 10
Specifically, we randomly mask out a part of the correct tokens in the source sentence and let the model learn to not only correct the original error tokens but also predict the masked tokens based on their context information. Our method enjoys several advantages: 1) it alleviates trivial copy; 2) it leverages effective training signals from correct tokens; 3) it is a plug-and-play module and can be applied to different models and tasks. Experiments on spelling error correction and speech recognition error correction on Mandarin datasets and grammar error correction on English datasets with both autoregressive and non-autoregressive generation models show that our method improves the correction accuracy consistently.

**MoSE: Modality Split and Ensemble for Multimodal Knowledge Graph Completion**
*Yu Zhao, Xiangrui Cai, Yike Wu, Haiwei Zhang, Ying Zhang, Guoqing Zhao and Ning Jiang* 15:30-17:00 (Hall A, Room D)
Multimodal knowledge graph completion (MKGC) aims to predict missing entities in MKGs. Previous works usually share relation representation across modalities. This results in mutual interference between modalities during training, since for a pair of entities, the relation from one modality probably contradicts that from another modality. Furthermore, making a unified prediction based on the shared relation representation treats the input in different modalities equally, while their importance to the MKGC task should be different. In this paper, we propose MoSE, a Modality Split representation learning and Ensemble inference framework for MKGC. Specifically, in the training phase, we learn modality-split relation embeddings for each modality instead of a single modality-shared one, which alleviates the modality interference. Based on these embeddings, in the inference phase, we first make modality-split predictions and then exploit various ensemble methods to combine the predictions with different weights, which models the modality importance dynamically. Experimental results on three KG datasets show that MoSE outperforms state-of-the-art MKGC methods. Codes are available at https://github.com/OreOZhao/MoSE4MKGC.

**ProofInfer: Generating Proof via Iterative Hierarchical Inference**
*Zichu Fei, Qi Zhang, Xin Zhou, Tao Gui and Xuanjing Huang* 15:30-17:00 (Hall A, Room D)
Proof generation focuses on deductive reasoning: given a hypothesis and a set of theories, including some supporting facts and logical rules expressed in natural language, the model generates a proof tree indicating how to deduce the hypothesis from given theories. Current models with state-of-the-art performance employ the stepwise method that adds an individual node to the proof step-by-step. However, these methods actually focus on generating several proof paths rather than a whole tree. During generation, they focus on the most relevant areas of the currently generated node while neglecting the rest of the proof tree. To address this problem, we propose ProofInfer, which generates the proof tree via iterative hierarchical inference. At each step, ProofInfer adds the entire layer to the proof, where all nodes in this layer are generated simultaneously. Since the conventional autoregressive generation architecture cannot simultaneously predict multiple nodes, ProofInfer employs text-to-text paradigm. To this end, we propose a divide-and-conquer algorithm to encode the proof tree as the plain text without losing structure information. Experimental results show that ProofInfer significantly improves performance on several widely-used datasets. In addition, ProofInfer still performs well with data-limited, achieving comparable performance to the state-of-the-art model with about 40% of the training data.

[CL] **Effective Approaches to Neural Query Language Identification**
*Xingzhang Ren, Baosong Yang, Dayiheng Liu, Haibo Zhang, Xiaoyu Lv, Liang Yao and Jun Xie* 15:30-17:00 (Hall A, Room D)
Query language identification (Q-LID) plays a crucial role in cross-lingual search engine. There exist two main challenges in Q-LID: 1) insufficient contextual information in queries for disambiguation; and 2) the lack of query-style training examples for low-resource languages. In this paper, we propose a neural Q-LID model by alleviating the above problems from both model architecture and data augmentation perspectives. Concretely, we build our model upon the advanced TRANSFORMER model. In order to enhance the discrimination of queries, a variety of external features, e.g. character, word as well as script, are fed into the model and fused by a multi-scale attention mechanism. Moreover, to remedy the low resource challenge in this task, a novel machine translation based strategy is proposed to automatically generate synthetic query style data for low-resource languages. We contribute the first Q-LID test set called QID-21, which consists of search queries in 21 languages. Experimental results reveal that our model yields better classification accuracy than strong baselines and existing LID systems on both query and traditional LID tasks.

[DEMO] **BotSIM: An End-to-End Bot Simulation Framework for Commercial Task-Oriented Dialog Systems**
*Guangsen Wang, Samson Tan, Shafiq Joty, Gang Wu, Jimmy Au and Steven C.H. Hoi* 15:30-17:00 (Hall A, Room D)
We present BotSIM, a data-efficient end-to-end Bot SIMulation framework for commercial task-oriented dialog (TOD) systems. BotSIM consists of three major components: 1) a Generator that can infer semantic-level dialog acts and entities from bot definitions and generate user queries via model-based paraphrasing; 2) an agenda-based dialog user Simulator (ABUS) to simulate conversations with the dialog agents;

3) a Remediator to analyze the simulated conversations, visualize the bot health reports and provide actionable remediation suggestions for bot troubleshooting and improvement. We demonstrate BotSIM's effectiveness in end-to-end evaluation, remediation and multi-intent dialog generation via case studies on two commercial bot platforms. BotSIM's "generation-simulation-remediation" paradigm accelerates the end-to-end bot evaluation and iteration process by: 1) reducing manual test cases creation efforts; 2) enabling a holistic gauge of the bot in terms of NLU and end-to-end performance via extensive dialog simulation; 3) improving the bot troubleshooting process with actionable suggestions. A demo of our system can be found at https://tinyurl.com/mryu74cd and a demo video at `https://youtu.be/qLPJm6\_UOKY`.

[DEMO] **TextBox 2.0: A Text Generation Library with Pre-trained Language Models**
*Tianyi Tang, Junyi Li, Zhipeng Chen, Yiwen HU, Zhuohao Yu, Wenxun Dai, Wayne Xin Zhao, Jian-Yun Nie and Ji-Rong Wen* 15:30-17:00 (Hall A, Room D)
To facilitate research on text generation, this paper presents a comprehensive and unified library, TextBox 2.0, focusing on the use of pre-trained language models (PLMs). To be comprehensive, our library covers 13 common text generation tasks and their corresponding 83 datasets and further incorporates 45 PLMs covering general, translation, Chinese, dialogue, controllable, distilled, prompting, and lightweight PLMs. We also implement 4 efficient training strategies and provide 4 generation objectives for pre-training new PLMs from scratch. To be unified, we design the interfaces to support the entire research pipeline (from data loading to training and evaluation), ensuring that each step can be fulfilled in a unified way. Despite the rich functionality, it is easy to use our library, either through the friendly Python API or command line. To validate the effectiveness of our library, we conduct extensive experiments and exemplify four types of research scenarios. The project is released at the link: https://github.com/RUCAIBox/TextBox#2.0.

## Virtual Portal 11

15:30-17:00 (Hall B)

**Curriculum Learning Meets Weakly Supervised Multimodal Correlation Learning**
*Sijie Mai, Ya Sun and Haifeng Hu* 15:30-17:00 (Hall B)
In the field of multimodal sentiment analysis (MSA), a few studies have leveraged the inherent modality correlation information stored in samples for self-supervised learning. However, they feed the training pairs in a random order without consideration of difficulty. Without human annotation, the generated training pairs of self-supervised learning often contain noise. If noisy or hard pairs are used for training at the easy stage, the model might be stuck in bad local optimum. In this paper, we inject curriculum learning into weakly supervised multimodal correlation learning. The weakly supervised correlation learning leverages the label information to generate scores for negative pairs to learn a more discriminative embedding space, where negative pairs are defined as two unimodal embeddings from different samples. To assist the correlation learning, we feed the training pairs to the model according to difficulty by the proposed curriculum learning, which consists of elaborately designed scoring and feeding functions. The scoring function computes the difficulty of pairs using pre-trained and current correlation predictors, where the pairs with large losses are defined as hard pairs. Notably, the hardest pairs are discarded in our algorithm, which are assumed as noisy pairs. Moreover, the feeding function takes the difference of correlation losses as feedback to determine the feeding actions ('stay', 'step back', or 'step forward'). The proposed method reaches state-of-the-art performance on MSA.

**Sentence Representation Learning with Generative Objective rather than Contrastive Objective**
*Bohong Wu and Hai Zhao* 15:30-17:00 (Hall B)
Though offering amazing contextualized token-level representations, current pre-trained language models take less attention on accurately acquiring sentence-level representation during their self-supervised pre-training. However, contrastive objectives which dominate the current sentence representation learning bring little linguistic interpretability and no performance guarantee on downstream semantic tasks. We instead propose a novel generative self-supervised learning objective based on phrase reconstruction. To overcome the drawbacks of previous generative methods, we carefully model intra-sentence structure by breaking down one sentence into pieces of important phrases. Empirical studies show that our generative learning achieves powerful enough performance improvement and outperforms the current state-of-the-art contrastive methods not only on the STS benchmarks, but also on downstream semantic retrieval and reranking tasks. Our code is available at https://github.com/chengzhipanpan/PaSeR.

**A Second Wave of UD Hebrew Treebanking and Cross-Domain Parsing**
*Amir Zeldes, Nick Howell, Noam Ordan and Yifat Ben Moshe* 15:30-17:00 (Hall B)
Foundational Hebrew NLP tasks such as segmentation, tagging and parsing, have relied to date on various versions of the Hebrew Treebank (HTB, Sima'an et al. 2001). However, the data in HTB, a single-source newswire corpus, is now over 30 years old, and does not cover many aspects of contemporary Hebrew on the web. This paper presents a new, freely available UD treebank of Hebrew stratified from a range of topics selected from Hebrew Wikipedia. In addition to introducing the corpus and evaluating the quality of its annotations, we deploy automatic validation tools based on grew (Guillaume, 2021), and conduct the first cross domain parsing experiments in Hebrew. We obtain new state-of-the-art (SOTA) results on UD NLP tasks, using a combination of the latest language modelling and some incremental improvements to existing transformer based approaches. We also release a new version of the UD HTB matching annotation scheme updates from our new corpus.

**MetaLogic: Logical Reasoning Explanations with Fine-Grained Structure**
*Yinya Huang, Hongming Zhang, Ruixin Hong, Xiaodan Liang, Changshui Zhang and Dong Yu* 15:30-17:00 (Hall B)
In this paper, we propose a comprehensive benchmark to investigate models' logical reasoning capabilities in complex real-life scenarios. Current explanation datasets often employ synthetic data with simple reasoning structures. Therefore, it cannot express more complex reasoning processes, such as the rebuttal to a reasoning step and the degree of certainty of the evidence. To this end, we propose a comprehensive logical reasoning explanation form. Based on the multi-hop chain of reasoning, the explanation form includes three main components: (1) The condition of rebuttal that the reasoning node can be challenged; (2) Logical formulae that uncover the internal texture of reasoning nodes; (3) Reasoning strength indicated by degrees of certainty. The fine-grained structure conforms to the real logical reasoning scenario, better fitting the human cognitive process but, simultaneously, is more challenging for the current models. We evaluate the current best models' performance on this new explanation form. The experimental results show that generating reasoning graphs remains a challenging task for current models, even with the help of giant pre-trained language models.

**Efficient Nearest Neighbor Emotion Classification with BERT-whitening**
*Wenbiao Yin and Lin Shang* 15:30-17:00 (Hall B)
Retrieval-based methods have been proven effective in many NLP tasks. Previous methods use representations from the pre-trained model for similarity search directly. However, the sentence representations from the pre-trained model like BERT perform poorly in retrieving semantically similar sentences, resulting in poor performance of the retrieval-based methods. In this paper, we propose kNN-EC, a simple and

efficient non-parametric emotion classification (EC) method using nearest neighbor retrieval. We use BERT-whitening to get better sentence semantics, ensuring that nearest neighbor retrieval works. Meanwhile, BERT-whitening can also reduce memory storage of datastore and accelerate retrieval speed, solving the efficiency problem of the previous methods. kNN-EC average improves the pre-trained model by 1.17 F1-macro on two emotion classification datasets.

**Capturing Global Structural Information in Long Document Question Answering with Compressive Graph Selector Network**
*Yuxiang Nie, Heyan Huang, Wei Wei and Xian-Ling Mao* 15:30-17:00 (Hall B)
Long document question answering is a challenging task due to its demands for complex reasoning over long text. Previous works usually take long documents as non-structured flat texts or only consider the local structure in long documents. However, these methods usually ignore the global structure of the long document, which is essential for long-range understanding. To tackle this problem, we propose Compressive Graph Selector Network (CGSN) to capture the global structure in a compressive and iterative manner. The proposed model mainly focuses on the evidence selection phase of long document question answering. Specifically, it consists of three modules: local graph network, global graph network and evidence memory network. Firstly, the local graph network builds the graph structure of the chunked segment in token, sentence, paragraph and segment levels to capture the short-term dependency of the text. Secondly, the global graph network selectively receives the information of each level from the local graph, compresses them into the global graph nodes and applies graph attention to the global graph nodes to build the long-range reasoning over the entire text in an iterative way. Thirdly, the evidence memory network is designed to alleviate the redundancy problem in the evidence selection by saving the selected result in the previous steps. Extensive experiments show that the proposed model outperforms previous methods on two datasets.

**DRLK: Dynamic Hierarchical Reasoning with Language Model and Knowledge Graph for Question Answering**
*Miao Zhang, Rufeng Dai, Ming Dong and Tingting He* 15:30-17:00 (Hall B)
In recent years, Graph Neural Network (GNN) approaches with enhanced knowledge graphs (KG) perform well in question answering (QA) tasks. One critical challenge is how to effectively utilize interactions between the QA context and KG. However, existing work only adopts the identical QA context representation to interact with multiple layers of KG, which results in a restricted interaction. In this paper, we propose DRLK (Dynamic Hierarchical Reasoning with Language Model and Knowledge Graphs), a novel model that utilizes dynamic hierarchical interactions between the QA context and KG for reasoning. DRLK extracts dynamic hierarchical features in the QA context, and performs inter-layer and intra-layer interactions on each iteration, allowing the KG representation to be grounded with the hierarchical features of the QA context. We conduct extensive experiments on four benchmark datasets in medical QA and commonsense reasoning. The experimental results demonstrate that DRLK achieves state-of-the-art performances on two benchmark datasets and performs competitively on the others.

**AEG: Argumentative Essay Generation via A Dual-Decoder Model with Content Planning**
*Jianzhu Bao, Yasheng Wang, Yitong Li, Fei Mi and Ruifeng Xu* 15:30-17:00 (Hall B)
Argument generation is an important but challenging task in computational argumentation. Existing studies have mainly focused on generating individual short arguments, while research on generating long and coherent argumentative essays is still under-explored. In this paper, we propose a new task, Argumentative Essay Generation (AEG). Given a writing prompt, the goal of AEG is to automatically generate an argumentative essay with strong persuasiveness. We construct a large-scale dataset, ArgEssay, for this new task and establish a strong model based on a dual-decoder Transformer architecture. Our proposed model contains two decoders, a planning decoder (PD) and a writing decoder (WD), where PD is used to generate a sequence for essay content planning and WD incorporates the planning information to write an essay. Further, we pre-train this model on a large news dataset to enhance the plan-and-write paradigm. Automatic and human evaluation results show that our model can generate more coherent and persuasive essays with higher diversity and less repetition compared to several baselines.

**Rethinking the Authorship Verification Experimental Setups**
*Florin Brad, Andrei Manolache, Elena Burceanu, Antonio Barbalau, Radu Tudor Ionescu and Marius Popescu* 15:30-17:00 (Hall B)
One of the main drivers of the recent advances in authorship verification is the PAN large-scale authorship dataset. Despite generating significant progress in the field, inconsistent performance differences between the closed and open test sets have been reported. To this end, we improve the experimental setup by proposing five new public splits over the PAN dataset, specifically designed to isolate and identify biases related to the text topic and to the author's writing style. We evaluate several BERT-like baselines on these splits, showing that such models are competitive with authorship verification state-of-the-art methods. Furthermore, using explainable AI, we find that these baselines are biased towards named entities. We show that models trained without the named entities obtain better results and generalize better when tested on DarkReddit, our new dataset for authorship verification.

**TIARA: Multi-grained Retrieval for Robust Question Answering over Large Knowledge Base**
*Yiheng Shu, Zhiwei Yu, Yuhan Li, Börje Karlsson, Tingting Ma, Yuzhong Qu and Chin-Yew Lin* 15:30-17:00 (Hall B)
Pre-trained language models (PLMs) have shown their effectiveness in multiple scenarios. However, KBQA remains challenging, especially regarding coverage and generalization settings. This is due to two main factors: i) understanding the semantics of both questions and relevant knowledge from the KB; ii) generating executable logical forms with both semantic and syntactic correctness. In this paper, we present a new KBQA model, TIARA, which addresses those issues by applying multi-grained retrieval to help the PLM focus on the most relevant KB context, viz., entities, exemplary logical forms, and schema items. Moreover, constrained decoding is used to control the output space and reduce generation errors. Experiments over important benchmarks demonstrate the effectiveness of our approach. TIARA outperforms previous SOTA, including those using PLMs or oracle entity annotations, by at least 4.1 and 1.1 F1 points on GrailQA and WebQuestionsSP, respectively. Specifically on GrailQA, TIARA outperforms previous models in all categories, with an improvement of 4.7 F1 points in zero-shot generalization.

**Rethinking Multi-Modal Alignment in Multi-Choice VideoQA from Feature and Sample Perspectives**
*Shaoning Xiao, Long Chen, Kaifeng Gao, Zhao Wang, Yi Yang, Zhimeng Zhang and Jun Xiao* 15:30-17:00 (Hall B)
Reasoning about causal and temporal event relations in videos is a new destination of Video Question Answering (VideoQA). The major stumbling block to achieve this purpose is the semantic gap between language and video since they are at different levels of abstraction. Existing efforts mainly focus on designing sophisticated architectures while utilizing frame- or object-level visual representations. In this paper, we reconsider the multi-modal alignment problem in VideoQA from feature and sample perspectives to achieve better performance. From the view of feature, we break down the video into trajectories and first leverage trajectory feature in VideoQA to enhance the alignment between two modalities. Moreover, we adopt a heterogeneous graph architecture and design a hierarchical framework to align both trajectory-level and frame-level visual feature with language feature. In addition, we found that VideoQA models are largely dependent on language priors and always neglect visual-language interactions. Thus, two effective yet portable training augmentation strategies are designed to strengthen the cross-modal correspondence ability of our model from the view of sample. Extensive results show that our method outperforms all the state-of-the-art models on the challenging NExT-QA benchmark.

**Hierarchical Multi-Label Classification of Scientific Documents**
*Mobashir Sadat and Cornelia Caragea* 15:30-17:00 (Hall B)
Automatic topic classification has been studied extensively to assist managing and indexing scientific documents in a digital collection. With

the large number of topics being available in recent years, it has become necessary to arrange them in a hierarchy. Therefore, the automatic classification systems need to be able to classify the documents hierarchically. In addition, each paper is often assigned to more than one relevant topic. For example, a paper can be assigned to several topics in a hierarchy tree. In this paper, we introduce a new dataset for hierarchical multi-label text classification (HMLTC) of scientific papers called SciHTC, which contains 186,160 papers and 1,234 categories from the ACM CCS tree. We establish strong baselines for HMLTC and propose a multi-task learning approach for topic classification with keyword labeling as an auxiliary task. Our best model achieves a Macro-F1 score of 34.57% which shows that this dataset provides significant research opportunities on hierarchical scientific topic classification. We make our dataset and code for all experiments publicly available.

**Symptom Identification for Interpretable Detection of Multiple Mental Disorders on Social Media**
*Zhiling Zhang, Siyuan Chen, Mengyue Wu and Kenny Zhu* 15:30-17:00 (Hall B)
Mental disease detection (MDD) from social media has suffered from poor generalizability and interpretability, due to lack of symptom modeling. This paper introduces PsySym, the first annotated symptom identification corpus of multiple psychiatric disorders, to facilitate further research progress. PsySym is annotated according to a knowledge graph of the 38 symptom classes related to 7 mental diseases complied from established clinical manuals and scales, and a novel annotation framework for diversity and quality. Experiments show that symptom-assisted MDD enabled by PsySym can outperform strong pure-text baselines. We also exhibit the convincing MDD explanations provided by symptom predictions with case studies, and point to their further potential applications.

**CISLR: Corpus for Indian Sign Language Recognition**
*Abhinav Joshi, Ashwani Bhat, Pradeep S, Priya Gole, Shashwat Gupta, Shreyansh Agarwal and Ashutosh Modi* 15:30-17:00 (Hall B)
Indian Sign Language, though used by a diverse community, still lacks well-annotated resources for developing systems that would enable sign language processing. In recent years researchers have actively worked for sign languages like American Sign Languages, however, Indian Sign language is still far from data-driven tasks like machine translation. To address this gap, in this paper, we introduce a new dataset CISLR (Corpus for Indian Sign Language Recognition) for word-level recognition in Indian Sign Language using videos. The corpus has a large vocabulary of around 4700 words covering different topics and domains. Further, we propose a baseline model for word recognition from sign language videos. To handle the low resource problem in the Indian Sign Language, the proposed model consists of a prototype-based one-shot learner that leverages resource rich American Sign Language to learn generalized features for improving predictions in Indian Sign Language. Our experiments show that gesture features learned in another sign language can help perform one-shot predictions in CISLR.

**Improving Tokenisation by Alternative Treatment of Spaces**
*Edward Gow-Smith, Harish Tayyar Madabushi, Carolina Scarton and Aline Villavicencio* 15:30-17:00 (Hall B)

[CL] **The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization**
*Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez and Montserrat Batet* 15:30-17:00 (Hall B)
We present a novel benchmark and associated evaluation metrics for assessing the performance of text anonymization methods. Text anonymization, defined as the task of editing a text document to prevent the disclosure of personal information, currently suffers from a shortage of privacy-oriented annotated text resources, making it difficult to properly evaluate the level of privacy protection offered by various anonymization methods. This paper presents TAB (Text Anonymization Benchmark), a new, open-source annotated corpus developed to address this shortage. The corpus comprises 1,268 English-language court cases from the European Court of Human Rights (ECHR) enriched with comprehensive annotations about the personal information appearing in each document, including their semantic category, identifier type, confidential attributes, and co-reference relations. Compared to previous work, the TAB corpus is designed to go beyond traditional de-identification (which is limited to the detection of predefined semantic categories), and explicitly marks which text spans ought to be masked in order to conceal the identity of the person to be protected. Along with presenting the corpus and its annotation layers, we also propose a set of evaluation metrics that are specifically tailored towards measuring the performance of text anonymization, both in terms of privacy protection and utility preservation. We illustrate the use of the benchmark and the proposed metrics by assessing the empirical performance of several baseline text anonymization models. The full corpus along with its privacy-oriented annotation guidelines, evaluation scripts and baseline models are available on: https://github.com/NorskRegnesentral/text-anonymization-benchmark

[DEMO] **AnEMIC: A Framework for Benchmarking ICD Coding Models**
*Juyong Kim, Abheesht Sharma, Suhas Shanbhogue, Jeremy Weiss and Pradeep Ravikumar* 15:30-17:00 (Hall B)
Diagnostic coding, or ICD coding, is the task of assigning diagnosis codes defined by the ICD (International Classification of Diseases) standard to patient visits based on clinical notes. The current process of manual ICD coding is time-consuming and often error-prone, which suggests the need for automatic ICD coding. However, despite the long history of automatic ICD coding, there have been no standardized frameworks for benchmarking ICD coding models. We open-source an easy-to-use tool named *AnEMIC*, which provides a streamlined pipeline for preprocessing, training, and evaluating for automatic ICD coding. We correct errors in preprocessing by existing works, and provide key models and weights trained on the correctly preprocessed datasets. We also provide an interactive demo performing real-time inference from custom inputs, and visualizations drawn from explainable AI to analyze the models. We hope the framework helps move the research of ICD coding forward and helps professionals explore the potential of ICD coding. The framework and the associated code are available here.

[DEMO] **MedConQA: Medical Conversational Question Answering System based on Knowledge Graphs**
*Fei Xia, Bin Li, Yixuan Weng, Shizhu He, Kang Liu, Bin Sun, Shutao Li and Jun Zhao* 15:30-17:00 (Hall B)
The medical conversational system can relieve doctors' burden and improve healthcare efficiency, especially during the COVID-19 pandemic. However, the existing medical dialogue systems have the problems of weak scalability, insufficient knowledge, and poor controllability. Thus, we propose a medical conversational question-answering (CQA) system based on the knowledge graph, namely MedConQA, which is designed as a pipeline framework to maintain high flexibility. Our system utilizes automated medical procedures, including medical triage, consultation, image-text drug recommendation, and record. Each module has been open-sourced as a tool, which can be used alone or in combination, with robust scalability. Besides, to conduct knowledge-grounded dialogues with users, we first construct a Chinese Medical Knowledge Graph (CMKG) and collect a large-scale Chinese Medical CQA (CMCQA) dataset, and we design a series of methods for reasoning more intellectually. Finally, we use several state-of-the-art (SOTA) techniques to keep the final generated response more controllable, which is further assured by hospital and professional evaluations. We have open-sourced related code, datasets, web pages, and tools, hoping to advance future research.

## Virtual Portal 12

15:30-17:00 (Collaboratorium)

**LVP-M3: Language-aware Visual Prompt for Multilingual Multimodal Machine Translation**
*Hongcheng Guo, Jiaheng Liu, Haoyang Huang, Jian Yang, Zhoujun Li, Dongdong Zhang and Zheng Cui* 15:30-17:00 (Collaboratorium)
Multimodal Machine Translation (MMT) focuses on enhancing text-only translation with visual features, which has attracted considerable attention from both natural language processing and computer vision communities. Recent advances still struggle to train a separate model for each language pair, which is costly and unaffordable when the number of languages increases in the real world. In other words, the multilingual multimodal machine translation (Multilingual MMT) task has not been investigated, which aims to handle the aforementioned issues by providing a shared semantic space for multiple languages. Besides, the image modality has no language boundaries, which is superior to bridging the semantic gap between languages. To this end, we first propose the Multilingual MMT task by establishing two new Multilingual MMT benchmark datasets covering seven languages. Then, an effective baseline LVP-M3 using visual prompts is proposed to support translations between different languages, which includes three stages (token encoding, language-aware visual prompt generation, and language translation). Extensive experimental results on our constructed benchmark datasets demonstrate the effectiveness of LVP-M3 method for Multilingual MMT.

**UniGeo: Unifying Geometry Logical Reasoning via Reformulating Mathematical Expression**
*Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen and Xiaodan Liang* 15:30-17:00 (Collaboratorium)
Geometry problem solving is a well-recognized testbed for evaluating the high-level multi-modal reasoning capability of deep models. In most existing works, two main geometry problems: calculation and proving, are usually treated as two specific tasks, hindering a deep model to unify its reasoning capability on multiple math tasks. However, in essence, these two tasks have similar problem representations and overlapped math knowledge which can improve the understanding and reasoning ability of a deep model on both two tasks. Therefore, we construct a large-scale Unified Geometry problem benchmark, UniGeo, which contains 4,998 calculation problems and 9,543 proving problems. Each proving problem is annotated with a multi-step proof with reasons and mathematical expressions. The proof can be easily reformulated as a proving sequence that shares the same formats with the annotated program sequence for calculation problems. Naturally, we also present a unified multi-task Geometric Transformer framework, Geoformer, to tackle calculation and proving problems simultaneously in the form of sequence generation, which finally shows the reasoning ability can be improved on both two tasks by unifying formulation. Furthermore, we propose a Mathematical Expression Pretraining (MEP) method that aims to predict the mathematical expressions in the problem solution, thus improving the Geoformer model. Experiments on the UniGeo demonstrate that our proposed Geoformer obtains state-of-the-art performance by outperforming task-specific model NGS with over 5.6

**PASTA: Table-Operations Aware Fact Verification via Sentence-Table Cloze Pre-training**
*Zihui Gu, Ju Fan, Nan Tang, Preslav Nakov, Xiaoman Zhao and Xiaoyong Du* 15:30-17:00 (Collaboratorium)
Fact verification has attracted a lot of attention recently, e.g., in journalism, marketing, and policymaking, as misinformation and dis- information can sway one's opinion and affect one's actions. While fact-checking is a hard task in general, in many cases, false statements can be easily debunked based on analytics over tables with reliable information. Hence, table- based fact verification has recently emerged as an important and growing research area. Yet, progress has been limited due to the lack of datasets that can be used to pre-train language models (LMs) to be aware of common table operations, such as aggregating a column or comparing tuples. To bridge this gap, this paper introduces PASTA for table-based fact verification via pre-training with synthesized sentence–table cloze questions. In particular, we design six types of common sentence–table cloze tasks, including Filter, Aggregation, Superlative, Comparative, Ordinal, and Unique, based on which we synthesize a large corpus consisting of 1.2 million sentence–table pairs from WikiTables. PASTA uses a recent pre-trained LM, DeBERTaV3, and further pre- trains it on our corpus. Our experimental results show that PASTA achieves new state-of-the-art (SOTA) performance on two table-based fact verification datasets TabFact and SEM-TAB- FACTS. In particular, on the complex set of TabFact, which contains multiple operations, PASTA largely outperforms previous SOTA by 4.7% (85.6% vs. 80.9%), and the gap between PASTA and human performance on the small test set is narrowed to just 1.5% (90.6% vs. 92.1%).

**An Anchor-based Relative Position Embedding Method for Cross-Modal Tasks**
*Ya Wang, Xingwu Sun, Lian Fengzong, ZhanHui Kang and Chengzhong Xu Xu* 15:30-17:00 (Collaboratorium)
Position Embedding (PE) is essential for transformer to capture the sequence ordering of input tokens. Despite its general effectiveness verified in Natural Language Processing (NLP) and Computer Vision (CV), its application in cross-modal tasks remains unexplored and suffers from two challenges: 1) the input text tokens and image patches are not aligned, 2) the encoding space of each modality is different, making it unavailable for feature comparison. In this paper, we propose a unified position embedding method for these problems, called AnChorbasEd Relative Position Embedding (ACE-RPE), in which we first introduce an anchor locating mechanism to bridge the semantic gap and locate anchors from different modalities. Then we conduct the distance calculation of each text token and image patch by computing their shortest paths from the located anchors. Last, we embed the anchor-based distance to guide the computation of cross-attention. In this way, it calculates cross-modal relative position embedding for cross-modal transformer. Benefiting from ACE-RPE, our method obtains new SOTA results on a wide range of benchmarks, such as Image-Text Retrieval on MS-COCO and Flickr30K, Visual Entailment on SNLI-VE, Visual Reasoning on NLVR2 and Weakly-supervised Visual Grounding on RefCOCO+.

**RACE: Retrieval-augmented Commit Message Generation**
*Ensheng Shi, Yanlin Wang, Wei Tao, Lun Du, Hongyu Zhang, Shi Han, Dongmei Zhang and Hongbin Sun* 15:30-17:00 (Collaboratorium)
Commit messages are important for software development and maintenance. Many neural network-based approaches have been proposed and shown promising results on automatic commit message generation. However, the generated commit messages could be repetitive or redundant. In this paper, we propose RACE, a new retrieval-augmented neural commit message generation method, which treats the retrieved similar commit as an exemplar and leverages it to generate an accurate commit message. As the retrieved commit message may not always accurately describe the content/intent of the current code diff, we also propose an exemplar guider, which learns the semantic similarity between the retrieved and current code diff and then guides the generation of commit message based on the similarity. We conduct extensive experiments on a large public dataset with five programming languages. Experimental results show that RACE can outperform all baselines. Furthermore, RACE can boost the performance of existing Seq2Seq models in commit message generation.

**Leveraging Locality in Abstractive Text Summarization**
*Yixin Liu, Ansong Ni, Linyong Nan, Budhaditya Deb, Chenguang Zhu, Ahmed Hassan Awadallah and Dragomir Radev* 15:30-17:00 (Collaboratorium)
Neural attention models have achieved significant improvements on many natural language processing tasks. However, the quadratic memory complexity of the self-attention module with respect to the input length hinders their applications in long text summarization. Instead of designing more efficient attention modules, we approach this problem by investigating if models with a restricted context can have competitive performance compared with the memory-efficient attention models that maintain a global context by treating the input as a single sequence. Our model is applied to individual pages, which contain parts of inputs grouped by the principle of locality, during both the encoding and decoding stages. We empirically investigated three kinds of locality in text summarization at different levels of granularity, ranging from sentences to documents. Our experimental results show that our model has a better performance compared with strong baseline models with efficient attention modules, and our analysis provides further insights into our locality-aware modeling strategy.

**A Span-based Multimodal Variational Autoencoder for Semi-supervised Multimodal Named Entity Recognition**
*Baohang Zhou, Ying Zhang, Kehui Song, Wenya Guo, Guoqing Zhao, hongbin wang and Xiaojie Yuan* 15:30-17:00 (Collaboratorium)
Multimodal named entity recognition (MNER) on social media is a challenging task which aims to extract named entities in free text and incorporate images to classify them into user-defined types. However, the annotation for named entities on social media demands a mount of human efforts. The existing semi-supervised named entity recognition methods focus on the text modal and are utilized to reduce labeling costs in traditional NER. However, the previous methods are not efficient for semi-supervised MNER. Because the MNER task is defined to combine the text information with image one and needs to consider the mismatch between the posted text and image. To fuse the text and image features for MNER effectively under semi-supervised setting, we propose a novel span-based multimodal variational autoencoder (SM-VAE) model for semi-supervised MNER. The proposed method exploits modal-specific VAEs to model text and image latent features, and utilizes product-of-experts to acquire multimodal features. In our approach, the implicit relations between labels and multimodal features are modeled by multimodal VAE. Thus, the useful information of unlabeled data can be exploited in our method under semi-supervised setting. Experimental results on two benchmark datasets demonstrate that our approach not only outperforms baselines under supervised setting, but also improves MNER performance with less labeled data than existing semi-supervised methods.

**Towards Unifying Reference Expression Generation and Comprehension**
*Duo Zheng, Tao Kong, Ya Jing, Jiaan Wang and Xiaojie WANG* 15:30-17:00 (Collaboratorium)
Reference Expression Generation (REG) and Comprehension (REC) are two highly correlated tasks. Modeling REG and REC simultaneously for utilizing the relation between them is a promising way to improve both. However, the problem of distinct inputs, as well as building connections between them in a single model, brings challenges to the design and training of the joint model. To address the problems, we propose a unified model for REG and REC, named UniRef. It unifies these two tasks with the carefully-designed Image-Region-Text Fusion layer (IRTF), which fuses the image, region and text via the image cross-attention and region cross-attention. Additionally, IRTF could generate pseudo input regions for the REC task to enable a uniform way for sharing the identical representation space across the REC and REG. We further propose Vision-conditioned Masked Language Modeling (VMLM) and Text-Conditioned Region Prediction (TRP) to pre-train UniRef model on multi-granular corpora. The VMLM and TRP are directly related to REG and REC, respectively, but could help each other. We conduct extensive experiments on three benchmark datasets, RefCOCO, RefCOCO+ and RefCOCOg. Experimental results show that our model outperforms previous state-of-the-art methods on both REG and REC.

**Assist Non-native Viewers: Multimodal Cross-Lingual Summarization for How2 Videos**
*Nayu Liu, Kaiwen Wei, Xian Sun, Hongfeng Yu, Fanglong Yao, li jin, Guo Zhi and Guangluan Xu* 15:30-17:00 (Collaboratorium)
Multimodal summarization for videos aims to generate summaries from multi-source information (videos, audio transcripts), which has achieved promising progress. However, existing works are restricted to monolingual video scenarios, ignoring the demands of non-native video viewers to understand the cross-language videos in practical applications. It stimulates us to propose a new task, named Multimodal Cross-Lingual Summarization for videos (MCLS), which aims to generate cross-lingual summaries from multimodal inputs of videos. First, to make it applicable to MCLS scenarios, we conduct a Video-guided Dual Fusion network (VDF) that integrates multimodal and cross-lingual information via diverse fusion strategies at both encoder and decoder. Moreover, to alleviate the problem of high annotation costs and limited resources in MCLS, we propose a triple-stage training framework to assist MCLS by transferring the knowledge from monolingual multimodal summarization data, which includes: 1) multimodal summarization on sufficient prevalent language videos with a VDF model; 2) knowledge distillation (KD) guided adjustment on bilingual transcripts; 3) multimodal summarization for cross-lingual videos with a KD induced VDF model. Experiment results on the reorganized How2 dataset show that the VDF model alone outperforms previous methods for multimodal summarization, and the performance further improves by a large margin via the proposed triple-stage training framework.

**Speaker Overlap-aware Neural Diarization for Multi-party Meeting Analysis**
*Zhihao Du, ShiLiang Zhang, Siqi Zheng and Zhi-Jie Yan* 15:30-17:00 (Collaboratorium)
Recently, hybrid systems of clustering and neural diarization models have been successfully applied in multi-party meeting analysis. However, current models always treat overlapped speaker diarization as a multi-label classification problem, where speaker dependency and overlaps are not well considered. To overcome the disadvantages, we reformulate overlapped speaker diarization task as a single-label prediction problem via the proposed power set encoding (PSE). Through this formulation, speaker dependency and overlaps can be explicitly modeled. To fully leverage this formulation, we further propose the speaker overlap-aware neural diarization (SOND) model, which consists of a context-independent (CI) scorer to model global speaker discriminability, a context-dependent scorer (CD) to model local discriminability, and a speaker combining network (SCN) to combine and reassign speaker activities. Experimental results show that using the proposed formulation can outperform the state-of-the-art methods based on target speaker voice activity detection, and the performance can be further improved with SOND, resulting in a 6.30% relative diarization error reduction.

**Extending Phrase Grounding with Pronouns in Visual Dialogues**
*Panzhong Lu, Xin Zhang, Meishan Zhang and Min Zhang* 15:30-17:00 (Collaboratorium)
Conventional phrase grounding aims to localize noun phrases mentioned in a given caption to their corresponding image regions, which has achieved great success recently. Apparently, sole noun phrase grounding is not enough for cross-modal visual language understanding. Here we extend the task by considering pronouns as well. First, we construct a dataset of phrase grounding with both noun phrases and pronouns to image regions. Based on the dataset, we test the performance of phrase grounding by using a state-of-the-art literature model of this line. Then, we enhance the baseline grounding model with coreference information which should help our task potentially, modeling the coreference structures with graph convolutional networks. Experiments on our dataset, interestingly, show that pronouns are easier to ground than noun phrases, where the possible reason might be that these pronouns are much less ambiguous. Additionally, our final model with coreference information can significantly boost the grounding performance of both noun phrases and pronouns.

**ClidSum: A Benchmark Dataset for Cross-Lingual Dialogue Summarization**
*Jiaan Wang, Fandong Meng, Ziyao Lu, Duo Zheng, Zhixu Li, Jianfeng Qu and Jie Zhou* 15:30-17:00 (Collaboratorium)
We present ClidSum, a benchmark dataset towards building cross-lingual summarization systems on dialogue documents. It consists of 67k+ dialogue documents and 112k+ annotated summaries in different target languages. Based on the proposed ClidSum, we introduce two benchmark settings for supervised and semi-supervised scenarios, respectively. We then build various baseline systems in different paradigms (pipeline and end-to-end) and conduct extensive experiments on ClidSum to provide deeper analyses. Furthermore, we propose mDialBART which extends mBART via further pre-training, where the multiple objectives help the pre-trained model capture the structural characteristics as well as key content in dialogues and the transformation from source to the target language. Experimental results show the superiority of mDialBART, as an end-to-end model, outperforms strong pipeline models on ClidSum. Finally, we discuss specific challenges that current approaches faced with this task and give multiple promising directions for future research. We have released the dataset and code at https://github.com/krystalan/ClidSum.

**Distilled Dual-Encoder Model for Vision-Language Understanding**
*Zekun Wang, Wenhui Wang, Haichao Zhu, ming liu, Bing Qin and Furu Wei* 15:30-17:00 (Collaboratorium)
On vision-language understanding (VLU) tasks, fusion-encoder vision-language models achieve superior results but sacrifice efficiency be-

cause of the simultaneous encoding of images and text. On the contrary, the dual encoder model that separately encodes images and text has the advantage in efficiency, while failing on VLU tasks due to the lack of deep cross-modal interactions. To get the best of both worlds, we propose DiDE, a framework that distills the knowledge of the fusion-encoder teacher model into the dual-encoder student model. Since the cross-modal interaction is the key to the superior performance of teacher model but is absent in the student model, we encourage the student not only to mimic the predictions of teacher, but also to calculate the cross-modal attention distributions and align with the teacher. Experimental results demonstrate that DiDE is competitive with the fusion-encoder teacher model in performance (only a 1% drop) while enjoying 4 times faster inference. Further analyses reveal that the proposed cross-modal attention distillation is crucial to the success of our framework.

**Do Children Texts Hold The Key To Commonsense Knowledge?**
*Julien Romero and Simon Razniewski* 15:30-17:00 (Collaboratorium)
Compiling comprehensive repositories of commonsense knowledge is a long-standing problem in AI. Many concerns revolve around the issue of reporting bias, i.e., that frequency in text sources is not a good proxy for relevance or truth. This paper explores whether children's texts hold the key to commonsense knowledge compilation, based on the hypothesis that such content makes fewer assumptions on the reader's knowledge, and therefore spells out commonsense more explicitly. An analysis with several corpora shows that children's texts indeed contain much more, and more typical commonsense assertions. Moreover, experiments show that this advantage can be leveraged in popular language-model-based commonsense knowledge extraction settings, where task-unspecific fine-tuning on small amounts of children texts (childBERT) already yields significant improvements. This provides a refreshing perspective different from the common trend of deriving progress from ever larger models and corpora.

**PEVL: Position-enhanced Pre-training and Prompt Tuning for Vision-language Models**
*Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua and Maosong Sun* 15:30-17:00 (Collaboratorium)
Vision-language pre-training (VLP) has shown impressive performance on a wide range of cross-modal tasks, where VLP models without reliance on object detectors are becoming the mainstream due to their superior computation efficiency and competitive performance. However, the removal of object detectors also deprives the capability of VLP models in explicit object modeling, which is essential to various position-sensitive vision-language (VL) tasks, such as referring expression comprehension and visual commonsense reasoning. To address the challenge, we introduce PEVL that enhances the pre-training and prompt tuning of VLP models with explicit object position modeling. Specifically, PEVL reformulates discretized object positions and language in a unified language modeling framework, which facilitates explicit VL alignment during pre-training, and also enables flexible prompt tuning for various downstream tasks. We show that PEVL enables state-of-the-art performance of detector-free VLP models on position-sensitive tasks such as referring expression comprehension and phrase grounding, and also improves the performance on position-insensitive tasks with grounded inputs. We make the data and code for this paper publicly available at https://github.com/thunlp/PEVL.

[TACL] **A Survey on Cross-Lingual Summarization**
*Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu and Jie Zhou* 15:30-17:00 (Collaboratorium)
Cross-lingual summarization is the task of generating a summary in one language (e.g., English) for the given document(s) in a different language (e.g., Chinese). Under the globalization background, this task has attracted increasing attention of the computational linguistics community. Nevertheless, there still remains a lack of comprehensive review for this task. Therefore, we present the first systematic critical review on the datasets, approaches, and challenges in this field. Specifically, we carefully organize existing datasets and approaches according to different construction methods and solution paradigms, respectively. For each type of datasets or approaches, we thoroughly introduce and summarize previous efforts and further compare them with each other to provide deeper analyses. In the end, we also discuss promising directions and offer our thoughts to facilitate future research. This survey is for both beginners and experts in cross-lingual summarization, and we hope it will serve as a starting point as well as a source of new ideas for researchers and engineers interested in this area.

## Poster Sessions 11 & 12

15:30-17:00 (Atrium)

---

**Understanding ME? Multimodal Evaluation for Fine-grained Visual Commonsense**
*Zhecan Wang, Haoxuan You, Yicheng He, Wenhao Li, Kai-Wei Chang and Shih-Fu Chang* 15:30-17:00 (Atrium)
Visual commonsense understanding requires Vision Language (VL) models to not only understand image and text but also cross-reference in-between to fully integrate and achieve comprehension of the visual scene described. Recently, various approaches have been developed and have achieved high performance on visual commonsense benchmarks. However, it is unclear whether the models really understand the visual scene and underlying commonsense knowledge due to limited evaluation data resources. To provide an in-depth analysis, we present a Multimodal Evaluation (ME) pipeline to automatically generate question-answer pairs to test models' understanding of the visual scene, text, and related knowledge. We then take a step further to show that training with the ME data boosts the model's performance in standard VCR evaluation. Lastly, our in-depth analysis and comparison reveal interesting findings: (1) semantically low-level information can assist the learning of high-level information but not the opposite; (2) visual information is generally under utilization compared with text.

**Multi-Label Intent Detection via Contrastive Task Specialization of Sentence Encoders**
*Ivan Vulić, Iñigo Casanueva, Georgios Spithourakis, Avishek Mondal, Tsung-Hsien Wen and Paweł Budzianowski* 15:30-17:00 (Atrium)
Deploying task-oriented dialog ToD systems for new domains and tasks requires natural language understanding models that are 1) resource-efficient and work under low-data regimes; 2) adaptable, efficient, and quick-to-train; 3) expressive and can handle complex ToD scenarios with multiple user intents in a single utterance. Motivated by these requirements, we introduce a novel framework for multi-label intent detection (mID): MultI-ConvFiT (Multi-Label Intent Detection via Contrastive Conversational Fine-Tuning). While previous work on efficient single-label intent detection learns a classifier on top of a fixed sentence encoder (SE), we propose to 1) transform general-purpose SEs into task-specialized SEs via contrastive fine-tuning on annotated multi-label data, 2) where task specialization knowledge can be stored into lightweight adapter modules without updating the original parameters of the input SE, and then 3) we build improved mID classifiers stacked on top of fixed specialized SEs. Our main results indicate that MultI-ConvFiT yields effective mID models, with large gains over non-specialized SEs reported across a spectrum of different mID datasets, both in low-data and high-data regimes.

[TACL] **Evaluating Attribution in Dialogue Systems: The BEGIN Benchmark**
*Nouha Dziri, Hannah Rashkin, Tal Linzen and David Reitter* 15:30-17:00 (Atrium)
Knowledge-grounded dialogue systems powered by large language models often generate responses that, while fluent, are not attributable to a relevant source of information. Progress towards models that do not exhibit this issue requires evaluation metrics that can quantify its prevalence. To this end, we introduce the Benchmark for Evaluation of Grounded INteraction (BEGIN), comprised of 12k dialogue turns generated

by neural dialogue systems trained on three knowledge-grounded dialogue corpora. We collect human annotations assessing the extent to which the models' responses can be attributed to the given background information. We then use BEGIN to analyze eight evaluation metrics. We find that these metrics rely on spurious correlations, do not reliably distinguish attributable abstractive responses from unattributable ones, and perform substantially worse when the knowledge source is longer. Our findings underscore the need for more sophisticated and robust evaluation metrics for knowledge-grounded dialogue. We make BEGIN publicly available at https://github.com/google/BEGIN-dataset .

**Is a Question Decomposition Unit All We Need?**
*Pruthvi Patel, Swaroop Mishra, Mihir Parmar and Chitta Baral* 15:30-17:00 (Atrium)
Large Language Models (LMs) have achieved state-of-the-art performance on many Natural Language Processing (NLP) benchmarks. With the growing number of new benchmarks, we build bigger and more complex LMs. However, building new LMs may not be an ideal option owing to the cost, time and environmental impact associated with it. We explore an alternative route: can we modify data by expressing it in terms of the model's strengths, so that a question becomes easier for models to answer? We investigate if humans can decompose a hard question into a set of simpler questions that are relatively easier for models to solve. We analyze a range of datasets involving various forms of reasoning and find that it is indeed possible to significantly improve model performance (24% for GPT3 and 29% for RoBERTa-SQuAD along with a symbolic calculator) via decomposition. Our approach provides a viable option to involve people in NLP research in a meaningful way. Our findings indicate that Human-in-the-loop Question Decomposition (HQD) can potentially provide an alternate path to building large LMs.

**EdgeFormer: A Parameter-Efficient Transformer for On-Device Seq2seq Generation**
*Tao Ge, Si-Qing Chen and Furu Wei* 15:30-17:00 (Atrium)
We introduce EdgeFormer – a parameter-efficient Transformer for on-device seq2seq generation under the strict computation and memory constraints. Compared with the previous parameter-efficient Transformers, EdgeFormer applies two novel principles for cost-effective parameterization, allowing it to perform better given the same parameter budget; moreover, EdgeFormer is further enhanced by layer adaptation innovation that is proposed for improving the network with shared layers.

Extensive experiments show EdgeFormer can effectively outperform previous parameter-efficient Transformer baselines and achieve competitive results under both the computation and memory constraints. Given the promising results, we release EdgeLM – the pretrained version of EdgeFormer, which is the first publicly available pretrained on-device seq2seq model that can be easily fine-tuned for seq2seq tasks with strong results, facilitating on-device seq2seq generation in practice.

**MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text**
*Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga and William Cohen* 15:30-17:00 (Atrium)
While language Models store a massive amount of world knowledge implicitly in their parameters, even very large models often fail to encode information about rare entities and events, while incurring huge computational costs. Recently, retrieval-augmented models, such as REALM, RAG, and RETRO, have incorporated world knowledge into language generation by leveraging an external non-parametric index and have demonstrated impressive performance with constrained model sizes. However, these methods are restricted to retrieving only textual knowledge, neglecting the ubiquitous amount of knowledge in other modalities like images – much of which contains information not covered by any text. To address this limitation, we propose the first Multimodal Retrieval-Augmented Transformer (MuRAG), which accesses an external non-parametric multimodal memory to augment language generation. MuRAG is pre-trained with a mixture of large-scale image-text and text-only corpora using a joint contrastive and generative loss. We perform experiments on two different datasets that require retrieving and reasoning over both images and text to answer a given query: WebQA, and MultimodalQA. Our results show that MuRAG achieves state-of-the-art accuracy, outperforming existing models by 10-20

**Debiasing Pretrained Text Encoders by Paying Attention to Paying Attention**
*Yacine Gaci, Boualem Benatallah, Fabio Casati and Khalid Benabdeslem* 15:30-17:00 (Atrium)
Natural Language Processing (NLP) models are found to exhibit discriminatory stereotypes across many social constructs, e.g. gender and race. In comparison to the progress made in reducing bias from static word embeddings, fairness in sentence-level text encoders received little consideration despite their wider applicability in contemporary NLP tasks. In this paper, we propose a debiasing method for pre-trained text encoders that both reduces social stereotypes, and inflicts next to no semantic damage. Unlike previous studies that directly manipulate the embeddings, we suggest to dive deeper into the operation of these encoders, and pay more attention to the way they pay attention to different social groups. We find that stereotypes are also encoded in the attention layer. Then, we work on model debiasing by redistributing the attention scores of a text encoder such that it forgets any preference to historically advantaged groups, and attends to all social classes with the same intensity. Our experiments confirm that reducing bias from attention effectively mitigates it from the model's text representations.

**SetGNER: General Named Entity Recognition as Entity Set Generation**
*Yuxin He and Buzhou Tang* 15:30-17:00 (Atrium)
Recently, joint recognition of flat, nested and discontinuous entities has received increasing attention. Motivated by the observation that the target output of NER is essentially a set of sequences, we propose a novel entity set generation framework for general NER scenes in this paper. Different from sequence-to-sequence NER methods, our method does not force the entities to be generated in a predefined order and can get rid of the problem of error propagation and inefficient decoding. Distinguished from the set-prediction NER framework, our method treats each entity as a sequence and is capable of recognizing discontinuous mentions. Given an input sentence, the model first encodes the sentence in word-level and detects potential entity mentions based on the encoder's output, then reconstructs entity mentions from the detected entity heads in parallel. To let the encoder of our model capture better right-to-left semantic structure, we also propose an auxiliary Inverse Generation Training task. Extensive experiments show that our model (w/o. Inverse Generation Training) outperforms state-of-the-art generative NER models by a large margin on two discontinuous NER datasets, two nested NER datasets and one flat NER dataset. Besides, the auxiliary Inverse Generation Training task is found to further improve the model's performance on the five datasets.

**Does Your Model Classify Entities Reasonably? Diagnosing and Mitigating Spurious Correlations in Entity Typing**
*Nan Xu, Fei Wang, Bangzheng Li, Mingtao Dong and Muhao Chen* 15:30-17:00 (Atrium)
Entity typing aims at predicting one or more words that describe the type(s) of a specific mention in a sentence. Due to shortcuts from surface patterns to annotated entity labels and biased training, existing entity typing models are subject to the problem of spurious correlations. To comprehensively investigate the faithfulness and reliability of entity typing methods, we first systematically define distinct kinds of model biases that are reflected mainly from spurious correlations. Particularly, we identify six types of existing model biases, including mention-context bias, lexical overlapping bias, named entity bias, pronoun bias, dependency bias, and overgeneralization bias. To mitigate model biases, we then introduce a counterfactual data augmentation method. By augmenting the original training set with their debiased counterparts, models are forced to fully comprehend sentences and discover the fundamental cues for entity typing, rather than relying on spurious correlations for shortcuts. Experimental results on the UFET dataset show our counterfactual data augmentation approach helps improve generalization of different entity typing models with consistently better performance on both the original and debiased test sets.

**Decoding a Neural Retriever's Latent Space for Query Suggestion**
*Leonard Adolphs, Michelle Chen Huebscher, Christian Buck, Sertan Girgin, Olivier Bachem, Massimiliano Ciaramita and Thomas Hofmann*

15:30-17:00 (Atrium)
Neural retrieval models have superseded classic bag-of-words methods such as BM25 as the retrieval framework of choice. However, neural systems lack the interpretability of bag-of-words models; it is not trivial to connect a query change to a change in the latent space that ultimately determines the retrieval results. To shed light on this embedding space, we learn a "query decoder" that, given a latent representation of a neural search engine, generates the corresponding query. We show that it is possible to decode a meaningful query from its latent representation and, when moving in the right direction in latent space, to decode a query that retrieves the relevant paragraph. In particular, the query decoder can be useful to understand "what should have been asked" to retrieve a particular paragraph from the collection. We employ the query decoder to generate a large synthetic dataset of query reformulations for MSMarco, leading to improved retrieval performance. On this data, we train a pseudo-relevance feedback (PRF) T5 model for the application of query suggestion that outperforms both query reformulation and PRF information retrieval baselines.

**GPS: Genetic Prompt Search for Efficient Few-Shot Learning**
*Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, wang yanggang, Haiyu Li and Zhilin Yang* 15:30-17:00 (Atrium)
Prompt-based techniques have demostrated great potential for improving the few-shot generalization of pretrained language models. However, their performance heavily relies on the manual design of prompts and thus requiring a lot of human efforts. In this paper, we introduce Genetic Prompt Search (GPS) to improve few-shot learning with prompts, which utilizes a genetic algorithm to automatically search for the best prompt. GPS is gradient-free and requires no update of model parameters but only a small validation set. Experiments on diverse datasets proved the effectiveness of GPS, which outperforms manual prompts by a large margin of 2.6 points. Our method is also better than other parameter-efficient tuning methods such as prompt tuning.

**Continual Training of Language Models for Few-Shot Learning**
*Zixuan Ke, Haowei Lin, Yijia Shao, Hu Xu, Lei Shu and Bing Liu* 15:30-17:00 (Atrium)
Recent work on applying large language models (LMs) achieves impressive performance in many NLP applications. Adapting or posttraining an LM using an unlabeled domain corpus can produce even better performance for end-tasks in the domain. This paper proposes the problem of continually extending an LM by incrementally post-train the LM with a sequence of unlabeled domain corpora to expand its knowledge without forgetting its previous skills. The goal is to improve the few-shot end-task learning in these domains. The resulting system is called CPT (Continual PostTraining), which to our knowledge, is the first continual post-training system. Experimental results verify its effectiveness.

**WeTS: A Benchmark for Translation Suggestion**
*Zhen Yang, Fandong Meng, Yingxue Zhang, Ernan Li and Jie Zhou* 15:30-17:00 (Atrium)
Translation suggestion (TS), which provides alternatives for specific words or phrases given the entire documents generated by machine translation (MT), has been proven to play a significant role in post-editing (PE). There are two main pitfalls for existing researches in this line. First, most conventional works only focus on the overall performance of PE but ignore the exact performance of TS, which makes the progress of PE sluggish and less explainable; Second, as no publicly available golden dataset exists to support in-depth research for TS, almost all of the previous works conduct experiments on their in-house datasets or the noisy datasets built automatically, which makes their experiments hard to be reproduced and compared. To break these limitations mentioned above and spur the research in TS, we create a benchmark dataset, called *WeTS*, which is a golden corpus annotated by expert translators on four translation directions. Apart from the golden corpus, we also propose several methods to generate synthetic corpora which can be used to improve the performance substantially through pre-training. As for the model, we propose the segment-aware self-attention based Transformer for TS. Experimental results show that our approach achieves the best results on all four directions, including English-to-German, German-to-English, Chinese-to-English, and English-to-Chinese.[4]

**T-Modules: Translation Modules for Zero-Shot Cross-Modal Machine Translation**
*Paul-Ambroise Duquenne, Hongyu Gong, Benoît Sagot and Holger Schwenk* 15:30-17:00 (Atrium)
We present a new approach to perform zero-shot cross-modal transfer between speech and text for translation tasks. Multilingual speech and text are encoded in a joint fixed-size representation space. Then, we compare different approaches to decode these multimodal and multilingual fixed-size representations, enabling zero-shot translation between languages and modalities. All our models are trained without the need of cross-modal labeled translation data. Despite a fixed-size representation, we achieve very competitive results on several text and speech translation tasks. In particular, we significantly improve the state-of-the-art for zero-shot speech translation on Must-C. Incorporating a speech decoder in our framework, we introduce the first results for zero-shot direct speech-to-speech and text-to-speech translation.

**Joint Completion and Alignment of Multilingual Knowledge Graphs**
*Soumen Chakrabarti, Harkanwar Singh, Shubham Lohiya, Prachi Jain and Mausam -* 15:30-17:00 (Atrium)
Knowledge Graph Completion (KGC) predicts missing facts in an incomplete Knowledge Graph (KG). Multilingual KGs associate entities and relations with surface forms written in different languages. An entity or relation may be associated with distinct IDs in different KGs, necessitating entity alignment (EA) and relation alignment (RA). Many effective algorithms have been proposed for completion and alignment as separate tasks. Here we show that these tasks are synergistic and best solved together. Our multitask approach starts with a state-of-the-art KG embedding scheme, but adds a novel relation representation based on sets of embeddings of (subject, object) entity pairs. This representation leads to a new relation alignment loss term based on a maximal bipartite matching between two sets of embedding vectors. This loss is combined with traditional KGC loss and optionally, losses based on text embeddings of entity (and relation) names. In experiments over KGs in seven languages, we find that our system achieves large improvements in KGC compared to a strong completion model that combines known facts in all languages. It also outperforms strong EA and RA baselines, underscoring the value of joint alignment and completion.

**BERT in Plutarch's Shadows**
*Ivan Yamshchikov, Alexey Tikhonov, Yorgos Pantis, Charlotte Schubert and Jürgen Jost* 15:30-17:00 (Atrium)
The extensive surviving corpus of the ancient scholar Plutarch of Chaeronea (ca. 45-120 CE) also contains several texts which, according to current scholarly opinion, did not originate with him and are therefore attributed to an anonymous author Pseudo-Plutarch. These include, in particular, the work Placita Philosophorum (Quotations and Opinions of the Ancient Philosophers), which is extremely important for the history of ancient philosophy. Little is known about the identity of that anonymous author and its relation to other authors from the same period. This paper presents a BERT language model for Ancient Greek. The model discovers previously unknown statistical properties relevant to these literary, philosophical, and historical problems and can shed new light on this authorship question. In particular, the Placita Philosophorum, together with one of the other Pseudo-Plutarch texts, shows similarities with the texts written by authors from an Alexandrian context (2nd/3rd century CE).

**Composing Ci with Reinforced Non-autoregressive Text Generation**
*Yan Song* 15:30-17:00 (Atrium)

[4]For reviewers, codes and corpus can be found in the attached files, and we will make them publicly available after the double-blind phase.

Composing Ci (also widely known as Song Ci), a special type of classical Chinese poetry, requires to follow particular format once their tune patterns are given. To automatically generate a well-formed Ci, text generation systems should strictly take into account pre-defined rigid formats (e.g., length and rhyme). Yet, most existing approaches regard Ci generation as a conventional sequence-to-sequence task and use autoregressive models, while it is challenging for such models to properly handle the constraints (according to tune patterns) of Ci during the generation process. Moreover, consider that with the format prepared, Ci generation can be operated by an efficient synchronous process, where autoregressive models are limited in doing so since they follow the character-by-character generation protocol. Therefore, in this paper, we propose to compose Ci through a non-autoregressive approach, which not only ensure that the generation process accommodates tune patterns by controlling the rhythm and essential meaning of each sentence, but also allow the model to perform synchronous generation. In addition, we further improve our approach by applying reinforcement learning to the generation process with the rigid constraints of Ci as well as the diversity in content serving as rewards, so as to further maintain the format and content requirement. Experiments on a collected Ci dataset confirm that our proposed approach outperforms strong baselines and previous studies in terms of both automatic evaluation metrics and human judgements.

**ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering**
*Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah and William Yang Wang* 15:30-17:00 (Atrium)
With the recent advance in large pre-trained language models, researchers have achieved record performances in NLP tasks that mostly focus on language pattern matching. The community is experiencing the shift of the challenge from how to model language to the imitation of complex reasoning abilities like human beings. In this work, we investigate the application domain of finance that involves real-world, complex numerical reasoning. We propose a new large-scale dataset, ConvFinQA, aiming to study the chain of numerical reasoning in conversational question answering. Our dataset poses great challenge in modeling long-range, complex numerical reasoning paths in real-world conversations. We conduct comprehensive experiments and analyses with both the neural symbolic methods and the prompting-based methods, to provide insights into the reasoning mechanisms of these two divisions. We believe our new dataset should serve as a valuable resource to push forward the exploration of real-world, complex reasoning tasks as the next research focus. Our dataset and code is publicly available at https://github.com/czyssrs/ConvFinQA.

**Uni-Parser: Unified Semantic Parser for Question Answering on Knowledge Base and Database**
*Ye Liu, Semih Yavuz, Rui Meng, Dragomir Radev, caiming xiong and Yingbo Zhou* 15:30-17:00 (Atrium)
Parsing natural language questions into executable logical forms is a useful and interpretable way to perform question answering on structured data such as knowledge bases (KB) or databases (DB). However, existing approaches on semantic parsing cannot adapt to both modalities, as they suffer from the exponential growth of the logical form candidates and can hardly generalize to unseen data. In this work, we propose Uni-Parser, a unified semantic parser for question answering (QA) on both KB and DB. We define the primitive (relation and entity in KB, and table name, column name and cell value in DB) as the essential element in our framework. The number of primitives grows only at a linear rate to the number of retrieved relations in KB and DB, preventing us from exponential logic form candidates. We leverage the generator to predict final logical forms by altering and composing top-ranked primitives with different operations (e.g. select, where, count). With sufficiently pruned search space by a contrastive primitive ranker, the generator is empowered to capture the composition of primitives enhancing its generalization ability. We achieve competitive results on multiple KB and DB QA benchmarks with more efficiency, especially in the compositional and zero-shot settings.

**GENIE: Toward Reproducible and Standardized Human Evaluation for Text Generation**
*Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith and Daniel Weld*15:30-17:00 (Atrium)
While often assumed a gold standard, effective human evaluation of text generation remains an important, open area for research. We revisit this problem with a focus on producing consistent evaluations that are reproducible—over time and across different populations. We study this goal in different stages of the human evaluation pipeline. In particular, we consider design choices for the annotation interface used to elicit human judgments and their impact on reproducibility. Furthermore, we develop an automated mechanism for maintaining annotator quality via a probabilistic model that detects and excludes noisy annotators. Putting these lessons together, we introduce GENIE: a system for running standardized human evaluations across different generation tasks. We instantiate GENIE with datasets representing four core challenges in text generation: machine translation, summarization, commonsense reasoning, and machine comprehension. For each task, GENIE offers a leaderboard that automatically crowdsources annotations for submissions, evaluating them along axes such as correctness, conciseness, and fluency. We have made the GENIE leaderboards publicly available, and have already ranked 50 submissions from 10 different research groups. We hope GENIE encourages further progress toward effective, standardized evaluations for text generation.

**Open World Classification with Adaptive Negative Samples**
*Ke Bai, Guoyin Wang, Jiwei Li, Sunghyun Park, Sungjin Lee, Puyang Xu, Ricardo Henao and Lawrence Carin* 15:30-17:00 (Atrium)
Open world classification is a task in natural language processing with key practical relevance and impact. Since the open or unknown category data only manifests in the inference phase, finding a model with a suitable decision boundary accommodating for the identification of known classes and discrimination of the open category is challenging. The performance of existing models is limited by the lack of effective open category data during the training stage or the lack of a good mechanism to learn appropriate decision boundaries. We propose an approach based on Adaptive Negative Samples (ANS) designed to generate effective synthetic open category samples in the training stage and without requiring any prior knowledge or external datasets. Empirically, we find a significant advantage in using auxiliary one-versus-rest binary classifiers, which effectively utilize the generated negative samples and avoid the complex threshold-seeking stage in previous works. Extensive experiments on three benchmark datasets show that ANS achieves significant improvements over state-of-the-art methods.

**FLUTE: Figurative Language Understanding through Textual Explanations**
*Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh and Smaranda Muresan* 15:30-17:00 (Atrium)
Figurative language understanding has been recently framed as a recognizing textual entailment (RTE) task (a.k.a. natural language inference (NLI)). However, similar to classical RTE/NLI datasets they suffer from spurious correlations and annotation artifacts. To tackle this problem, work on NLI has built explanation-based datasets such as eSNLI, allowing us to probe whether language models are right for the right reasons. Yet no such data exists for figurative language, making it harder to assess genuine understanding of such expressions. To address this issue, we release FLUTE, a dataset of 9,000 figurative NLI instances with explanations, spanning four categories: Sarcasm, Simile, Metaphor, and Idioms. We collect the data through a Human-AI collaboration framework based on GPT-3, crowd workers, and expert annotators. We show how utilizing GPT-3 in conjunction with human annotators (novices and experts) can aid in scaling up the creation of datasets even for such complex linguistic phenomena as figurative language. The baseline performance of the T5 model fine-tuned on FLUTE shows that our dataset can bring us a step closer to developing models that understand figurative language through textual explanations.

**Breakpoint Transformers for Modeling and Tracking Intermediate Beliefs**
*Kyle Richardson, Ronen Tamari, Oren Sultan, Dafna Shahaf, Reut Tsarfaty and Ashish Sabharwal* 15:30-17:00 (Atrium)
Can we teach models designed for language understanding tasks to track and improve their beliefs through intermediate points in text? Besides making their inner workings more transparent, this would also help make models more reliable and consistent. To this end, we propose

a representation learning framework called breakpoint modeling that allows for efficient and robust learning of this type. Given any text encoder and data marked with intermediate states (breakpoints) along with corresponding textual queries viewed as true/false propositions (i.e., the candidate intermediate beliefs of a model), our approach trains models in an efficient and end-to-end fashion to build intermediate representations that facilitate direct querying and training of beliefs at arbitrary points in text, alongside solving other end-tasks. We evaluate breakpoint modeling on a diverse set of NLU tasks including relation reasoning on Cluttr and narrative understanding on bAbI. Using novel proposition prediction tasks alongside these end-tasks, we show the benefit of our T5-based breakpoint transformer over strong conventional representation learning approaches in terms of processing efficiency, belief accuracy, and belief consistency, all with minimal to no degradation on the end-task. To show the feasibility of incorporating our belief tracker into more complex reasoning pipelines, we also obtain state-of-the-art performance on the three-tiered reasoning challenge for the recent TRIP benchmark (23-32% absolute improvement on Tasks 2-3).

**Sentiment-Aware Word and Sentence Level Pre-training for Sentiment Analysis**
*Shuai Fan, Chen Lin, Haonan Li, Zhenghao Lin, Jinsong Su, Hang Zhang, Yeyun Gong, JIan Guo and Nan Duan* 15:30-17:00 (Atrium)
Most existing pre-trained language representation models (PLMs) are sub-optimal in sentiment analysis tasks, as they capture the sentiment information from word-level while under-considering sentence-level information. In this paper, we propose SentiWSP, a novel Sentiment-aware pre-trained language model with combined Word-level and Sentence-level Pre-training tasks. The word level pre-training task detects replaced sentiment words, via a generator-discriminator framework, to enhance the PLM's knowledge about sentiment words. The sentence level pre-training task further strengthens the discriminator via a contrastive learning framework, with similar sentences as negative samples, to encode sentiments in a sentence. Extensive experimental results show that SentiWSP achieves new state-of-the-art performance on various sentence-level and aspect-level sentiment classification benchmarks. We have made our code and model publicly available at https://github.com/XMUDM/SentiWSP.

**Improving Aspect Sentiment Quad Prediction via Template-Order Data Augmentation**
*Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai and Shiwan Zhao* 15:30-17:00 (Atrium)
Recently, aspect sentiment quad prediction (ASQP) has become a popular task in the field of aspect-level sentiment analysis. Previous work utilizes a predefined template to paraphrase the original sentence into a structure target sequence, which can be easily decoded as quadruplets of the form (aspect category, aspect term, opinion term, sentiment polarity). The template involves the four elements in a fixed order. However, we observe that this solution contradicts with the order-free property of the ASQP task, since there is no need to fix the template order as long as the quadruplet is extracted correctly. Inspired by the observation, we study the effects of template orders and find that some orders help the generative model achieve better performance. It is hypothesized that different orders provide various views of the quadruplet. Therefore, we propose a simple but effective method to identify the most proper orders, and further combine multiple proper templates as data augmentation to improve the ASQP task. Specifically, we use the pre-trained language model to select the orders with minimal entropy. By fine-tuning the pre-trained language model with these template orders, our approach improves the performance of quad prediction, and outperforms state-of-the-art methods significantly in low-resource settings.

**CTRLsum: Towards Generic Controllable Text Summarization**
*Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani and Caiming Xiong* 15:30-17:00 (Atrium)
Current summarization systems yield generic summaries that are disconnected from users' preferences and expectations. To address this limitation, we present CTRLsum, a generic framework to control generated summaries through a set of keywords. During training keywords are extracted automatically without requiring additional human annotations. At test time CTRLsum features a control function to map control signal to keywords; through engineering the control function, the same trained model is able to be applied to control summaries on various dimensions, while neither affecting the model training process nor the pretrained models. We additionally explore the combination of keywords and text prompts for more control tasks. Experiments demonstrate the effectiveness of CTRLsum on three domains of summarization datasets and five control tasks: (1) entity-centric and (2) length-controllable summarization, (3) contribution summarization on scientific papers, (4) invention purpose summarization on patent filings, and (5) question-guided summarization on news articles. Moreover, when used in a standard, unconstrained summarization setting, CTRLsum is comparable or better than strong pretrained systems.

# Main Conference: Sunday, December 11, 2022

## Session 11 - 09:00-10:30

### Machine Translation

09:00-10:30 (Hall A, Room A)

**The Importance of Being Parameters: An Intra-Distillation Method for Serious Gains**
*Haoran Xu, Philipp Koehn and Kenton Murray* 09:00-09:15 (Hall A, Room A)
Recent model pruning methods have demonstrated the ability to remove redundant parameters without sacrificing model performance. Common methods remove redundant parameters according to the parameter sensitivity, a gradient-based measure reflecting the contribution of the parameters. In this paper, however, we argue that redundant parameters can be trained to make beneficial contributions. We first highlight the large sensitivity (contribution) gap among high-sensitivity and low-sensitivity parameters and show that the model generalization performance can be significantly improved after balancing the contribution of all parameters. Our goal is to balance the sensitivity of all parameters and encourage all of them to contribute equally. We propose a general task-agnostic method, namely intra-distillation, appended to the regular training loss to balance parameter sensitivity. Moreover, we also design a novel adaptive learning method to control the strength of intra-distillation loss for faster convergence. Our experiments show the strong effectiveness of our methods on machine translation, natural language understanding, and zero-shot cross-lingual transfer across up to 48 languages, e.g., a gain of 3.54 BLEU on average across 8 language pairs from the IWSLT'14 dataset.

**Non-Parametric Domain Adaptation for End-to-End Speech Translation**
*Yichao Du, Weizhi Wang, Zhirui Zhang, Boxing Chen, Tong Xu, Jun Xie and Enhong Chen* 09:15-09:30 (Hall A, Room A)
The end-to-end speech translation (E2E-ST) has received increasing attention due to the potential of its less error propagation, lower latency and fewer parameters. However, the effectiveness of neural-based approaches to this task is severely limited by the available training corpus, especially for domain adaptation where in-domain triplet data is scarce or nonexistent. In this paper, we propose a novel non-parametric method that leverages in-domain text translation corpus to achieve domain adaptation for E2E-ST systems. To this end, we first incorporate an additional encoder into the pre-trained E2E-ST model to realize text translation modeling, based on which the decoder's output representations for text and speech translation tasks are unified by reducing the correspondent representation mismatch in available triplet training data. During domain adaptation, a k-nearest-neighbor (kNN) classifier is introduced to produce the final translation distribution using the external datastore built by the domain-specific text translation corpus, while the universal output representation is adopted to perform a similarity search. Experiments on the Europarl-ST benchmark demonstrate that when in-domain text translation data is involved only, our proposed approach significantly improves baseline by 12.82 BLEU on average in all translation directions, even outperforming the strong in-domain fine-tuning strategy.

**Information-Transport-based Policy for Simultaneous Translation**
*Shaolei Zhang and Yang Feng* 09:30-09:45 (Hall A, Room A)
Simultaneous translation (ST) outputs translation while receiving the source inputs, and hence requires a policy to determine whether to translate a target token or wait for the next source token. The major challenge of ST is that each target token can only be translated based on the current received source tokens, where the received source information will directly affect the translation quality. So naturally, how much source information is received for the translation of the current target token is supposed to be the pivotal evidence for the ST policy to decide between translating and waiting. In this paper, we treat the translation as information transport from source to target and accordingly propose an Information-Transport-based Simultaneous Translation (ITST). ITST quantifies the transported information weight from each source token to the current target token, and then decides whether to translate the target token according to its accumulated received information. Experiments on both text-to-text ST and speech-to-text ST (a.k.a., streaming speech translation) tasks show that ITST outperforms strong baselines and achieves state-of-the-art performance.

**Multilingual Machine Translation with Hyper-Adapters**
*Christos Baziotis, Mikel Artetxe, James Cross and Shruti Bhosale* 09:45-10:00 (Hall A, Room A)
Multilingual machine translation suffers from negative interference across languages. A common solution is to relax parameter sharing with language-specific modules like adapters. However, adapters of related languages are unable to transfer information, and their total number of parameters becomes prohibitively expensive as the number of languages grows. In this work, we overcome these drawbacks using hyper-adapters – hyper-networks that generate adapters from language and layer embeddings. While past work had poor results when scaling hyper-networks, we propose a rescaling fix that significantly improves convergence and enables training larger hyper-networks. We find that hyper-adapters are more parameter efficient than regular adapters, reaching the same performance with up to 12 times less parameters. When using the same number of parameters and FLOPS, our approach consistently outperforms regular adapters. Also, hyper-adapters converge faster than alternative approaches and scale better than regular dense networks. Our analysis shows that hyper-adapters learn to encode language relatedness, enabling positive transfer across languages.

**Continual Learning of Neural Machine Translation within Low Forgetting Risk Regions**
*Shuhao Gu, Bojie Hu and Yang Feng* 10:00-10:15 (Hall A, Room A)
This paper considers continual learning of large-scale pretrained neural machine translation model without accessing the previous training data or introducing model separation. We argue that the widely used regularization-based methods, which perform multi-objective learning with an auxiliary loss, suffer from the misestimate problem and cannot always achieve a good balance between the previous and new tasks. To solve the problem, we propose a two-stage training method based on the local features of the real loss. We first search low forgetting risk regions, where the model can retain the performance on the previous task as the parameters are updated, to avoid the catastrophic forgetting problem. Then we can continually train the model within this region only with the new training data to fit the new task. Specifically, we propose two methods to search the low forgetting risk regions, which are based on the curvature of loss and the impacts of the parameters on the model output, respectively. We conduct experiments on domain adaptation and more challenging language adaptation tasks, and the experimental results show that our method can achieve significant improvements compared with several strong baselines.

**Distill The Image to Nowhere: Inversion Knowledge Distillation for Multimodal Machine Translation**
*RU Peng, Yawen Zeng and Jake Zhao* 10:15-10:30 (Hall A, Room A)
Past works on multimodal machine translation (MMT) elevate bilingual setup by incorporating additional aligned vision information. How-

ever, an image-must requirement of the multimodal dataset largely hinders MMT's development — namely that it demands an aligned form of [image, source text, target text]. This limitation is generally troublesome during the inference phase especially when the aligned image is not provided as in the normal NMT setup. Thus, in this work, we introduce IKD-MMT, a novel MMT framework to support the image-free inference phase via an inversion knowledge distillation scheme. In particular, a multimodal feature generator is executed with a knowledge distillation module, which directly generates the multimodal feature from (only) source texts as the input. While there have been a few prior works entertaining the possibility to support image-free inference for machine translation, their performances have yet to rival the image-must translation. In our experiments, we identify our method as the first image-free approach to comprehensively rival or even surpass (almost) all image-must frameworks, and achieved the state-of-the-art result on the often-used Multi30k benchmark. Our code and data are available at: https://github.com/pengr/IKD-mmt/tree/master..

## Commonsense Reasoning

09:00-10:30 (Hall A, Room B)

**Using Commonsense Knowledge to Answer Why-Questions**
*Yash Kumar Lal, Niket Tandon, Tanvi Aggarwal, Horace Liu, Nathanael Chambers, Raymond Mooney and Niranjan Balasubramanian*09:00-09:15 (Hall A, Room B)
Answering questions in narratives about why events happened often requires commonsense knowledge external to the text. What aspects of this knowledge are available in large language models? What aspects can be made accessible via external commonsense resources? We study these questions in the context of answering questions in the TellMeWhy dataset using COMET as a source of relevant commonsense relations. We analyze the effects of model size (T5 and GPT3) along with methods of injecting knowledge (COMET) into these models. Results show that the largest models, as expected, yield substantial improvements over base models. Injecting external knowledge helps models of various sizes, but the amount of improvement decreases with larger model size. We also find that the format in which knowledge is provided is critical, and that smaller models benefit more from larger amounts of knowledge. Finally, we develop an ontology of knowledge types and analyze the relative coverage of the models across these categories.

**Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations**
*Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras and Yejin Choi* 09:15-09:30 (Hall A, Room B)
Pre-trained language models (LMs) struggle with consistent reasoning; recently, prompting LMs to generate explanations that self-guide the inference has emerged as a promising direction to amend this. However, these approaches are fundamentally bounded by the correctness of explanations, which themselves are often noisy and inconsistent. In this work, we develop Maieutic Prompting, which aims to infer a correct answer to a question even from the unreliable generations of LM. Maieutic Prompting induces a tree of explanations abductively (e.g. X is true, because ...) and recursively, then frames the inference as a satisfiability problem over these explanations and their logical relations. We test Maieutic Prompting for true/false QA on three challenging benchmarks that require complex commonsense reasoning. Maieutic Prompting achieves up to 20% better accuracy than state-of-the-art prompting methods, and as a fully unsupervised approach, performs competitively with supervised models. We also show that Maieutic Prompting improves robustness in inference while providing interpretable rationales.

**Language Models of Code are Few-Shot Commonsense Learners**
*Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang and Graham Neubig* 09:30-09:45 (Hall A, Room B)
We address the general task of structured commonsense reasoning: given a natural language input, the goal is to generate a graph such as an event or a reasoning-graph. To employ large language models (LMs) for this task, existing approaches 'serialize' the output graph as a flat list of nodes and edges. Although feasible, these serialized graphs strongly deviate from the natural language corpora that LMs were pre-trained on, hindering LMs from generating them correctly. In this paper, we show that when we instead frame structured commonsense reasoning tasks as code generation tasks, pre-trained LMs of code are better structured commonsense reasoners than LMs of natural language, even when the downstream task does not involve source code at all. We demonstrate our approach across three diverse structured commonsense reasoning tasks. In all these natural language tasks, we show that using our approach, a code generation LM (codex) outperforms natural-LMs that are fine-tuned on the target task (T5) and other strong LMs such as GPT-3 in the few-shot setting.

**Enhancing Self-Consistency and Performance of Pre-Trained Language Models through Natural Language Inference**
*Eric Mitchell, Joseph Noh, Siyan Li, Will Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn and Christopher Manning* 09:45-10:00 (Hall A, Room B)
While large pre-trained language models are powerful, their predictions often lack logical consistency across test inputs. For example, a state-of-the-art Macaw question-answering (QA) model answers Yes to Is a sparrow a bird? and Does a bird have feet? but answers No to Does a sparrow have feet?. To address this failure mode, we propose a framework, Consistency Correction through Relation Detection, or ConCoRD, for boosting the consistency and accuracy of pre-trained NLP models using pre-trained natural language inference (NLI) models without fine-tuning or re-training. Given a batch of test inputs, ConCoRD samples several candidate outputs for each input and instantiates a factor graph that accounts for both the model's belief about the likelihood of each answer choice in isolation and the NLI model's beliefs about pair-wise answer choice compatibility. We show that a weighted MaxSAT solver can efficiently compute high-quality answer choices under this factor graph, improving over the raw model's predictions. Our experiments demonstrate that ConCoRD consistently boosts accuracy and consistency of off-the-shelf closed-book QA and VQA models using off-the-shelf NLI models, notably increasing accuracy of LXMERT on ConVQA by 5% absolute. See the project website (https://ericmitchell.ai/emnlp-2022-concord/) for code and data.

**EvEntS ReaLM: Event Reasoning of Entity States via Language Models**
*Evangelia Spiliopoulou, Artidoro Pagnoni, Yonatan Bisk and Eduard Hovy* 10:00-10:15 (Hall A, Room B)
This paper investigates models of event implications. Specifically, how well models predict entity state-changes, by targeting their understanding of physical attributes. Nominally, Large Language models (LLM) have been exposed to procedural knowledge about how objects interact, yet our benchmarking shows they fail to reason about the world. Conversely, we also demonstrate that existing approaches often misrepresent the surprising abilities of LLMs via improper task encodings and that proper model prompting can dramatically improve performance of reported baseline results across multiple tasks. In particular, our results indicate that our prompting technique is especially useful for unseen attributes (out-of-domain) or when only limited data is available.

**GeoMLAMA: Geo-Diverse Commonsense Probing on Multilingual Pre-Trained Language Models**
*Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li and Kai-Wei Chang* 10:15-10:30 (Hall A, Room B)
Recent work has shown that Pre-trained Language Models (PLMs) store the relational knowledge learned from data and utilize it for per-

forming downstream tasks. However, commonsense knowledge across different regions may vary. For instance, the color of bridal dress is white in American weddings whereas it is red in Chinese weddings. In this paper, we introduce a benchmark dataset, Geo-diverse Commonsense Multilingual Language Models Analysis (GeoMLAMA), for probing the diversity of the relational knowledge in multilingual PLMs. GeoMLAMA contains 3125 prompts in English, Chinese, Hindi, Persian, and Swahili, with a wide coverage of concepts shared by people from American, Chinese, Indian, Iranian and Kenyan cultures. We benchmark 11 standard multilingual PLMs on GeoMLAMA. Interestingly, we find that 1) larger multilingual PLMs variants do not necessarily store geo-diverse concepts better than its smaller variant; 2) multilingual PLMs are not intrinsically biased towards knowledge from the Western countries (the United States); 3) the native language of a country may not be the best language to probe its knowledge and 4) a language may better probe knowledge about a non-native country than its native country.

## Interpretability, Interactivity, and Analysis of Models for NLP 2

09:00-10:30 (Hall A, Room C)

**A Multilingual Perspective Towards the Evaluation of Attribution Methods in Natural Language Inference**
*Kerem Zaman and Yonatan Belinkov* 09:00-09:12 (Hall A, Room C)
Most evaluations of attribution methods focus on the English language. In this work, we present a multilingual approach for evaluating attribution methods for the Natural Language Inference (NLI) task in terms of faithfulness and plausibility. First, we introduce a novel cross-lingual strategy to measure faithfulness based on word alignments, which eliminates the drawbacks of erasure-based evaluations. We then perform a comprehensive evaluation of attribution methods, considering different output mechanisms and aggregation methods. Finally, we augment the XNLI dataset with highlight-based explanations, providing a multilingual NLI dataset with highlights, to support future exNLP studies. Our results show that attribution methods performing best for plausibility and faithfulness are different.

**Robustness of Demonstration-based Learning Under Limited Data Scenario**
*Hongxin Zhang, Yanzhe Zhang, Ruiyi Zhang and Diyi Yang* 09:12-09:24 (Hall A, Room C)
Demonstration-based learning has shown great potential in stimulating pretrained language models' ability under limited data scenario. Simply augmenting the input with some demonstrations can significantly improve performance on few-shot NER. However, why such demonstrations are beneficial for the learning process remains unclear since there is no explicit alignment between the demonstrations and the predictions. In this paper, we design pathological demonstrations by gradually removing intuitively useful information from the standard ones to take a deep dive of the robustness of demonstration-based sequence labeling and show that (1) demonstrations composed of random tokens still make the model a better few-shot learner; (2) the length of random demonstrations and the relevance of random tokens are the main factors affecting the performance; (3) demonstrations increase the confidence of model predictions on captured superficial patterns. We have publicly released our code at https://github.com/SALT-NLP/RobustDemo.

**Entailer: Answering Questions with Faithful and Truthful Chains of Reasoning**
*Oyvind Tafjord, Bhavana Dalvi Mishra and Peter Clark* 09:24-09:36 (Hall A, Room C)
Our goal is a question-answering (QA) system that can show how its answers are implied by its own internal beliefs via a systematic chain of reasoning. Such a capability would allow better understanding of why a model produced the answer it did. Our approach is to recursively combine a trained backward-chaining model, capable of generating a set of premises entailing an answer hypothesis, with a verifier that checks that the model itself believes those premises (and the entailment itself) through self-querying. To our knowledge, this is the first system to generate multistep chains that are both faithful (the answer follows from the reasoning) and truthful (the chain reflects the system's own internal beliefs). In evaluation using two different datasets, users judge that a majority (70%+) of generated chains clearly show how an answer follows from a set of facts - substantially better than a high-performance baseline - while preserving answer accuracy. By materializing model beliefs that systematically support an answer, new opportunities arise for understanding the model's system of belief, and diagnosing and correcting its misunderstandings when an answer is wrong.

**Revisiting Parameter-Efficient Tuning: Are We Really There Yet?**
*Guanzheng Chen, Fangyu Liu, Zaiqiao Meng and Shangsong Liang* 09:36-09:48 (Hall A, Room C)
Parameter-Efficient Tuning (PETuning) methods have been deemed by many as the new paradigm for using pretrained language models (PLMs). By tuning just a fraction amount of parameters comparing to full model finetuning, PETuning methods claim to have achieved performance on par with or even better than finetuning. In this work, we take a step back and re-examine these PETuning methods by conducting the first comprehensive investigation into the training and evaluation of them. We found the problematic validation and testing practice in current studies, when accompanied by the instability nature of PETuning methods, has led to unreliable conclusions. When being compared under a truly fair evaluation protocol, PETuning cannot yield consistently competitive performance while finetuning remains to be the best-performing method in medium- and high-resource settings. We delve deeper into the cause of the instability and observed that the number of trainable parameters and training iterations are two main factors: reducing trainable parameters and prolonging training iterations may lead to higher stability in PETuning methods.

**Word Order Matters When You Increase Masking**
*Karim Lasri, Alessandro Lenci and Thierry Poibeau* 09:48-10:00 (Hall A, Room C)
Word order, an essential property of natural languages, is injected in Transformer-based neural language models using position encoding. However, recent experiments have shown that explicit position encoding is not always useful, since some models without such feature managed to achieve state-of-the art performance on some tasks. To understand better this phenomenon, we examine the effect of removing position encodings on the pre-training objective itself (i.e., masked language modelling), to test whether models can reconstruct position information from co-occurrences alone. We do so by controlling the amount of masked tokens in the input sentence, as a proxy to affect the importance of position information for the task. We find that the necessity of position information increases with the amount of masking, and that masked language models without position encodings are not able to reconstruct this information on the task. These findings point towards a direct relationship between the amount of masking and the ability of Transformers to capture order-sensitive aspects of language using position encoding.

**Stop Measuring Calibration When Humans Disagree**
*Joris Baan, Wilker Aziz, Barbara Plank and Raquel Fernandez* 10:00-10:12 (Hall A, Room C)
Calibration is a popular framework to evaluate whether a classifier knows when it does not know - i.e., its predictive probabilities are a good indication of how likely a prediction is to be correct. Correctness is commonly estimated against the human majority class. Recently, calibration to human majority has been measured on tasks where humans inherently disagree about which class applies. We show that measuring calibration to human majority given inherent disagreements is theoretically problematic, demonstrate this empirically on the ChaosNLI

dataset, and derive several instance-level measures of calibration that capture key statistical properties of human judgements - including class frequency, ranking and entropy.

**Are Hard Examples also Harder to Explain? A Study with Human and Model-Generated Explanations**
*Swarnadeep Saha, Peter Hase, Nazneen Rajani and Mohit Bansal* 10:12-10:24 (Hall A, Room C)
Recent work on explainable NLP has shown that few-shot prompting can enable large pre-trained language models (LLMs) to generate grammatical and factual natural language explanations for data labels. In this work, we study the connection between explainability and sample hardness by investigating the following research question – "Are LLMs and humans equally good at explaining data labels for both easy and hard samples?" We answer this question by first collecting human-written explanations in the form of generalizable commonsense rules on the task of Winograd Schema Challenge (Winogrande dataset). We compare these explanations with those generated by GPT-3 while varying the hardness of the test samples as well as the in-context samples. We observe that (1) GPT-3 explanations are as grammatical as human explanations regardless of the hardness of the test samples, (2) for easy examples, GPT-3 generates highly supportive explanations but human explanations are more generalizable, and (3) for hard examples, human explanations are significantly better than GPT-3 explanations both in terms of label-supportiveness and generalizability judgements. We also find that hardness of the in-context examples impacts the quality of GPT-3 explanations. Finally, we show that the supportiveness and generalizability aspects of human explanations are also impacted by sample hardness, although by a much smaller margin than models.

## NLP Applications 2 & TACL

09:00-10:30 (Hall A, Room D)

**Metric-guided Distillation: Distilling Knowledge from the Metric to Ranker and Retriever for Generative Commonsense Reasoning**
*Xingwei He, Yeyun Gong, A-Long Jin, Weizhen Qi, Hang Zhang, Jian Jiao, Bartuer Zhou, Biao Cheng, SM Yiu and Nan Duan* 09:00-09:15 (Hall A, Room D)
Commonsense generation aims to generate a realistic sentence describing a daily scene under the given concepts, which is very challenging, since it requires models to have relational reasoning and compositional generalization capabilities. Previous work focuses on retrieving prototype sentences for the provided concepts to assist generation. They first use a sparse retriever to retrieve candidate sentences, then re-rank the candidates with a ranker. However, the candidates returned by their ranker may not be the most relevant sentences, since the ranker treats all candidates equally without considering their relevance to the reference sentences of the given concepts. Another problem is that re-ranking is very expensive, but only using retrievers will seriously degrade the performance of their generation models. To solve these problems, we propose the metric distillation rule to distill knowledge from the metric (e.g., BLEU) to the ranker. We further transfer the critical knowledge summarized by the distilled ranker to the retriever. In this way, the relevance scores of candidate sentences predicted by the ranker and retriever will be more consistent with their quality measured by the metric. Experimental results on the CommonGen benchmark verify the effectiveness of our proposed method: (1) Our generation model with the distilled ranker achieves a new state-of-the-art result. (2) Our generation model with the distilled retriever even surpasses the previous SOTA.

**Segmenting Numerical Substitution Ciphers**
*Nada Aldarrab and Jonathan May* 09:15-09:30 (Hall A, Room D)
Deciphering historical substitution ciphers is a challenging problem. Example problems that have been previously studied include detecting cipher type, detecting plaintext language, and acquiring the substitution key for segmented ciphers. However, attacking unsegmented ciphers is still a challenging task. Segmentation (i.e. finding substitution units) is essential for cracking those ciphers. In this work, we propose the first automatic methods to segment those ciphers using Byte Pair Encoding (BPE) and unigram language models. Our methods achieve an average segmentation error of 2

**Deconfounding Legal Judgment Prediction for European Court of Human Rights Cases Towards Better Alignment with Experts**
*T.Y.S.S Santosh, Shanshan Xu, Oana Ichim and Matthias Grabmair* 09:30-09:45 (Hall A, Room D)
This work demonstrates that Legal Judgement Prediction systems without expert-informed adjustments can be vulnerable to shallow, distracting surface signals that arise from corpus construction, case distribution, and confounding factors. To mitigate this, we use domain expertise to strategically identify statistically predictive but legally irrelevant information. We adopt adversarial training to prevent the system from relying on it. We evaluate our deconfounded models by employing interpretability techniques and comparing to expert annotations. Quantitative experiments and qualitative analysis show that our deconfounded model consistently aligns better with expert rationales than baselines trained for prediction only. We further contribute a set of reference expert annotations to the validation and testing partitions of an existing benchmark dataset of European Court of Human Rights cases.

**PLM-based World Models for Text-based Games**
*Minsoo Kim, Yeonjoon Jung, Dohyeon Lee and Seung-won Hwang* 09:45-10:00 (Hall A, Room D)
World models have improved the ability of reinforcement learning agents to operate in a sample efficient manner, by being trained to predict plausible changes in the underlying environment. As the core tasks of world models are future prediction and commonsense understanding, our claim is that pre-trained language models (PLMs) already provide a strong base upon which to build world models. Worldformer is a recently proposed world model for text-based game environments, based only partially on PLM and transformers. Our distinction is to fully leverage PLMs as actionable world models in text-based game environments, by reformulating generation as constrained decoding which decomposes actions into verb templates and objects. We show that our model improves future valid action prediction and graph change prediction. Additionally, we show that our model better reflects commonsense than standard PLM.

**ConReader: Exploring Implicit Relations in Contracts for Contract Clause Extraction**
*Weiwen Xu, Yang Deng, Wenqiang Lei, Wenlong ZHAO, Tat-Seng Chua and Wai Lam* 10:00-10:15 (Hall A, Room D)
We study automatic Contract Clause Extraction (CCE) by modeling implicit relations in legal contracts. Existing CCE methods mostly treat contracts as plain text, creating a substantial barrier to understanding contracts of high complexity. In this work, we first comprehensively analyze the complexity issues of contracts and distill out three implicit relations commonly found in contracts, namely, 1) Long-range Context Relation that captures the correlations of distant clauses; 2) Term-Definition Relation that captures the relation between important terms with their corresponding definitions, and 3) Similar Clause Relation that captures the similarities between clauses of the same type. Then we propose a novel framework ConReader to exploit the above three relations for better contract understanding and improving CCE. Experimental results show that ConReader makes the prediction more interpretable and achieves new state-of-the-art on two CCE tasks in both conventional and zero-shot settings.

[TACL] **Modeling Non-Cooperative Dialogue: Theoretical and Empirical Insights**
*Anthony Sicilia, Tristan Maidment, Pat Healy and Malihe Alikhani* 10:15-10:30 (Hall A, Room D)
Investigating cooperativity of interlocutors is central in studying pragmatics of dialogue. Models of conversation that only assume cooperative agents fail to explain the dynamics of strategic conversations. Thus, we investigate the ability of agents to identify non-cooperative interlocutors while completing a concurrent visual-dialogue task. Within this novel setting, we study the optimality of communication strategies for achieving this multi-task objective. We use the tools of learning theory to develop a theoretical model for identifying non-cooperative interlocutors and apply this theory to analyze different communication strategies. We also introduce a corpus of non-cooperative conversations about images in the GuessWhat?! dataset proposed by De Vries et al. (2017). We use reinforcement learning to implement multiple communication strategies in this context and find empirical results validate our theory.

## Unsupervised and Weakly Supervised Methods

09:00-10:30 (Hall B)

**Bilingual Lexicon Induction for Low-Resource Languages using Graph Matching via Optimal Transport**
*Kelly Marchisio, Ali Saad-Eldin, Kevin Duh, carey priebe and Philipp Koehn* 09:00-09:15 (Hall B)
Bilingual lexicons form a critical component of various natural language processing applications, including unsupervised and semisupervised machine translation and crosslingual information retrieval. In this work, we improve bilingual lexicon induction performance across 40 language pairs with a graph-matching method based on optimal transport. The method is especially strong with low amounts of supervision.

**Zero-Shot Text Classification with Self-Training**
*Ariel Gera, Alon Halfon, Eyal Shnarch, Yotam Perlitz, Liat Ein-Dor and Noam Slonim* 09:15-09:30 (Hall B)
Recent advances in large pretrained language models have increased attention to zero-shot text classification. In particular, models finetuned on natural language inference datasets have been widely adopted as zero-shot classifiers due to their promising results and off-the-shelf availability. However, the fact that such models are unfamiliar with the target task can lead to instability and performance issues. We propose a plug-and-play method to bridge this gap using a simple self-training approach, requiring only the class names along with an unlabeled dataset, and without the need for domain expertise or trial and error. We show that fine-tuning the zero-shot classifier on its most confident predictions leads to significant performance gains across a wide range of text classification tasks, presumably since self-training adapts the zero-shot model to the task at hand.

**Fine-grained Category Discovery under Coarse-grained supervision with Hierarchical Weighted Self-contrastive Learning**
*Wenbin An, Feng Tian, Ping Chen, Siliang Tang, Qinghua Zheng and QianYing Wang* 09:30-09:45 (Hall B)
Novel category discovery aims at adapting models trained on known categories to novel categories. Previous works only focus on the scenario where known and novel categories are of the same granularity. In this paper, we investigate a new practical scenario called Fine-grained Category Discovery under Coarse-grained supervision (FCDC). FCDC aims at discovering fine-grained categories with only coarse-grained labeled data, which can adapt models to categories of different granularity from known ones and reduce significant labeling cost. It is also a challenging task since supervised training on coarse-grained categories tends to focus on inter-class distance (distance between coarse-grained classes) but ignore intra-class distance (distance between fine-grained sub-classes) which is essential for separating fine-grained categories. Considering most current methods cannot transfer knowledge from coarse-grained level to fine-grained level, we propose a hierarchical weighted self-contrastive network by building a novel weighted self-contrastive module and combining it with supervised learning in a hierarchical manner. Extensive experiments on public datasets show both effectiveness and efficiency of our model over compared methods.

**Learning Instructions with Unlabeled Data for Zero-Shot Cross-Task Generalization**
*Yuxian Gu, Pei Ke, Xiaoyan Zhu and Minlie Huang* 09:45-10:00 (Hall B)
Training language models to learn from human instructions for zero-shot cross-task generalization has attracted much attention in NLP communities. Recently, instruction tuning (IT), which fine-tunes a pre-trained language model on a massive collection of tasks described via human-craft instructions, has been shown effective in instruction learning for unseen tasks. However, IT relies on a large amount of human-annotated samples, which restricts its generalization. Unlike labeled data, unlabeled data are often massive and cheap to obtain. In this work, we study how IT can be improved with unlabeled data. We first empirically explore the IT performance trends versus the number of labeled data, instructions, and training tasks. We find it critical to enlarge the number of training instructions, and the instructions can be underutilized due to the scarcity of labeled data. Then, we propose Unlabeled Data Augmented Instruction Tuning (UDIT) to take better advantage of the instructions during IT by constructing pseudo-labeled data from unlabeled plain texts. We conduct extensive experiments to show UDIT's effectiveness in various scenarios of tasks and datasets. We also comprehensively analyze the key factors of UDIT to investigate how to better improve IT with unlabeled data. The code is publicly available at https://github.com/thu-coai/UDIT.

**Learning to Adapt to Low-Resource Paraphrase Generation**
*Zhigen Li, Yanmeng Wang, Rizhao Fan, Ye Wang, Jianfeng Li and Shaojun Wang* 10:00-10:15 (Hall B)
Paraphrase generation is a longstanding NLP task and achieves great success with the aid of large corpora. However, transferring a paraphrasing model to another domain encounters the problem of domain shifting especially when the data is sparse. At the same time, widely using large pre-trained language models (PLMs) faces the overfitting problem when training on scarce labeled data. To mitigate these two issues, we propose, LAPA, an effective adapter for PLMs optimized by meta-learning. LAPA has three-stage training on three types of related resources to solve this problem: 1. pre-training PLMs on unsupervised corpora, 2. inserting an adapter layer and meta-training on source domain labeled data, and 3. fine-tuning adapters on a small amount of target domain labeled data. This method enables paraphrase generation models to learn basic language knowledge first, then learn the paraphrasing task itself later, and finally adapt to the target task. Our experimental results demonstrate that LAPA achieves state-of-the-art in supervised, unsupervised, and low-resource settings on three benchmark datasets. With only 2% of trainable parameters and 1% labeled data of the target task, our approach can achieve a competitive performance with previous work.

**ToKen: Task Decomposition and Knowledge Infusion for Few-Shot Hate Speech Detection**
*Badr AlKhamissi, Faisal Ladhak, Srinivasan Iyer, Veselin Stoyanov, Zornitsa Kozareva, Xian Li, Pascale Fung, Lambert Mathias, Asli Celikyilmaz and Mona Diab* 10:15-10:30 (Hall B)
Hate speech detection is complex; it relies on commonsense reasoning, knowledge of stereotypes, and an understanding of social nuance that differs from one culture to the next. It is also difficult to collect a large-scale hate speech annotated dataset. In this work, we frame this problem as a few-shot learning task, and show significant gains with decomposing the task into its "constituent" parts. In addition, we see that infusing knowledge from reasoning datasets (e.g. ATOMIC2020) improves the performance even further. Moreover, we observe that the trained models generalize to out-of-distribution datasets, showing the superiority of task decomposition and knowledge infusion compared to

previously used methods. Concretely, our method outperforms the baseline by 17.83% absolute gain in the 16-shot case.

## Industry 3

09:00-10:30 (Collaboratorium)

[INDUSTRY] **Improving Precancerous Case Characterization via Transformer-based Ensemble Learning**
*Yizhen Zhong, Jiajie Xiao, Thomas Vetterli, Mahan Matin, Ellen Loo, Jimmy Lin, Richard Bourgon and Ofer Shapira* 09:00-09:15 (Collaboratorium)
The application of natural language processing (NLP) to cancer pathology reports has been focused on detecting cancer cases, largely ignoring precancerous cases. Improving the characterization of precancerous adenomas assists in developing diagnostic tests for early cancer detection and prevention, especially for colorectal cancer (CRC). Here we developed transformer-based deep neural network NLP models to perform the CRC phenotyping, with the goal of extracting precancerous lesion attributes and distinguishing cancer and precancerous cases. We achieved 0.914 macro-F1 scores for classifying patients into negative, non-advanced adenoma, advanced adenoma and CRC. We further improved the performance to 0.923 using an ensemble of classifiers for cancer status classification and lesion size named-entity recognition (NER). Our results demonstrated the potential of using NLP to leverage real-world health record data to facilitate the development of diagnostic tests for early cancer prevention.

[INDUSTRY] **Improving Large-Scale Conversational Assistants using Model Interpretation based Training Sample Selection**
*Stefan Schroedl, Manoj Kumar, Kiana Hajebi, Morteza Ziyadi, Sriram Venkatapathy, Anil Ramakrishna, Rahul Gupta and Pradeep Natarajan* 09:15-09:30 (Collaboratorium)
This paper presents an approach to identify samples from live traffic where the customer implicitly communicated satisfaction with Alexa's responses, by leveraging interpretations of model behavior. Such customer signals are noisy and adding a large number of samples from live traffic to training set makes re-training infeasible. Our work addresses these challenges by identifying a small number of samples that grow training set by 0.05% while producing statistically significant improvements in both offline and online tests.

[INDUSTRY] **CoCoID: Learning Contrastive Representations and Compact Clusters for Semi-Supervised Intent Discovery**
*Qian Cao, Deyi Xiong, Qinlong Wang and Xia Peng* 09:30-09:45 (Collaboratorium)
Intent discovery is to mine new intents from user utterances, which are not present in the set of manually predefined intents. Previous approaches to intent discovery usually automatically cluster novel intents with prior knowledge from intent-labeled data in a semi-supervised way. In this paper, we focus on the discriminative user utterance representation learning and the compactness of the learned intent clusters. We propose a novel semi-supervised intent discovery framework CoCoID with two essential components: contrastive user utterance representation learning and intra-cluster knowledge distillation. The former attempts to detect similar and dissimilar intents from a minibatch-wise perspective. The latter regularizes the predictive distribution of the model over samples in a cluster-wise way. We conduct experiments on both real-life challenging datasets (i.e., CLINC and BANKING) that are curated to emulate the true environment of commercial/production systems and traditional datasets (i.e., StackOverflow and DBPedia) to evaluate the proposed CoCoID. Experiment results demonstrate that our model substantially outperforms state-of-the-art intent discovery models (12 baselines) by over 1.4 ACC and ARI points and 1.1 NMI points across the four datasets. Further analyses suggest that CoCoID is able to learn contrastive representations and compact clusters for intent discovery.

[INDUSTRY] **Automatic Scene-based Topic Channel Construction System for E-Commerce**
*Peng Lin, Yanyan Zou, Lingfei Wu, Mian Ma, Zhuoye Ding and Bo Long* 09:45-10:00 (Collaboratorium)
Scene marketing that well demonstrates user interests within a certain scenario has proved effective for offline shopping. To conduct scene marketing for e-commerce platforms, this work presents a novel product form, scene-based topic channel which typically consists of a list of diverse products belonging to the same usage scenario and a topic title that describes the scenario with marketing words. As manual construction of channels is time-consuming due to billions of products as well as dynamic and diverse customers' interests, it is necessary to leverage AI techniques to automatically construct channels for certain usage scenarios and even discover novel topics. To be specific, we first frame the channel construction task as a two-step problem, i.e., scene-based topic generation and product clustering, and propose an E-commerce Scene-based Topic Channel construction system (i.e., ESTC) to achieve automated production, consisting of scene-based topic generation model for the e-commerce domain, product clustering on the basis of topic similarity, as well as quality control based on automatic model filtering and human screening. Extensive offline experiments and online A/B test validates the effectiveness of such a novel product form as well as the proposed system. In addition, we also introduce the experience of deploying the proposed system on a real-world e-commerce recommendation platform.

[INDUSTRY] **Gaining Insights into Unrecognized User Utterances in Task-Oriented Dialog Systems**
*Ella Rabinovich, Matan Vetzler, David Boaz, Vineet Kumar, Gaurav Pandey and Ateret Anaby Tavor* 10:00-10:15 (Collaboratorium)
The rapidly growing market demand for automatic dialogue agents capable of goal-oriented behavior has caused many tech-industry leaders to invest considerable efforts into task-oriented dialog systems. The success of these systems is highly dependent on the accuracy of their intent identification – the process of deducing the goal or meaning of the user's request and mapping it to one of the known intents for further processing. Gaining insights into unrecognized utterances – user requests the systems fails to attribute to a known intent – is therefore a key process in continuous improvement of goal-oriented dialog systems. We present an end-to-end pipeline for processing unrecognized user utterances, deployed in a real-world, commercial task-oriented dialog system, including a specifically-tailored clustering algorithm, a novel approach to cluster representative extraction, and cluster naming. We evaluated the proposed components, demonstrating their benefits in the analysis of unrecognized user requests.

[INDUSTRY] **CGF: Constrained Generation Framework for Query Rewriting in Conversational AI**
*Jie Hao, Yang Liu, Xing Fan, Saurabh Gupta, Saleh Soltan, Rakesh Chada, Pradeep Natarajan, Chenlei Guo and Gokhan Tur* 10:15-10:30 (Collaboratorium)
In conversational AI agents, Query Rewriting (QR) plays a crucial role in reducing user frictions and satisfying their daily demands. User frictions are caused by various reasons, such as errors in the conversational AI system, users' accent or their abridged language. In this work, we present a novel Constrained Generation Framework (CGF) for query rewriting at both global and personalized levels. It is based on the encoder-decoder framework, where the encoder takes the query and its previous dialogue turns as the input to form a context-enhanced representation, and the decoder uses constrained decoding to generate the rewrites based on the pre-defined global or personalized constrained decoding space. Extensive offline and online A/B experiments show that the proposed CGF significantly boosts the query rewriting performance.

## Poster Sessions 13 & 14

09:00-10:30 (Atrium)

**Discourse-Aware Soft Prompting for Text Generation**
*Marjan Ghazvininejad, Vladimir Karpukhin, Vera Gor and Asli Celikyilmaz* 09:00-10:30 (Atrium)
Current efficient fine-tuning methods (e.g., adapters, prefix-tuning, etc.) have optimized conditional text generation via training a small set of extra parameters of the neural language model, while freezing the rest for efficiency. While showing strong performance on some generation tasks, they don't generalize across all generation tasks. We show that soft-prompt based conditional text generation can be improved with simple and efficient methods that simulate modeling the discourse structure of human written text. We investigate two design choices: First, we apply hierarchical blocking on the prefix parameters to simulate a higher-level discourse structure of human written text. Second, we apply attention sparsity on the prefix parameters at different layers of the network and learn sparse transformations on the softmax-function. We show that structured design of prefix parameters yields more coherent, faithful and relevant generations than the baseline prefix-tuning on all generation tasks.

**Mitigating Data Sparsity for Short Text Topic Modeling by Topic-Semantic Contrastive Learning**
*Xiaobao Wu, Anh Tuan Luu and Xinshuai Dong* 09:00-10:30 (Atrium)
To overcome the data sparsity issue in short text topic modeling, existing methods commonly rely on data augmentation or the data characteristic of short texts to introduce more word co-occurrence information. However, most of them do not make full use of the augmented data or the data characteristic: they insufficiently learn the relations among samples in data, leading to dissimilar topic distributions of semantically similar text pairs. To better address data sparsity, in this paper we propose a novel short text topic modeling framework, Topic-Semantic Contrastive Topic Model (TSCTM). To sufficiently model the relations among samples, we employ a new contrastive learning method with efficient positive and negative sampling strategies based on topic semantics. This contrastive learning method refines the representations, enriches the learning signals, and thus mitigates the sparsity issue. Extensive experimental results show that our TSCTM outperforms state-of-the-art baselines regardless of the data augmentation availability, producing high-quality topics and topic distributions.

**Analyzing and Evaluating Faithfulness in Dialogue Summarization**
*Bin Wang, Chen Zhang, Yan Zhang, Yiming Chen and Haizhou Li* 09:00-10:30 (Atrium)
Dialogue summarization is abstractive in nature, making it suffer from factual errors. The factual correctness of summaries has the highest priority before practical applications. Many efforts have been made to improve faithfulness in text summarization. However, there is a lack of systematic study on dialogue summarization systems. In this work, we first perform the fine-grained human analysis on the faithfulness of dialogue summaries and observe that over 35% of generated summaries are faithfully inconsistent respective the source dialogues. Furthermore, we present a new model-level faithfulness evaluation method. It examines generation models with multi-choice questions created by rule-based transformations. Experimental results show that our evaluation schema is a strong proxy for the factual correctness of summarization models. The human-annotated faithfulness samples and the evaluation toolkit are released to facilitate future research toward faithful dialogue summarization.

**BotsTalk: Machine-sourced Framework for Automatic Curation of Large-scale Multi-skill Dialogue Datasets**
*Minju Kim, Chaehyeong Kim, Yong Ho Song, Seung-won Hwang and Jinyoung Yeo* 09:00-10:30 (Atrium)
To build open-domain chatbots that are able to use diverse communicative skills, we propose a novel framework BotsTalk, where multiple agents grounded to the specific target skills participate in a conversation to automatically annotate multi-skill dialogues. We further present Blended Skill BotsTalk (BSBT), a large-scale multi-skill dialogue dataset comprising 300K conversations. Through extensive experiments, we demonstrate that our dataset can be effective for multi-skill dialogue systems which require an understanding of skill blending as well as skill grounding. Our code and data are available at https://github.com/convei-lab/BotsTalk.

**LittleBird: Efficient Faster & Longer Transformer for Question Answering**
*Minchul Lee, Kijong Han and Myeong Cheol Shin* 09:00-10:30 (Atrium)
BERT has shown a lot of sucess in a wide variety of NLP tasks. But it has a limitation dealing with long inputs due to its attention mechanism. Longformer, ETC and BigBird addressed this issue and effectively solved the quadratic dependency problem. However we find that these models are not sufficient, and propose LittleBird, a novel model based on BigBird with improved speed and memory footprint while maintaining accuracy. In particular, we devise a more flexible and efficient position representation method based on Attention with Linear Biases(ALiBi). We also show that replacing the method of global information represented in the BigBird with pack and unpack attention is more effective. The proposed model can work on long inputs even after being pre-trained on short inputs, and can be trained efficiently reusing existing pre-trained language model for short inputs. This is a significant benefit for low-resource languages where large amounts of long text data are difficult to obtain. As a result, our experiments show that LittleBird works very well in a variety of languages, achieving high performance in question answering tasks, particularly in KorQuAD2.0, Korean Question Answering Dataset for long paragraphs.

**How "Multi" is Multi-Document Summarization?**
*Ruben Wolhandler, Arie Cattan, Ori Ernst and Ido Dagan* 09:00-10:30 (Atrium)
The task of multi-document summarization (MDS) aims at models that, given multiple documents as input, are able to generate a summary that combines disperse information, originally spread __across__ these documents. Accordingly, it is expected that both reference summaries in MDS datasets, as well as system summaries, would indeed be based on such dispersed information. In this paper, we argue for quantifying and assessing this expectation. To that end, we propose an automated measure for evaluating the degree to which a summary is "disperse", in the sense of the number of source documents needed to cover its content. We apply our measure to empirically analyze several popular MDS datasets, with respect to their reference summaries, as well as the output of state-of-the-art systems. Our results show that certain MDS datasets barely require combining information from multiple documents, where a single document often covers the full summary content. Overall, we advocate using our metric for assessing and improving the degree to which summarization datasets require combining multi-document information, and similarly how summarization models actually meet this challenge.

**Transformer-based Entity Typing in Knowledge Graphs**
*Zhiwei Hu, Victor Gutierrez-Basulto, Zhiliang Xiang, Ru Li and Jeff Pan* 09:00-10:30 (Atrium)
We investigate the knowledge graph entity typing task which aims at inferring plausible entity types. In this paper, we propose a novel Transformer-based Entity Typing (TET) approach, effectively encoding the content of neighbours of an entity by means of a transformer mechanism. More precisely, TET is composed of three different mechanisms: a local transformer allowing to infer missing entity types by independently encoding the information provided by each of its neighbours; a global transformer aggregating the information of all neighbours of an entity into a single long sequence to reason about more complex entity types; and a context transformer integrating neighbours content in a differentiated way through information exchange between neighbour pairs, while preserving the graph structure. Furthermore, TET uses information about class membership of types to semantically strengthen the representation of an entity. Experiments on two real-world

datasets demonstrate the superior performance of TET compared to the state-of-the-art.

**NewsClaims: A New Benchmark for Claim Detection from News with Attribute Knowledge**
*Revanth Gangi Reddy, Sai Chetan Chinthakindi, Zhenhailong Wang, Yi Fung, Kathryn Conger, Ahmed ELsayed, Martha Palmer, Preslav Nakov, Eduard Hovy, Kevin Small and Heng Ji* 09:00-10:30 (Atrium)
Claim detection and verification are crucial for news understanding and have emerged as promising technologies for mitigating misinformation and disinformation in the news. However, most existing work has focused on claim sentence analysis while overlooking additional crucial attributes (e.g., the claimer and the main object associated with the claim).In this work, we present NewsClaims, a new benchmark for attribute-aware claim detection in the news domain. We extend the claim detection problem to include extraction of additional attributes related to each claim and release 889 claims annotated over 143 news articles. NewsClaims aims to benchmark claim detection systems in emerging scenarios, comprising unseen topics with little or no training data. To this end, we see that zero-shot and prompt-based baselines show promising performance on this benchmark, while still considerably behind human performance.

**IsoVec: Controlling the Relative Isomorphism of Word Embedding Spaces**
*Kelly Marchisio, Neha Verma, Kevin Duh and Philipp Koehn* 09:00-10:30 (Atrium)
The ability to extract high-quality translation dictionaries from monolingual word embedding spaces depends critically on the geometric similarity of the spaces—their degree of "isomorphism." We address the root-cause of faulty cross-lingual mapping: that word embedding training resulted in the underlying spaces being non-isomorphic. We incorporate global measures of isomorphism directly into the skipgram loss function, successfully increasing the relative isomorphism of trained word embedding spaces and improving their ability to be mapped to a shared cross-lingual space. The result is improved bilingual lexicon induction in general data conditions, under domain mismatch, and with training algorithm dissimilarities. We release IsoVec at https://github.com/kellymarchisio/isovec.

**Natural Logic-guided Autoregressive Multi-hop Document Retrieval for Fact Verification**
*Rami Aly and Andreas Vlachos* 09:00-10:30 (Atrium)
A key component of fact verification is the evidence retrieval, often from multiple documents. Recent approaches use dense representations and condition the retrieval of each document on the previously retrieved ones. The latter step is performed over all the documents in the collection, requiring storing their dense representations in an index, thus incurring a high memory footprint. An alternative paradigm is retrieve-and-rerank, where documents are retrieved using methods such as BM25, their sentences are reranked, and further documents are retrieved conditioned on these sentences, reducing the memory requirements. However, such approaches can be brittle as they rely on heuristics and assume hyperlinks between documents.

We propose a novel retrieve-and-rerank method for multi-hop retrieval, that consists of a retriever that jointly scores documents in the knowledge source and sentences from previously retrieved documents using an autoregressive formulation and is guided by a proof system based on natural logic that dynamically terminates the retrieval process if the evidence is deemed sufficient.

This method exceeds or is on par with the current state-of-the-art on FEVER, HoVer and FEVEROUS-S, while using 5 to 10 times less memory than competing systems. Evaluation on an adversarial dataset indicates improved stability of our approach compared to commonly deployed threshold-based methods. Finally, the proof system helps humans predict model decisions correctly more often than using the evidence alone.

**CEFR-Based Sentence Difficulty Annotation and Assessment**
*Yuki Arase, Satoru Uchida and Tomoyuki Kajiwara* 09:00-10:30 (Atrium)
Controllable text simplification is a crucial assistive technique for language learning and teaching. One of the primary factors hindering its advancement is the lack of a corpus annotated with sentence difficulty levels based on language ability descriptions. To address this problem, we created the CEFR-based Sentence Profile (CEFR-SP) corpus, containing 17k English sentences annotated with the levels based on the Common European Framework of Reference for Languages assigned by English-education professionals. In addition, we propose a sentence-level assessment model to handle unbalanced level distribution because the most basic and highly proficient sentences are naturally scarce. In the experiments in this study, our method achieved a macro-F1 score of 84.5% in the level assessment, thus outperforming strong baselines employed in readability assessment.

**Factorizing Content and Budget Decisions in Abstractive Summarization of Long Documents**
*Marcio Fonseca, Yftah Ziser and Shay B. Cohen* 09:00-10:30 (Atrium)
We argue that disentangling content selection from the budget used to cover salient content improves the performance and applicability of abstractive summarizers. Our method, FactorSum, does this disentanglement by factorizing summarization into two steps through an energy function: (1) generation of abstractive summary views covering salient information in subsets of the input document (document views); (2) combination of these views into a final summary, following a budget and content guidance. This guidance may come from different sources, including from an advisor model such as BART or BigBird, or in oracle mode – from the reference. This factorization achieves significantly higher ROUGE scores on multiple benchmarks for long document summarization, namely PubMed, arXiv, and GovReport. Most notably, our model is effective for domain adaptation. When trained only on PubMed samples, it achieves a 46.29 ROUGE-1 score on arXiv, outperforming PEGASUS trained in domain by a large margin. Our experimental results indicate that the performance gains are due to more flexible budget adaptation and processing of shorter contexts provided by partial document views.

**Understanding and Improving Knowledge Distillation for Quantization Aware Training of Large Transformer Encoders**
*Minsoo Kim, Sihwa Lee, Suk-Jin Hong, Du-Seong Chang and Jungwook Choi* 09:00-10:30 (Atrium)
Knowledge distillation (KD) has been a ubiquitous method for model compression to strengthen the capability of a lightweight model with the transferred knowledge from the teacher. In particular, KD has been employed in quantization-aware training (QAT) of Transformer encoders like BERT to improve the accuracy of the student model with the reduced-precision weight parameters. However, little is understood about which of the various KD approaches best fits the QAT of Transformers. In this work, we provide an in-depth analysis of the mechanism of KD on attention recovery of quantized large Transformers. In particular, we reveal that the previously adopted MSE loss on the attention score is insufficient for recovering the self-attention information. Therefore, we propose two KD methods; attention-map and attention-output losses. Furthermore, we explore the unification of both losses to address task-dependent preference between attention-map and output losses. The experimental results on various Transformer encoder models demonstrate that the proposed KD methods achieve state-of-the-art accuracy for QAT with sub-2-bit weight quantization.

**Synergy with Translation Artifacts for Training and Inference in Multilingual Tasks**
*Jaehoon Oh, Jongwoo Ko and Se-Young Yun* 09:00-10:30 (Atrium)
Translation has played a crucial role in improving the performance on multilingual tasks: (1) to generate the target language data from the source language data for training and (2) to generate the source language data from the target language data for inference. However, prior works have not considered the use of both translations simultaneously. This paper shows that combining them can synergize the results on various multilingual sentence classification tasks. We empirically find that translation artifacts stylized by translators are the main factor of the performance gain. Based on this analysis, we adopt two training methods, SupCon and MixUp, considering translation artifacts. Furthermore,

we propose a cross-lingual fine-tuning algorithm called MUSC, which uses SupCon and MixUp jointly and improves the performance. Our code is available at https://github.com/jongwooko/MUSC.

**Topical Segmentation of Spoken Narratives: A Test Case on Holocaust Survivor Testimonies**
*Eitan Wagner, Renana Keydar, Amit Pinchevski and Omri Abend* 09:00-10:30 (Atrium)
The task of topical segmentation is well studied, but previous work has mostly addressed it in the context of structured, well-defined segments, such as segmentation into paragraphs, chapters, or segmenting text that originated from multiple sources. We tackle the task of segmenting running (spoken) narratives, which poses hitherto unaddressed challenges. As a test case, we address Holocaust survivor testimonies, given in English. Other than the importance of studying these testimonies for Holocaust research, we argue that they provide an interesting test case for topical segmentation, due to their unstructured surface level, relative abundance (tens of thousands of such testimonies were collected), and the relatively confined domain that they cover. We hypothesize that boundary points between segments correspond to low mutual information between the sentences proceeding and following the boundary. Based on this hypothesis, we explore a range of algorithmic approaches to the task, building on previous work on segmentation that uses generative Bayesian modeling and state-of-the-art neural machinery. Compared to manually annotated references, we find that the developed approaches show considerable improvements over previous work.

**Enhancing Multilingual Language Model with Massive Multilingual Knowledge Triples**
*Linlin Liu, Xin Li, Ruidan He, Lidong Bing, Shafiq Joty and Luo Si* 09:00-10:30 (Atrium)
Knowledge-enhanced language representation learning has shown promising results across various knowledge-intensive NLP tasks. However, prior methods are limited in efficient utilization of multilingual knowledge graph (KG) data for language model (LM) pretraining. They often train LMs with KGs in indirect ways, relying on extra entity/relation embeddings to facilitate knowledge injection. In this work, we explore methods to make better use of the multilingual annotation and language agnostic property of KG triples, and present novel knowledge based multilingual language models (KMLMs) trained directly on the knowledge triples. We first generate a large amount of multilingual synthetic sentences using the Wikidata KG triples. Then based on the intra- and inter-sentence structures of the generated data, we design pretraining tasks to enable the LMs to not only memorize the factual knowledge but also learn useful logical patterns. Our pretrained KMLMs demonstrate significant performance improvements on a wide range of knowledge-intensive cross-lingual tasks, including named entity recognition (NER), factual knowledge retrieval, relation classification, and a newly designed logical reasoning task.

**IDK-MRC: Unanswerable Questions for Indonesian Machine Reading Comprehension**
*Rifki Afina Putri and Alice Oh* 09:00-10:30 (Atrium)
Machine Reading Comprehension (MRC) has become one of the essential tasks in Natural Language Understanding (NLU) as it is often included in several NLU benchmarks (Liang et al., 2020; Wilie et al., 2020). However, most MRC datasets only have answerable question type, overlooking the importance of unanswerable questions. MRC models trained only on answerable questions will select the span that is most likely to be the answer, even when the answer does not actually exist in the given passage (Rajpurkar et al., 2018). This problem especially remains in medium- to low-resource languages like Indonesian. Existing Indonesian MRC datasets (Purwarianti et al., 2007; Clark et al., 2020) are still inadequate because of the small size and limited question types, i.e., they only cover answerable questions. To fill this gap, we build a new Indonesian MRC dataset called I(n)don'tKnow- MRC (IDK-MRC) by combining the automatic and manual unanswerable question generation to minimize the cost of manual dataset construction while maintaining the dataset quality. Combined with the existing answerable questions, IDK-MRC consists of more than 10K questions in total. Our analysis shows that our dataset significantly improves the performance of Indonesian MRC models, showing a large improvement for unanswerable questions.

**Cross-stitching Text and Knowledge Graph Encoders for Distantly Supervised Relation Extraction**
*Qin Dai, Benjamin Heinzerling and Kentaro Inui* 09:00-10:30 (Atrium)
Bi-encoder architectures for distantly-supervised relation extraction are designed to make use of the complementary information found in text and knowledge graphs (KG). However, current architectures suffer from two drawbacks. They either do not allow any sharing between the text encoder and the KG encoder at all, or, in case of models with KG-to-text attention, only share information in one direction. Here, we introduce cross-stitch bi-encoders, which allow full interaction between the text encoder and the KG encoder via a cross-stitch mechanism. The cross-stitch mechanism allows sharing and updating representations between the two encoders at any layer, with the amount of sharing being dynamically controlled via cross-attention-based gates. Experimental results on two relation extraction benchmarks from two different domains show that enabling full interaction between the two encoders yields strong improvements.

**Open-domain Video Commentary Generation**
*Edison Marrese-Taylor, Yumi Hamazono, Tatsuya Ishigaki, Goran Topić, Yusuke Miyao, Ichiro Kobayashi and Hiroya Takamura*09:00-10:30 (Atrium)
Live commentary plays an important role in sports broadcasts and video games, making spectators more excited and immersed. In this context, though approaches for automatically generating such commentary have been proposed in the past, they have been generally concerned with specific fields, where it is possible to leverage domain-specific information. In light of this, we propose the task of generating video commentary in an open-domain fashion. We detail the construction of a new large-scale dataset of transcribed commentary aligned with videos containing various human actions in a variety of domains, and propose approaches based on well-known neural architectures to tackle the task. To understand the strengths and limitations of current approaches, we present an in-depth empirical study based on our data. Our results suggest clear trade-offs between textual and visual inputs for the models and highlight the importance of relying on external knowledge in this open-domain setting, resulting in a set of robust baselines for our task.

**Tutoring Helps Students Learn Better: Improving Knowledge Distillation for BERT with Tutor Network**
*Junho Kim, Jun-Hyung Park, Mingyu Lee, Wing-Lam Mok, Joon-Young Choi and SangKeun Lee* 09:00-10:30 (Atrium)
Pre-trained language models have achieved remarkable successes in natural language processing tasks, coming at the cost of increasing model size. To address this issue, knowledge distillation (KD) has been widely applied to compress language models. However, typical KD approaches for language models have overlooked the difficulty of training examples, suffering from incorrect teacher prediction transfer and sub-efficient training. In this paper, we propose a novel KD framework, Tutor-KD, which improves the distillation effectiveness by controlling the difficulty of training examples during pre-training. We introduce a tutor network that generates samples that are easy for the teacher but difficult for the student, with training on a carefully designed policy gradient method. Experimental results show that Tutor-KD significantly and consistently outperforms the state-of-the-art KD methods with variously sized student models on the GLUE benchmark, demonstrating that the tutor can effectively generate training examples for the student.

**Unifying Data Perspectivism and Personalization: An Application to Social Norms**
*Joan Plepi, Béla Neuendorf, Lucie Flek and Charles Welch* 09:00-10:30 (Atrium)
Instead of using a single ground truth for language processing tasks, several recent studies have examined how to represent and predict the labels of the set of annotators. However, often little or no information about annotators is known, or the set of annotators is small. In this work, we examine a corpus of social media posts about conflict from a set of 13k annotators and 210k judgements of social norms. We provide a novel experimental setup that applies personalization methods to the modeling of annotators and compare their effectiveness for predicting

the perception of social norms. We further provide an analysis of performance across subsets of social situations that vary by the closeness of the relationship between parties in conflict, and assess where personalization helps the most.

**Debiasing Masks: A New Framework for Shortcut Mitigation in NLU**
*Johannes Mario Meissner, Saku Sugawara and Akiko Aizawa* 09:00-10:30 (Atrium)
Debiasing language models from unwanted behaviors in Natural Language Understanding (NLU) tasks is a topic with rapidly increasing interest in the NLP community. Spurious statistical correlations in the data allow models to perform shortcuts and avoid uncovering more advanced and desirable linguistic features. A multitude of effective debiasing approaches has been proposed, but flexibility remains a major issue. For the most part, models must be retrained to find a new set of weights with debiased behavior. We propose a new debiasing method in which we identify debiased pruning masks that can be applied to a finetuned model. This enables the selective and conditional application of debiasing behaviors. We assume that bias is caused by a certain subset of weights in the network; our method is, in essence, a mask search to identify and remove biased weights. Our masks show equivalent or superior performance to the standard counterparts, while offering important benefits. Pruning masks can be stored with high efficiency in memory, and it becomes possible to switch among several debiasing behaviors (or revert back to the original biased model) at inference time. Finally, it opens the doors to further research on how biases are acquired by studying the generated masks. For example, we observed that the early layers and attention heads were pruned more aggressively, possibly hinting towards the location in which biases may be encoded.

**EUR-Lex-Sum: A Multi- and Cross-lingual Dataset for Long-form Summarization in the Legal Domain**
*Dennis Aumiller, Ashish Chouhan and Michael Gertz* 09:00-10:30 (Atrium)
Existing summarization datasets come with two main drawbacks: (1) They tend to focus on overly exposed domains, such as news articles or wiki-like texts, and (2) are primarily monolingual, with few multilingual datasets. In this work, we propose a novel dataset, called EUR-Lex-Sum, based on manually curated document summaries of legal acts from the European Union law platform (EUR-Lex). Documents and their respective summaries exist as cross-lingual paragraph-aligned data in several of the 24 official European languages, enabling access to various cross-lingual and lower-resourced summarization setups. We obtain up to 1,500 document/summary pairs per language, including a subset of 375 cross-lingually aligned legal acts with texts available in *all* 24 languages. In this work, the data acquisition process is detailed and key characteristics of the resource are compared to existing summarization resources. In particular, we illustrate challenging sub-problems and open questions on the dataset that could help the facilitation of future research in the direction of domain-specific cross-lingual summarization. Limited by the extreme length and language diversity of samples, we further conduct experiments with suitable extractive monolingual and cross-lingual baselines for future work. Code for the extraction as well as access to our data and baselines is available online at: [https://github.com/achouhan93/eur-lex-sum](https://github.com/achouhan93/eur-lex-sum).

**Hyper-X: A Unified Hypernetwork for Multi-Task Multilingual Transfer**
*Ahmet Üstün, Arianna Bisazza, Gosse Bouma, Gertjan van Noord and Sebastian Ruder* 09:00-10:30 (Atrium)
Massively multilingual models are promising for transfer learning across tasks and languages. However, existing methods are unable to fully leverage training data when it is available in different task-language combinations. To exploit such heterogeneous supervision, we propose Hyper-X, a single hypernetwork that unifies multi-task and multilingual learning with efficient adaptation. It generates weights for adapter modules conditioned on both tasks and language embeddings. By learning to combine task and language-specific knowledge, our model enables zero-shot transfer for unseen languages and task-language combinations. Our experiments on a diverse set of languages demonstrate that Hyper-X achieves the best or competitive gain when a mixture of multiple resources is available, while on par with strong baseline in the standard scenario. Hyper-X is also considerably more efficient in terms of parameters and resources compared to methods that train separate adapters. Finally, Hyper-X consistently produces strong results in few-shot scenarios for new languages, showing the versatility of our approach beyond zero-shot transfer.

**Entity-Focused Dense Passage Retrieval for Outside-Knowledge Visual Question Answering**
*Jialin Wu and Raymond Mooney* 09:00-10:30 (Atrium)
Most Outside-Knowledge Visual Question Answering (OK-VQA) systems employ a two-stage framework that first retrieves external knowledge given the visual question and then predicts the answer based on the retrieved content. However, the retrieved knowledge is often inadequate. Retrievals are frequently too general and fail to cover specific knowledge needed to answer the question. Also, the naturally available supervision (whether the passage contains the correct answer) is weak and does not guarantee question relevancy. To address these issues, we propose an Entity-Focused Retrieval (EnFoRe) model that provides stronger supervision during training and recognizes question-relevant entities to help retrieve more specific knowledge. Experiments show that our EnFoRe model achieves superior retrieval performance on OK-VQA, the currently largest outside-knowledge VQA dataset. We also combine the retrieved knowledge with state-of-the-art VQA models, and achieve a new state-of-the-art performance on OK-VQA.

**Multimodal Robustness for Neural Machine Translation**
*Yuting Zhao and Ioan Calapodescu* 09:00-10:30 (Atrium)
In this paper, we look at the case of a Generic text-to-text NMT model that has to deal with data coming from various modalities, like speech, images, or noisy text extracted from the web. We propose a two-step method, based on composable adapters, to deal with this problem of Multimodal Robustness. In a first step, we separately learn domain adapters and modality specific adapters, to deal with noisy input coming from various sources: ASR, OCR, or noisy text (UGC). In a second step, we combine these components at runtime via dynamic routing or, when the source of noise is unknown, via two new transfer learning mechanisms (Fast Fusion and Multi Fusion). We show that our method provides a flexible, state-of-the-art, architecture able to deal with noisy multimodal inputs.

**Generalizing over Long Tail Concepts for Medical Term Normalization**
*Beatrice Portelli, Simone Scaboro, Enrico Santus, Hooman Sedghamiz, Emmanuele Chersoni and Giuseppe Serra* 09:00-10:30 (Atrium)
Medical term normalization consists in mapping a piece of text to a large number of output classes. Given the small size of the annotated datasets and the extremely long tail distribution of the concepts, it is of utmost importance to develop models that are capable to generalize to scarce or unseen concepts. An important attribute of most target ontologies is their hierarchical structure. In this paper we introduce a simple and effective learning strategy that leverages such information to enhance the generalizability of both discriminative and generative models. The evaluation shows that the proposed strategy produces state-of-the-art performance on seen concepts and consistent improvements on unseen ones, allowing also for efficient zero-shot knowledge transfer across text typologies and datasets.

**Disentangling Uncertainty in Machine Translation Evaluation**
*Chrysoula Zerva, Taisiya Glushkova, Ricardo Rei and André F. T. Martins* 09:00-10:30 (Atrium)
Trainable evaluation metrics for machine translation (MT) exhibit strong correlation with human judgements, but they are often hard to interpret and might produce unreliable scores under noisy or out-of-domain data. Recent work has attempted to mitigate this with simple uncertainty quantification techniques (Monte Carlo dropout and deep ensembles), however these techniques (as we show) are limited in several ways – for example, they are unable to distinguish between different kinds of uncertainty, and they are time and memory consuming. In this paper, we propose more powerful and efficient uncertainty predictors for MT evaluation, and we assess their ability to target different

sources of aleatoric and epistemic uncertainty. To this end, we develop and compare training objectives for the COMET metric to enhance it with an uncertainty prediction output, including heteroscedastic regression, divergence minimization, and direct uncertainty prediction. Our experiments show improved results on uncertainty prediction for the WMT metrics task datasets, with a substantial reduction in computational costs. Moreover, they demonstrate the ability of these predictors to address specific uncertainty causes in MT evaluation, such as low quality references and out-of-domain data.

**POQue: Asking Participant-specific Outcome Questions for a Deeper Understanding of Complex Events**
*Sai Vallurupalli, Sayontan Ghosh, Katrin Erk, Niranjan Balasubramanian and Francis Ferraro* 09:00-10:30 (Atrium)
Knowledge about outcomes is critical for complex event understanding but is hard to acquire. We show that by pre-identifying a participant in a complex event, crowdworkers are able to (1) infer the collective impact of salient events that make up the situation, (2) annotate the volitional engagement of participants in causing the situation, and (3) ground the outcome of the situation in state changes of the participants. By creating a multi-step interface and a careful quality control strategy, we collect a high quality annotated dataset of 8K short newswire narratives and ROCStories with high inter-annotator agreement (0.74-0.96 weighted Fleiss Kappa). Our dataset, POQUe (Participant Outcome Questions), enables the exploration and development of models that address multiple aspects of semantic understanding. Experimentally, we show that current language models lag behind human performance in subtle ways through our task formulations that target abstract and specific comprehension of a complex event, its outcome, and a participant's influence over the event culmination.

**Perturbation Augmentation for Fairer NLP**
*Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela and Adina Williams* 09:00-10:30 (Atrium)
Unwanted and often harmful social biases are becoming ever more salient in NLP research, affecting both models and datasets. In this work, we ask whether training on demographically perturbed data leads to fairer language models. We collect a large dataset of human annotated text perturbations and train a neural perturbation model, which we show outperforms heuristic alternatives. We find that (i) language models (LMs) pre-trained on demographically perturbed corpora are typically more fair, and (ii) LMs finetuned on perturbed GLUE datasets exhibit less demographic bias on downstream tasks, and (iii) fairness improvements do not come at the expense of performance on downstream tasks. Lastly, we discuss outstanding questions about how best to evaluate the (un)fairness of large language models. We hope that this exploration of neural demographic perturbation will help drive more improvement towards fairer NLP.

**The Aligned Multimodal Movie Treebank: An audio, video, dependency-parse treebank**
*Adam Yaari, Jan DeWitt, Henry Hu, Bennett Stankovits, Sue Felshin, Yevgeni Berzak, Helena Aparicio, Boris Katz, Ignacio Cases and Andrei Barbu* 09:00-10:30 (Atrium)
Treebanks have traditionally included only text and were derived from written sources such as newspapers or the web. We introduce the Aligned Multimodal Movie Treebank (AMMT), an English language treebank derived from dialog in Hollywood movies which includes transcriptions of the audio-visual streams with word-level alignment, as well as part of speech tags and dependency parses in the Universal Dependencies formalism. AMMT consists of 31,264 sentences and 218,090 words, that will amount to the 3rd largest UD English treebank and the only multimodal treebank in UD. To help with the web-based annotation effort, we also introduce the Efficient Audio Alignment Annotator (EAAA), a companion tool that enables annotators to significantly speed-up their annotation processes.

**Evaluating and Improving Factuality in Multimodal Abstractive Summarization**
*David Wan and Mohit Bansal* 09:00-10:30 (Atrium)
Current metrics for evaluating factuality for abstractive document summarization have achieved high correlations with human judgment, but they do not account for the vision modality and thus are not adequate for vision-and-language summarization. We propose CLIPBERTSCORE, a simple weighted combination of CLIPScore and BERTScore to leverage the robustness and strong factuality detection performance between image-summary and document-summary, respectively. Next, due to the lack of meta-evaluation benchmarks to evaluate the quality of multimodal factuality metrics, we collect human judgments of factuality with respect to documents and images. We show that this simple combination of two metrics in the zero-shot setting achieves higher correlations than existing factuality metrics for document summarization, outperforms an existing multimodal summarization metric, and performs competitively with strong multimodal factuality metrics specifically fine-tuned for the task. Our thorough analysis demonstrates the robustness and high correlation of CLIPBERTSCORE and its components on four factuality metric-evaluation benchmarks. Finally, we demonstrate two practical downstream applications of our CLIPBERTSCORE metric: for selecting important images to focus on during training, and as a reward for reinforcement learning to improve factuality of multimodal summary generation w.r.t automatic and human evaluation.

**Revisiting Pre-trained Language Models and their Evaluation for Arabic Natural Language Processing**
*Abbas Ghaddar, Yimeng Wu, Sunyam Bagga, Ahmad Rashid, Khalil Bibi, Mehdi Rezagholizadeh, Chao Xing, Yasheng Wang, Xinyu Duan, Zhefeng Wang, baoxing Huai, Xin Jiang, Qun Liu and Phillippe Langlais* 09:00-10:30 (Atrium)
There is a growing body of work in recent years to develop pre-trained language models (PLMs) for the Arabic language. This work addresses two major problems in existing Arabic PLMs that limit the progress of the Arabic NLU and NLG fields. First, existing Arabic PLMs are not well-explored and their pre-training can be improved significantly using a more methodical approach. Second, there is a lack of systematic and reproducible evaluation of these models in the literature. We revisit both the pre-training and evaluation of Arabic PLMs. In terms of pre-training, we explore the impact of the quality of the pretraining data, the size of the model, and the incorporation of character-level information on Arabic PLM. As a result, we release three new Arabic BERT-style models ( JABER, Char-JABER, and SABER), and two T5-style models (AT5S and AT5B). In terms of evaluation, we conduct a comprehensive empirical study to systematically evaluate the performance of existing state-of-the-art models on ALUE, a leaderboard-powered benchmark for Arabic NLU tasks, and on a subset of the Arabic generative tasks. We show that our models significantly outperform existing Arabic PLMs and achieve a new state-of-the-art performance on discriminative and generative Arabic NLU and NLG tasks. Our models and source code to reproduce results will be made available upon acceptance.

**MABEL: Attenuating Gender Bias using Textual Entailment Data**
*Jacqueline He, Mengzhou Xia, Christiane Fellbaum and Danqi Chen* 09:00-10:30 (Atrium)
Pre-trained language models encode undesirable social biases, which are further exacerbated in downstream use. To this end, we propose MABEL (a Method for Attenuating Gender Bias using Entailment Labels), an intermediate pre-training approach for mitigating gender bias in contextualized representations. Key to our approach is the use of a contrastive learning objective on counterfactually augmented, gender-balanced entailment pairs from natural language inference (NLI) datasets. We also introduce an alignment regularizer that pulls identical entailment pairs along opposite gender directions closer. We extensively evaluate our approach on intrinsic and extrinsic metrics, and show that MABEL outperforms previous task-agnostic debiasing approaches in terms of fairness. It also preserves task performance after fine-tuning on downstream tasks. Together, these findings demonstrate the suitability of NLI data as an effective means of bias mitigation, as opposed to only using unlabeled sentences in the literature. Finally, we identify that existing approaches often use evaluation settings that are insufficient or inconsistent. We make an effort to reproduce and compare previous methods, and call for unifying the evaluation settings across gender debiasing methods for better future comparison.

**"Covid vaccine is against Covid but Oxford vaccine is made at Oxford!" Semantic Interpretation of Proper Noun Compounds**

*Keshav Kolluru, Gabriel Stanovsky and Mausam* - 09:00-10:30 (Atrium)
Proper noun compounds, e.g., "Covid vaccine", convey information in a succinct manner (a "Covid vaccine" is a "vaccine that immunizes against the Covid disease"). These are commonly used in short-form domains, such as news headlines, but are largely ignored in information-seeking applications. To address this limitation, we release a new manually annotated dataset, ProNCI, consisting of 22.5K proper noun compounds along with their free-form semantic interpretations. ProNCI is 60 times larger than prior noun compound datasets and also includes non-compositional examples, which have not been previously explored. We experiment with various neural models for automatically generating the semantic interpretations from proper noun compounds, ranging from few-shot prompting to supervised learning, with varying degrees of knowledge about the constituent nouns. We find that adding targeted knowledge, particularly about the common noun, results in performance gains of upto 2.8%. Finally, we integrate our model generated interpretations with an existing Open IE system and observe an 7.5% increase in yield at a precision of 85%. The dataset and code are available at https://github.com/dair-iitd/pronci.

**Pneg: Prompt-based Negative Response Generation for Dialogue Response Selection Task**
*Nyoungwoo Lee, ChaeHun Park, Ho-Jin Choi and Jaegul Choo* 09:00-10:30 (Atrium)
In retrieval-based dialogue systems, a response selection model acts as a ranker to select the most appropriate response among several candidates. However, such selection models tend to rely on context-response content similarity, which makes models vulnerable to adversarial responses that are semantically similar but not relevant to the dialogue context. Recent studies have shown that leveraging these adversarial responses as negative training samples is useful for improving the discriminating power of the selection model. Nevertheless, collecting human-written adversarial responses is expensive, and existing synthesizing methods often have limited scalability. To overcome these limitations, this paper proposes a simple but efficient method for generating adversarial negative responses leveraging a large-scale language model. Experimental results on dialogue selection tasks show that our method outperforms other methods of synthesizing adversarial negative responses. These results suggest that our method can be an effective alternative to human annotators in generating adversarial responses. Our code and dataset will be released if the paper is accepted.

**SLICER: Sliced Fine-Tuning for Low-Resource Cross-Lingual Transfer for Named Entity Recognition**
*Fabian David Schmidt, Ivan Vulić and Goran Glavaš* 09:00-10:30 (Atrium)
Large multilingual language models generally demonstrate impressive results in zero-shot cross-lingual transfer, yet often fail to successfully transfer to low-resource languages, even for token-level prediction tasks like named entity recognition (NER). In this work, we introduce a simple yet highly effective approach for improving zero-shot transfer for NER to low-resource languages. We observe that NER fine-tuning in the source language decontextualizes token representations, i.e., tokens increasingly attend to themselves. This increased reliance on token information itself, we hypothesize, triggers a type of overfitting to properties that NE tokens within the source languages share, but are generally not present in NE mentions of target languages. As a remedy, we propose a simple yet very effective sliced fine-tuning for NER (SLICER) that forces stronger token contextualization in the Transformer: we divide the transformed token representations and classifier into disjoint slices that are then independently classified during training. We evaluate SLICER on two standard benchmarks for NER that involve low-resource languages, WikiANN and MasakhaNER, and show that it (i) indeed reduces decontextualization (i.e., extent to which NE tokens attend to themselves), consequently (ii) yielding consistent transfer gains, especially prominent for low-resource target languages distant from the source language.

**Faithful Knowledge Graph Explanations in Commonsense Question Answering**
*Guy Aglionby and Simone Teufel* 09:00-10:30 (Atrium)
Knowledge graphs are commonly used as sources of information in commonsense question answering, and can also be used to express explanations for the model's answer choice. A common way of incorporating facts from the graph is to encode them separately from the question, and then combine the two representations to select an answer. In this paper, we argue that highly faithful graph-based explanations cannot be extracted from existing models of this type. Such explanations will not include reasoning done by the transformer encoding the question, so will be incomplete. We confirm this theory with a novel proxy measure for faithfulness and propose two architecture changes to address the problem. Our findings suggest a path forward for developing architectures for faithful graph-based explanations.

**KOLD: Korean Offensive Language Dataset**
*Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park and Alice Oh* 09:00-10:30 (Atrium)
Recent directions for offensive language detection are hierarchical modeling, identifying the type and the target of offensive language, and interpretability with offensive span annotation and prediction. These improvements are focused on English and do not transfer well to other languages because of cultural and linguistic differences. In this paper, we present the Korean Offensive Language Dataset (KOLD) comprising 40,429 comments, which are annotated hierarchically with the type and the target of offensive language, accompanied by annotations of the corresponding text spans. We collect the comments from NAVER news and YouTube platform and provide the titles of the articles and videos as the context information for the annotation process. We use these annotated comments as training data for Korean BERT and RoBERTa models and find that they are effective at offensiveness detection, target classification, and target span detection while having room for improvement for target group classification and offensive span detection. We discover that the target group distribution differs drastically from the existing English datasets, and observe that providing the context information improves the model performance in offensiveness detection (+0.3), target classification (+1.5), and target group classification (+13.1). We publicly release the dataset and baseline models.

**ECTSum: A New Benchmark Dataset For Bullet Point Summarization of Long Earnings Call Transcripts**
*Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh and Pawan Goyal* 09:00-10:30 (Atrium)
Despite tremendous progress in automatic summarization, state-of-the-art methods are predominantly trained to excel in summarizing short newswire articles, or documents with strong layout biases such as scientific articles or government reports. Efficient techniques to summarize financial documents, discussing facts and figures, have largely been unexplored, majorly due to the unavailability of suitable datasets. In this work, we present ECTSum, a new dataset with transcripts of earnings calls (ECTs), hosted by publicly traded companies, as documents, and experts-written short telegram-style bullet point summaries derived from corresponding Reuters articles. ECTs are long unstructured documents without any prescribed length limit or format. We benchmark our dataset with state-of-the-art summarization methods across various metrics evaluating the content quality and factual consistency of the generated summaries. Finally, we present a simple yet effective approach, ECT-BPS, to generate a set of bullet points that precisely capture the important facts discussed in the calls.

**IndicXNLI: Evaluating Multilingual Inference for Indian Languages**
*Divyanshu Aggarwal, Vivek Gupta and Anoop Kunchukuttan* 09:00-10:30 (Atrium)
While Indic NLP has made rapid advances recently in terms of the availability of corpora and pre-trained models, benchmark datasets on standard NLU tasks are limited. To this end, we introduce INDICXNLI, an NLI dataset for 11 Indic languages. It has been created by high-quality machine translation of the original English XNLI dataset and our analysis attests to the quality of INDICXNLI. By finetuning different pre-trained LMs on this INDICXNLI, we analyze various cross-lingual transfer techniques with respect to the impact of the choice of language models, languages, multi-linguality, mix-language input, etc. These experiments provide us with useful insights into the behaviour of pre-trained models for a diverse set of languages.

**PreQuEL: Quality Estimation of Machine Translation Outputs in Advance**
*Shachar Don-Yehiya, Leshem Choshen and Omri Abend* 09:00-10:30 (Atrium)
We present the task of PreQuEL, Pre-(Quality-Estimation) Learning. A PreQuEL system predicts how well a given sentence will be translated, without recourse to the actual translation, thus eschewing unnecessary resource allocation when translation quality is bound to be low. PreQuEL can be defined relative to a given MT system (e.g., some industry service) or generally relative to the state-of-the-art. From a theoretical perspective, PreQuEL places the focus on the source text, tracing properties, possibly linguistic features, that make a sentence harder to machine translate.

We develop a baseline model for the task and analyze its performance. We also develop a data augmentation method (from parallel corpora), that improves results substantially. We show that this augmentation method can improve the performance of the Quality-Estimation task as well. We investigate the properties of the input text that our model is sensitive to, by testing it on challenge sets and different languages. We conclude that it is aware of syntactic and semantic distinctions, and correlates and even over-emphasizes the importance of standard NLP features.

**Improving Embeddings Representations for Comparing Higher Education Curricula: A Use Case in Computing**
*Jeffri Murrugarra-Llerena, Fernando Alva-Manchego and Nils Murrugarra-LLerena* 09:00-10:30 (Atrium)
We propose an approach for comparing curricula of study programs in higher education. Pre-trained word embeddings are fine-tuned in a study program classification task, where each curriculum is represented by the names and content of its courses. By combining metric learning with a novel course-guided attention mechanism, our method obtains more accurate curriculum representations than strong baselines. Experiments on a new dataset with curricula of computing programs demonstrate the intuitive power of our approach via attention weights, topic modeling, and embeddings visualizations. We also present a use case comparing computing curricula from USA and Latin America to showcase the capabilities of our improved embeddings representations.

**WeDef: Weakly Supervised Backdoor Defense for Text Classification**
*Lesheng Jin, Zihan Wang and Jingbo Shang* 09:00-10:30 (Atrium)
Existing backdoor defense methods are only effective for limited trigger types. To defend different trigger types at once, we start from the class-irrelevant nature of the poisoning process and propose a novel weakly supervised backdoor defense framework WeDef. Recent advances in weak supervision make it possible to train a reasonably accurate text classifier using only a small number of user-provided, class-indicative seed words. Such seed words shall be considered independent of the triggers. Therefore, a weakly supervised text classifier trained by only the poisoned documents without their labels will likely have no backdoor. Inspired by this observation, in WeDef, we define the reliability of samples based on whether the predictions of the weak classifier agree with their labels in the poisoned training set. We further improve the results through a two-phase sanitization: (1) iteratively refine the weak classifier based on the reliable samples and (2) train a binary poison classifier by distinguishing the most unreliable samples from the most reliable samples. Finally, we train the sanitized model on the samples that the poison classifier predicts as benign. Extensive experiments show that WeDef is effective against popular trigger-based attacks (e.g., words, sentences, and paraphrases), outperforming existing defense methods.

**Pseudo-Relevance for Enhancing Document Representation**
*Jihyuk Kim, Seung-won Hwang, Seoho Song, Hyeseon Ko and Young-In Song* 09:00-10:30 (Atrium)
This paper studies how to enhance the document representation for the bi-encoder approach in dense document retrieval. The bi-encoder, separately encoding a query and a document as a single vector, is favored for high efficiency in large-scale information retrieval, compared to more effective but complex architectures. To combine the strength of the two, the multi-vector representation of documents for bi-encoder, such as ColBERT preserving all token embeddings, has been widely adopted. Our contribution is to reduce the size of the multi-vector representation, without compromising the effectiveness, supervised by query logs. Our proposed solution decreases the latency and the memory footprint, up to 8- and 3-fold, validated on MSMARCO and real-world search query logs.

**Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings**
*Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp and Georg Rehm* 09:00-10:30 (Atrium)
Learning scientific document representations can be substantially improved through contrastive learning objectives, where the challenge lies in creating positive and negative training samples that encode the desired similarity semantics. Prior work relies on discrete citation relations to generate contrast samples. However, discrete citations enforce a hard cut-off to similarity. This is counter-intuitive to similarity-based learning and ignores that scientific papers can be very similar despite lacking a direct citation - a core problem of finding related research. Instead, we use controlled nearest neighbor sampling over citation graph embeddings for contrastive learning. This control allows us to learn continuous similarity, to sample hard-to-learn negatives and positives, and also to avoid collisions between negative and positive samples by controlling the sampling margin between them. The resulting method SciNCL outperforms the state-of-the-art on the SciDocs benchmark. Furthermore, we demonstrate that it can train (or tune) language models sample-efficiently and that it can be combined with recent training-efficient methods. Perhaps surprisingly, even training a general-domain language model this way outperforms baselines pretrained in-domain.

**SPE: Symmetrical Prompt Enhancement for Fact Probing**
*Yiyuan Li, Tong Che, Yezhen Wang, Zhengbao Jiang, Caiming Xiong and Snigdha Chaturvedi* 09:00-10:30 (Atrium)
Pretrained language models (PLMs) have been shown to accumulate factual knowledge during pretraining (Petroni et al. 2019). Recent works probe PLMs for the extent of this knowledge through prompts either in discrete or continuous forms. However, these methods do not consider symmetry of the task: object prediction and subject prediction. In this work, we propose Symmetrical Prompt Enhancement (SPE), a continuous prompt-based method for factual probing in PLMs that leverages the symmetry of the task by constructing symmetrical prompts for subject and object prediction. Our results on a popular factual probing dataset, LAMA, show significant improvement of SPE over previous probing methods.

**Offer a Different Perspective: Modeling the Belief Alignment of Arguments in Multi-party Debates**
*Suzanna Sia, Kokil Jaidka, Hansin Ahuja, Niyati Chhaya and Kevin Duh* 09:00-10:30 (Atrium)
In contexts where debate and deliberation are the norm, the participants are regularly presented with new information that conflicts with their original beliefs. When required to update their beliefs (belief alignment), they may choose arguments that align with their worldview (confirmation bias). We test this and competing hypotheses in a constraint-based modeling approach to predict the winning arguments in multi-party interactions in the Reddit Change My View and Intelligence Squared debates datasets. We adopt a hierarchical generative Variational Autoencoder as our model and impose structural constraints that reflect competing hypotheses about the nature of argumentation. Our findings suggest that in most settings, predictive models that anticipate winning arguments to be further from the initial argument of the opinion holder are more likely to succeed.

**Generating Literal and Implied Subquestions to Fact-check Complex Claims**
*Jifan Chen, Aniruddh Sriram, Eunsol Choi and Greg Durrett* 09:00-10:30 (Atrium)
Verifying political claims is a challenging task, as politicians can use various tactics to subtly misrepresent the facts for their agenda. Existing

automatic fact-checking systems fall short here, and their predictions like "half-true" are not very useful in isolation, since it is unclear which parts of a claim are true and which are not. In this work, we focus on decomposing a complex claim into a comprehensive set of yes-no subquestions whose answers influence the veracity of the claim. We present CLAIMDECOMP, a dataset of decompositions for over 1000 claims. Given a claim and its verification paragraph written by fact-checkers, our trained annotators write subquestions covering both explicit propositions of the original claim and its implicit facets, such as asking about additional political context that changes our view of the claim's veracity. We study whether state-of-the-art models can generate such subquestions, showing that these models generate reasonable questions to ask, but predicting the comprehensive set of subquestions from the original claim without evidence remains challenging. We further show that these subquestions can help identify relevant evidence to fact-check the full claim and derive the veracity through their answers, suggesting that they can be useful pieces of a fact-checking pipeline.

**Analyzing the Mono- and Cross-Lingual Pretraining Dynamics of Multilingual Language Models**
*Terra Blevins, Hila Gonen and Luke Zettlemoyer* 09:00-10:30 (Atrium)
The emergent cross-lingual transfer seen in multilingual pretrained models has sparked significant interest in studying their behavior. However, because these analyses have focused on fully trained multilingual models, little is known about the dynamics of the multilingual pretraining process. We investigate when these models acquire their in-language and cross-lingual abilities by probing checkpoints taken from throughout XLM-R pretraining, using a suite of linguistic tasks. Our analysis shows that the model achieves high in-language performance early on, with lower-level linguistic skills acquired before more complex ones. In contrast, the point in pretraining when the model learns to transfer cross-lingually differs across language pairs. Interestingly, we also observe that, across many languages and tasks, the final model layer exhibits significant performance degradation over time, while linguistic knowledge propagates to lower layers of the network. Taken together, these insights highlight the complexity of multilingual pretraining and the resulting varied behavior for different languages over time.

**Not to Overfit or Underfit the Source Domains? An Empirical Study of Domain Generalization in Question Answering**
*Md Arafat Sultan, Avi Sil and Radu Florian* 09:00-10:30 (Atrium)
Machine learning models are prone to overfitting their training (source) domains, which is commonly believed to be the reason why they falter in novel target domains. Here we examine the contrasting view that multi-source domain generalization (DG) is first and foremost a problem of mitigating source domain underfitting: models not adequately learning the signal already present in their multi-domain training data. Experiments on a reading comprehension DG benchmark show that as a model learns its source domains better—using familiar methods such as knowledge distillation (KD) from a bigger model—its zero-shot out-of-domain utility improves at an even faster pace. Improved source domain learning also demonstrates superior out-of-domain generalization over three popular existing DG approaches that aim to limit overfitting. Our implementation of KD-based domain generalization is available via PrimeQA at: https://ibm.biz/domain-generalization-with-kd.

**Logical Reasoning with Span-Level Predictions for Interpretable and Robust NLI Models**
*Joe Stacey, Pasquale Minervini, Haim Dubossarsky and Marek Rei* 09:00-10:30 (Atrium)
Current Natural Language Inference (NLI) models achieve impressive results, sometimes outperforming humans when evaluating on in-distribution test sets. However, as these models are known to learn from annotation artefacts and dataset biases, it is unclear to what extent the models are learning the task of NLI instead of learning from shallow heuristics in their training data.
We address this issue by introducing a logical reasoning framework for NLI, creating highly transparent model decisions that are based on logical rules. Unlike prior work, we show that improved interpretability can be achieved without decreasing the predictive accuracy. We almost fully retain performance on SNLI, while also identifying the exact hypothesis spans that are responsible for each model prediction.
Using the e-SNLI human explanations, we verify that our model makes sensible decisions at a span level, despite not using any span labels during training. We can further improve model performance and the span-level decisions by using the e-SNLI explanations during training. Finally, our model is more robust in a reduced data setting. When training with only 1,000 examples, out-of-distribution performance improves on the MNLI matched and mismatched validation sets by 13% and 16% relative to the baseline. Training with fewer observations yields further improvements, both in-distribution and out-of-distribution.

**On Measuring the Intrinsic Few-Shot Hardness of Datasets**
*Xinran Zhao, Shikhar Murty and Christopher Manning* 09:00-10:30 (Atrium)
While advances in pre-training have led to dramatic improvements in few-shot learning of NLP tasks, there is limited understanding of what drives successful few-shot adaptation in datasets. In particular, given a new dataset and a pre-trained model, what properties of the dataset make it few-shot learnable, and are these properties independent of the specific adaptation techniques used? We consider an extensive set of recent few-shot learning methods and show that their performance across a large number of datasets is highly correlated, showing that few-shot hardness may be intrinsic to datasets, for a given pre-trained model. To estimate intrinsic few-shot hardness, we then propose a simple and lightweight metric called Spread that captures the intuition that few-shot learning is made possible by exploiting feature-space invariances between training and test samples. Our metric better accounts for few-shot hardness compared to existing notions of hardness and is 8-100x faster to compute.

**ProsocialDialog: A Prosocial Backbone for Conversational Agents**
*Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi and Maarten Sap* 09:00-10:30 (Atrium)
Most existing dialogue systems fail to respond properly to potentially unsafe user utterances by either ignoring or passively agreeing with them. To address this issue, we introduce ProsocialDialog, the first large-scale multi-turn dialogue dataset to teach conversational agents to respond to problematic content following social norms. Covering diverse unethical, problematic, biased, and toxic situations, ProsocialDialog contains responses that encourage prosocial behavior, grounded in commonsense social rules (i.e., rules-of-thumb, RoTs). Created via a human-AI collaborative framework, ProsocialDialog consists of 58K dialogues, with 331K utterances, 160K unique RoTs, and 497K dialogue safety labels accompanied by free-form rationales.
With this dataset, we introduce a dialogue safety detection module, Canary, capable of generating RoTs given conversational context, and a socially-informed dialogue agent, Prost. Empirical results show that Prost generates more socially acceptable dialogues compared to other state-of-the-art language and dialogue models in both in-domain and out-of-domain settings. Additionally, Canary effectively guides conversational agents and off-the-shelf language models to generate significantly more prosocial responses. Our work highlights the promise and importance of creating and steering conversational AI to be socially responsible.

**Scientific Paper Extractive Summarization Enhanced by Citation Graphs**
*Xiuying Chen, Mingzhe Li, Shen Gao, Rui Yan, Xin Gao and Xiangliang Zhang* 09:00-10:30 (Atrium)
In a citation graph, adjacent paper nodes share related scientific terms and topics. The graph thus conveys unique structure information of document-level relatedness that can be utilized in the paper summarization task, for exploring beyond the intra-document information. In this work, we focus on leveraging citation graphs to improve scientific paper extractive summarization under different settings. We first propose a Multi-granularity Unsupervised Summarization model (MUS) as a simple and low-cost solution to the task. MUS finetunes a pre-trained encoder model on the citation graph by link prediction tasks. Then, the abstract sentences are extracted from the corresponding paper considering multi-granularity information. Preliminary results demonstrate that citation graph is helpful even in a simple unsupervised framework. Motivated by this, we next propose a Graph-based Supervised Summarizationmodel (GSS) to achieve more accurate results on the task when

large-scale labeled data are available. Apart from employing the link prediction as an auxiliary task, GSS introduces a gated sentence encoder and a graph information fusion module to take advantage of the graph information to polish the sentence representation. Experiments on a public benchmark dataset show that MUS and GSS bring substantial improvements over the prior state-of-the-art model.

**Mixture of Attention Heads: Selecting Attention Heads Per Token**
*Xiaofeng Zhang, Yikang Shen, Zeyu Huang, Jie Zhou, Wenge Rong and Zhang Xiong* 09:00-10:30 (Atrium)
Mixture-of-Experts (MoE) networks have been proposed as an efficient way to scale up model capacity and implement conditional computing. However, the study of MoE components mostly focused on the feedforward layer in Transformer architecture. This paper proposes the Mixture of Attention Heads (MoA), a new architecture that combines multi-head attention with the MoE mechanism. MoA includes a set of attention heads that each has its own set of parameters. Given an input, a router dynamically selects a subset of k attention heads per token. This conditional computation schema allows MoA to achieve stronger performance than the standard multi-head attention layer. Furthermore, the sparsely gated MoA can easily scale up the number of attention heads and the number of parameters while preserving computational efficiency. Despite performance improvements, MoA also automatically differentiates heads' utilities, providing a new perspective to discuss the model's interpretability. We conducted experiments on several important tasks, including Machine Translation and Masked Language Modeling. Experiments have shown promising results on several tasks against strong baselines that involve large and very deep models.

**On the Calibration of Massively Multilingual Language Models**
*Kabir Ahuja, Sunayana Sitaram, Sandipan Dandapat and Monojit Choudhury* 09:00-10:30 (Atrium)
Massively Multilingual Language Models (MMLMs) have recently gained popularity due to their surprising effectiveness in cross-lingual transfer. While there has been much work in evaluating these models for their performance on a variety of tasks and languages, little attention has been paid on how well calibrated these models are with respect to the confidence in their predictions. We first investigate the calibration of MMLMs in the zero-shot setting and observe a clear case of miscalibration in low-resource languages or those which are typologically diverse from English. Next, we empirically show that calibration methods like temperature scaling and label smoothing do reasonably well in improving calibration in the zero-shot scenario. We also find that few-shot examples in the language can further help reduce calibration errors, often substantially. Overall, our work contributes towards building more reliable multilingual models by highlighting the issue of their miscalibration, understanding what language and model-specific factors influence it, and pointing out the strategies to improve the same.

**Momentum Contrastive Pre-training for Question Answering**
*Minda Hu, Muzhi Li, Yasheng Wang and Irwin King* 09:00-10:30 (Atrium)
Existing pre-training methods for extractive Question Answering (QA) generate cloze-like queries different from natural questions in syntax structure, which could overfit pre-trained models to simple keyword matching. In order to address this problem, we propose a novel Momentum Contrastive pRe-training fOr queStion anSwering (MCROSS) method for extractive QA. Specifically, MCROSS introduces a momentum contrastive learning framework to align the answer probability between cloze-like and natural query-passage sample pairs. Hence, the pre-trained models can better transfer the knowledge learned in cloze-like samples to answering natural questions. Experimental results on three benchmarking QA datasets show that our method achieves noticeable improvement compared with all baselines in both supervised and zero-shot scenarios.

[INDUSTRY] **Dense Feature Memory Augmented Transformers for COVID-19 Vaccination Search Classification**
*Jai Prakash Gupta, Yi Tay, Chaitanya Kamath, Vinh Tran, Donald Metzler, Shailesh Bavadekar, Mimi Sun and Evgeniy Gabrilovich* 09:00-10:30 (Atrium)
With the devastating outbreak of COVID-19, vaccines are one of the crucial lines of defense against mass infection in this global pandemic. Given the protection they provide, vaccines are becoming mandatory in certain social and professional settings. This paper presents a classification model for detecting COVID-19 vaccination related search queries, a machine learning model that is used to generate search insights for COVID-19 vaccinations. The proposed method combines and leverages advancements from modern state-of-the-art (SOTA) natural language understanding (NLU) techniques such as pretrained Transformers with traditional dense features. We propose a novel approach of considering dense features as memory tokens that the model can attend to. We show that this new modeling approach enables a significant improvement to the Vaccine Search Insights (VSI) task, improving a strong well-established gradient-boosting baseline by relative +15% improvement in F1 score and +14% in precision.

[INDUSTRY] **Deploying Unified BERT Moderation Model for E-Commerce Reviews**
*Ravindra Nayak and Nikesh Garera* 09:00-10:30 (Atrium)
Moderation of user-generated e-commerce content has become crucial due to the large and diverse user base on the platforms. Product reviews and ratings have become an integral part of the shopping experience to build trust among users. Due to the high volume of reviews generated on a vast catalog of products, manual moderation is infeasible, making machine moderation a necessity. In this work, we described our deployed system and models for automated moderation of user-generated content. At the heart of our approach, we outline several rejection reasons for review & rating moderation and explore a unified BERT model to moderate them. We convey the importance of product vertical embeddings for the relevancy of the review for a given product and highlight the advantages of pre-training the BERT models with monolingual data to cope with the domain gap in the absence of huge labelled datasets. We observe a 4.78% F1 increase with less labelled data and a 2.57% increase in F1 score on the review data compared to the publicly available BERT-based models. Our best model In-House-BERT-vertical sends only 5.89% of total reviews to manual moderation and has been deployed in production serving live traffic for millions of users.

[INDUSTRY] **End-to-End Speech to Intent Prediction to improve E-commerce Customer Support Voicebot in Hindi and English**
*Abhinav Goyal, Anupam Singh and Nikesh Lucky Garera* 09:00-10:30 (Atrium)
Automation of on-call customer support relies heavily on accurate and efficient speech-to-intent (S2I) systems. Building such systems using multi-component pipelines can pose various challenges because they require large annotated datasets, have higher latency, and have complex deployment. These pipelines are also prone to compounding errors. To overcome these challenges, we discuss an end-to-end (E2E) S2I model for customer support voicebot task in a bilingual setting. We show how we can solve E2E intent classification by leveraging a pre-trained automatic speech recognition (ASR) model with slight modification and fine-tuning on small annotated datasets. Experimental results show that our best E2E model outperforms a conventional pipeline by a relative 27% on the F1 score.

[INDUSTRY] **Deploying a Retrieval based Response Model for Task Oriented Dialogues**
*Lahari Poddar, György Szarvas, Cheng Wang, Jorge Balazs, Pavel Danchenko and Patrick Ernst* 09:00-10:30 (Atrium)
Task-oriented dialogue systems in industry settings need to have high conversational capability, be easily adaptable to changing situations and conform to business constraints. This paper describes a 3-step procedure to develop a conversational model that satisfies these criteria and can efficiently scale to rank a large set of response candidates. First, we provide a simple algorithm to semi-automatically create a high-coverage template set from historic conversations without any annotation. Second, we propose a neural architecture that encodes the dialogue context and applicable business constraints as profile features for ranking the next turn. Third, we describe a two-stage learning strategy with self-supervised training, followed by supervised fine-tuning on limited data collected through a human-in-the-loop platform. Finally, we describe

offline experiments and present results of deploying our model with human-in-the-loop to converse with live customers online.

### Demo Session 5

09:00-10:30 (Atrium)

[DEMO] **SPEAR : Semi-supervised Data Programming in Python**
*Guttu Sai Abhishek, Harshad Ingole, Parth Laturia, Vineeth Dorna, Ayush Maheshwari, Ganesh Ramakrishnan and Rishabh K. Iyer* 09:00-10:30 (Atrium)
We present SPEAR, an open-source python library for data programming with semi supervision. The package implements several recent data programming approaches including facility to programmatically label and build training data. SPEAR facilitates weak supervision in the form of heuristics (or rules) and association of noisy labels to the training dataset. These noisy labels are aggregated to assign labels to the unlabeled data for downstream tasks. We have implemented several label aggregation approaches that aggregate the noisy labels and then train using the noisily labeled set in a cascaded manner. Our implementation also includes other approaches that jointly aggregate and train the model for text classification tasks. Thus, in our python package, we integrate several cascade and joint data-programming approaches while also providing the facility of data programming by letting the user define labeling functions or rules. The code and tutorial notebooks are available at https://github.com/decile-team/spear. Further, extensive documentation can be found at https://spear-decile.readthedocs.io/. Video tutorials demonstrating the usage of our package are available https://youtube.com/playlist?list=PLW8agt_HvkVnOJoJAqBpaerFb-z-ZlqlP. We also present some real-world use cases of SPEAR.

[DEMO] **Label Sleuth: From Unlabeled Text to a Classifier in a Few Hours**
*Eyal Shnarch, Alon Halfon, Ariel Gera, Marina Danilevsky, Yannis Katsis, Leshem Choshen, Martin Santillan Cooper, Dina Epelboim, Zheng Zhang, Dakuo Wang, Lucy Yip, Liat Ein-Dor, Lena Dankin, Ilya Shnayderman, Ranit Aharonov, Yunyao Li, Naftali Liberman, Philip Levin Slesarev, Gwilym Newton, Shila Ofek-Koifman, Noam Slonim and Yoav Katz* 09:00-10:30 (Atrium)
Text classification can be useful in many real-world scenarios, saving a lot of time for end users. However, building a classifier generally requires coding skills and ML knowledge, which poses a significant barrier for many potential users. To lift this barrier we introduce Label Sleuth, a free open source system for labeling and creating text classifiers. This system is unique for: being a no-code system, making NLP accessible for non-experts; guiding its users throughout the entire labeling process until they obtain their desired classifier, making the process efficient; from cold start to a classifier in a few hours; being open for configuration and extension by developers. By open sourcing Label Sleuth we hope to build a community of users and developers that will widen the utilization of NLP models.

[DEMO] **FALTE: A Toolkit for Fine-grained Annotation for Long Text Evaluation**
*Tanya Goyal, Junyi Jessy Li and Greg Durrett* 09:00-10:30 (Atrium)
A growing swath of NLP research is tackling problems related to generating long text, including tasks such as open-ended story generation, summarization, dialogue, and more. However, we currently lack appropriate tools to evaluate these long outputs of generation models: classic automatic metrics such as ROUGE have been shown to perform poorly, and newer learned metrics do not necessarily work well for all tasks and domains of text. Human rating and error analysis remains a crucial component for any evaluation of long text generation. In this paper, we introduce FALTE, a web-based annotation toolkit designed to address this shortcoming. Our tool allows researchers to collect fine-grained judgments of text quality from crowdworkers using an error taxonomy specific to the downstream task. Using the task interface, annotators can select and assign error labels to text span selections in an incremental paragraph-level annotation workflow. The latter functionality is designed to simplify the document-level task into smaller units and reduce cognitive load on the annotators. Our tool has previously been used to run a large-scale annotation study that evaluates the coherence of long generated summaries, demonstrating its utility.

[DEMO] **ALToolbox: A Set of Tools for Active Learning Annotation of Natural Language Texts**
*Akim Tsvigun, Leonid Sanochkin, Daniil Larionov, Gleb Kuzmin, Artem Vazhentsev, Ivan Lazichny, Nikita Khromov, Danil Kireev, Aleksandr Rubashevskii, Olga O. Shahmatova, Dmitry V. Dylov, Igor Galitskiy and Artem Shelmanov* 09:00-10:30 (Atrium)
We present ALToolbox – an open-source framework for active learning (AL) annotation in natural language processing. Currently, the framework supports text classification, sequence tagging, and seq2seq tasks. Besides state-of-the-art query strategies, ALToolbox provides a set of tools that help to reduce computational overhead and duration of AL iterations and increase annotated data reusability. The framework aims to support data scientists and researchers by providing an easy-to-deploy GUI annotation tool directly in the Jupyter IDE and an extensible benchmark for novel AL methods. We prepare a small demonstration of ALToolbox capabilities available at http://demo.nlpresearch.group. A demo video for ALToolbox is provided at: http://demo-video.nlpresearch.group. The code of the framework is published at https://github.com/AIRI-Institute/al_toolbox under the MIT license.

[DEMO] **A Pipeline for Generating, Annotating and Employing Synthetic Data for Real World Question Answering**
*Matt Maufe, James Ravenscroft, Rob Procter and Maria Liakata* 09:00-10:30 (Atrium)
Question Answering (QA) is a growing area of research, often used to facilitate the extraction of information from within documents. State-of-the-art QA models are usually pre-trained on domain-general corpora like Wikipedia and thus tend to struggle on out-of-domain documents without fine-tuning. We demonstrate that synthetic domain-specific datasets can be generated easily using domain-general models, while still providing significant improvements to QA performance. We present two new tools for this task: A flexible pipeline for validating the synthetic QA data and training down stream models on it, and an online interface to facilitate human annotation of this generated data. Using this interface, crowdworkers labelled 1117 synthetic QA pairs, which we then used to fine-tune downstream models and improve domain-specific QA performance by 8.75 F1.

## Session 12 - 11:00-12:30

### Question Answering 2

11:00-12:30 (Hall A, Room A)

**Learning to Decompose: Hypothetical Question Decomposition Based on Comparable Texts**
*Ben Zhou, Kyle Richardson, Xiaodong Yu and Dan Roth* 11:00-11:15 (Hall A, Room A)

Explicit decomposition modeling, which involves breaking down complex tasks into more straightforward and often more interpretable subtasks, has long been a central theme in developing robust and interpretable NLU systems. However, despite the many datasets and resources built as part of this effort, the majority have small-scale annotations and limited scope, which is insufficient to solve general decomposition tasks. In this paper, we look at large-scale intermediate pre-training of decomposition-based transformers using distant supervision from comparable texts, particularly large-scale parallel news. We show that with such intermediate pre-training, developing robust decomposition-based models for a diverse range of tasks becomes more feasible. For example, on semantic parsing, our model, DecompT5, improves 20% to 30% on two datasets, Overnight and TORQUE, over the baseline language model. We further use DecompT5 to build a novel decomposition-based QA system named DecompEntail, improving over state-of-the-art models, including GPT-3, on both HotpotQA and StrategyQA by 8% and 4%, respectively.

**TaCube: Pre-computing Data Cubes for Answering Numerical-Reasoning Questions over Tabular Data**
*Fan Zhou, Mengkang Hu, Haoyu Dong, zhoujun cheng, Fan Cheng, Shi Han and Dongmei Zhang* 11:15-11:30 (Hall A, Room A)
Existing auto-regressive pre-trained language models (PLMs) like T5 and BART, have been well applied to table question answering by UNIFIEDSKG and TAPEX, respectively, and demonstrated state-of-the-art results on multiple benchmarks. However, auto-regressive PLMs are challenged by recent emerging numerical reasoning datasets, such as TAT-QA, due to the error-prone implicit calculation. In this paper, we present TaCube, to pre-compute aggregation/arithmetic results for the table in advance, so that they are handy and readily available for PLMs to answer numerical reasoning questions. TaCube systematically and comprehensively covers a collection of computational operations over table segments. By simply concatenating TaCube to the input sequence of PLMs, it shows significant experimental effectiveness. TaCube promotes the F1 score from 49.6% to 66.2% on TAT-QA and achieves new state-of-the-art results on WikiTQ (59.6% denotation accuracy). TaCube's improvements on numerical reasoning cases are even more notable: on TAT-QA, TaCube promotes the exact match accuracy of BART-large by 39.6% on sum, 52.5% on average, 36.6% on substraction, and 22.2% on division. We believe that TaCube is a general and portable pre-computation solution that can be potentially integrated to various numerical reasoning frameworks

**Rich Knowledge Sources Bring Complex Knowledge Conflicts: Recalibrating Models to Reflect Conflicting Evidence**
*Hung-Ting Chen, Michael Zhang and Eunsol Choi* 11:30-11:45 (Hall A, Room A)
Question answering models can use rich knowledge sources — up to one hundred retrieved passages and parametric knowledge in the large-scale language model (LM). Prior work assumes information in such knowledge sources is consistent with each other, paying little attention to how models blend information stored in their LM parameters with that from retrieved evidence documents. In this paper, we simulate knowledge conflicts (i.e., where parametric knowledge suggests one answer and different passages suggest different answers) and examine model behaviors. We find retrieval performance heavily impacts which sources models rely on, and current models mostly rely on non-parametric knowledge in their best-performing settings. We discover a troubling trend that contradictions among knowledge sources affect model confidence only marginally. To address this issue, we present a new calibration study, where models are discouraged from presenting any single answer when presented with multiple conflicting answer candidates in retrieved evidences.

**QA Domain Adaptation using Hidden Space Augmentation and Self-Supervised Contrastive Adaptation**
*Zhenrui Yue, Huimin Zeng, Bernhard Kratzwald, Stefan Feuerriegel and Dong Wang* 11:45-12:00 (Hall A, Room A)
Question answering (QA) has recently shown impressive results for answering questions from customized domains. Yet, a common challenge is to adapt QA models to an unseen target domain. In this paper, we propose a novel self-supervised framework called QADA for QA domain adaptation. QADA introduces a novel data augmentation pipeline used to augment training QA samples. Different from existing methods, we enrich the samples via hidden space augmentation. For questions, we introduce multi-hop synonyms and sample augmented token embeddings with Dirichlet distributions. For contexts, we develop an augmentation method which learns to drop context spans via a custom attentive sampling strategy. Additionally, contrastive learning is integrated in the proposed self-supervised adaptation framework QADA. Unlike existing approaches, we generate pseudo labels and propose to train the model via a novel attention-based contrastive adaptation method. The attention weights are used to build informative features for discrepancy estimation that helps the QA model separate answers and generalize across source and target domains. To the best of our knowledge, our work is the first to leverage hidden space augmentation and attention-based contrastive adaptation for self-supervised domain adaptation in QA. Our evaluation shows that QADA achieves considerable improvements on multiple target datasets over state-of-the-art baselines in QA domain adaptation.

**Retrieval as Attention: End-to-end Learning of Retrieval and Reading within a Single Transformer**
*Zhengbao Jiang, Luyu Gao, Zhiruo Wang, Jun Araki, Haibo Ding, Jamie Callan and Graham Neubig* 12:00-12:15 (Hall A, Room A)
Systems for knowledge-intensive tasks such as open-domain question answering (QA) usually consist of two stages: efficient retrieval of relevant documents from a large corpus and detailed reading of the selected documents. This is usually done through two separate models, a retriever that encodes the query and finds nearest neighbors, and a reader based on Transformers. These two components are usually modeled separately, which necessitates a cumbersome implementation and is awkward to optimize in an end-to-end fashion. In this paper, we revisit this design and eschew the separate architecture and training in favor of a single Transformer that performs retrieval as attention (RAA), and end-to-end training solely based on supervision from the end QA task. We demonstrate for the first time that an end-to-end trained single Transformer can achieve both competitive retrieval and QA performance on in-domain datasets, matching or even slightly outperforming state-of-the-art dense retrievers and readers. Moreover, end-to-end adaptation of our model significantly boosts its performance on out-of-domain datasets in both supervised and unsupervised settings, making our model a simple and adaptable end-to-end solution for knowledge-intensive tasks.

**Generating Information-Seeking Conversations from Unlabeled Documents**
*Gangwoo Kim, Sungdong Kim, Kang Min Yoo and Jaewoo Kang* 12:15-12:30 (Hall A, Room A)
Synthesizing datasets for conversational question answering (CQA) from unlabeled documents remains challenging due to its interactive nature. Moreover, while modeling information needs is an essential key, only few studies have discussed it. In this paper, we introduce a novel framework, **SimSeek**, (**Sim**ulating information-**Seek**ing conversation from unlabeled documents), and compare its two variants. In our baseline, **SimSeek-sym**, a questioner generates follow-up questions upon the predetermined answer by an answerer. On the contrary, **SimSeek-asym** first generates the question and then finds its corresponding answer under the conversational context. Our experiments show that they can synthesize effective training resources for CQA and conversational search tasks. As a result, conversations from **SimSeek-asym** not only make more improvements in our experiments but also are favorably reviewed in a human evaluation. We finally release a large-scale resource of synthetic conversations, **Wiki-SimSeek**, containing 2 million CQA pairs built upon Wikipedia documents. With the dataset, our CQA model achieves the state-of-the-art performance on a recent CQA benchmark, QuAC. The code and dataset are available at https://github.com/naver-ai/simseek

## Morphology, Syntax, Linguistics, Psycholinguistics & TACL

11:00-12:30 (Hall A, Room B)

**Unsupervised Boundary-Aware Language Model Pretraining for Chinese Sequence Labeling**
*Peijie Jiang, Dingkun Long, Yanzhao Zhang, Pengjun Xie, Meishan Zhang and Min Zhang* 11:00-11:15 (Hall A, Room B)
Boundary information is critical for various Chinese language processing tasks, such as word segmentation, part-of-speech tagging, and named entity recognition. Previous studies usually resorted to the use of a high-quality external lexicon, where lexicon items can offer explicit boundary information. However, to ensure the quality of the lexicon, great human effort is always necessary, which has been generally ignored. In this work, we suggest unsupervised statistical boundary information instead, and propose an architecture to encode the information directly into pre-trained language models, resulting in Boundary-Aware BERT (BABERT). We apply BABERT for feature induction of Chinese sequence labeling tasks. Experimental results on ten benchmarks of Chinese sequence labeling demonstrate that BABERT can provide consistent improvements on all datasets. In addition, our method can complement previous supervised lexicon exploration, where further improvements can be achieved when integrated with external lexicon information.

**SynGEC: Syntax-Enhanced Grammatical Error Correction with a Tailored GEC-Oriented Parser**
*Yue Zhang, Bo Zhang, Zhenghua Li, Zuyi Bao, Chen Li and Min Zhang* 11:15-11:30 (Hall A, Room B)
This work proposes a syntax-enhanced grammatical error correction (GEC) approach named SynGEC that effectively incorporates dependency syntactic information into the encoder part of GEC models. The key challenge for this idea is that off-the-shelf parsers are unreliable when processing ungrammatical sentences. To confront this challenge, we propose to build a tailored GEC-oriented parser (GOPar) using parallel GEC training data as a pivot. First, we design an extended syntax representation scheme that allows us to represent both grammatical errors and syntax in a unified tree structure. Then, we obtain parse trees of the source incorrect sentences by projecting trees of the target correct sentences. Finally, we train GOPar with such projected trees. For GEC, we employ the graph convolution network to encode source-side syntactic information produced by GOPar, and fuse them with the outputs of the Transformer encoder. Experiments on mainstream English and Chinese GEC datasets show that our proposed SynGEC approach consistently and substantially outperforms strong baselines and achieves competitive performance. Our code and data are all publicly available at https://github.com/HillZhang1999/SynGEC.

**Unbiased and Efficient Sampling of Dependency Trees**
*Miloš Stanojević* 11:30-11:45 (Hall A, Room B)
Most computational models of dependency syntax consist of distributions over spanning trees. However, the majority of dependency treebanks require that every valid dependency tree has a single edge coming out of the ROOT node, a constraint that is not part of the definition of spanning trees. For this reason all standard inference algorithms for spanning trees are sub-optimal for inference over dependency trees. Zmigrod et al (2021) proposed algorithms for sampling with and without replacement from the dependency tree distribution that incorporate the single-root constraint. In this paper we show that their fastest algorithm for sampling with replacement, Wilson-RC, is in fact producing biased samples and we provide two alternatives that are unbiased. Additionally, we propose two algorithms (one incremental, one parallel) that reduce the asymptotic runtime of algorithm for sampling k trees without replacement to $O(kn^3)$. These algorithms are both asymptotically and practically more efficient.

**A Comprehensive Comparison of Neural Networks as Cognitive Models of Inflection**
*Adam Wiemerslage, Shiran Dudy and Katharina Kann* 11:45-12:00 (Hall A, Room B)
Neural networks have long been at the center of a debate around the cognitive mechanism by which humans process inflectional morphology. This debate has gravitated into NLP by way of the question: Are neural networks a feasible account for human behavior in morphological inflection? We address that question by measuring the correlation between human judgments and neural network probabilities for unknown word inflections. We test a larger range of architectures than previously studied on two important tasks for the cognitive processing debate: English past tense, and German number inflection. We find evidence that the Transformer may be a better account of human behavior than LSTMs on these datasets, and that LSTM features known to increase inflection accuracy do not always result in more human-like behavior.

[TACL] **Morphology Without Borders: Clause-Level Morphology**
*Omer Goldman and Reut Tsarfaty* 12:00-12:15 (Hall A, Room B)
Morphological tasks use large multi-lingual datasets that organize words into inflection tables, which then serve as training and evaluation data for various tasks. However, a closer inspection of these data reveals profound cross-linguistic inconsistencies, that arise from the lack of a clear linguistic and operational definition of what is a word, and that severely impair the universality of the derived tasks. To overcome this deficiency, we propose to view morphology as a clauselevel phenomenon, rather than word-level. It is anchored in a fixed yet inclusive set of features, that encapsulates all functions realized in a saturated clause. We deliver MIGHTYMORPH, a novel dataset for clause-level morphology covering 4 typologically-different languages: English, German, Turkish and Hebrew. We use this dataset to derive 3 clause-level morphological tasks: inflection, reinflection and analysis. Our experiments show that the clause-level tasks are substantially harder than the respective word-level tasks, while having comparable complexity across languages. Furthermore, redefining morphology to the clause-level provides a neat interface with contextualized language models (LMs) and allows assessing the morphological knowledge encoded in these models and their usability for morphological tasks. Taken together, this work opens up new horizons in the study of computational morphology, leaving ample space for studying neural morphology cross-linguistically.

[TACL] **Transformer Grammars: Augmenting Transformer Language Models with Syntactic Inductive Biases at Scale**
*Laurent Sartran, Samuel Barrett, Adhiguna Kuncoro, Miloš Stanojević, Phil Blunsom and Chris Dyer* 12:15-12:30 (Hall A, Room B)
We introduce Transformer Grammars – a novel class of Transformer language models that combine: (i) the expressive power, scalability, and strong performance of Transformer language models, and (ii) recursive syntactic compositions, which here are implemented through a special attention mask. We find that Transformer Grammars outperform various strong baselines on multiple syntax-sensitive language modeling evaluation metrics, in addition to sentence-level language modeling perplexity. Nevertheless, we find that the recursive syntactic composition bottleneck harms perplexity on document-level language modeling, providing evidence that a different kind of memory mechanism – that works independently of syntactic structures – plays an important role in the processing of long-form text.

## Dialog and Interactive Systems 2

11:00-12:30 (Hall A, Room C)

**MetaASSIST: Robust Dialogue State Tracking with Meta Learning**
*Fanghua Ye, Xi Wang, Jie Huang, Shenghui Li, Samuel Stern and Emine Yilmaz* 11:00-11:15 (Hall A, Room C)
Existing dialogue datasets contain lots of noise in their state annotations. Such noise can hurt model training and ultimately lead to poor gener-

alization performance. A general framework named ASSIST has recently been proposed to train robust dialogue state tracking (DST) models. It introduces an auxiliary model to generate pseudo labels for the noisy training set. These pseudo labels are combined with vanilla labels by a common fixed weighting parameter to train the primary DST model. Notwithstanding the improvements of ASSIST on DST, tuning the weighting parameter is challenging. Moreover, a single parameter shared by all slots and all instances may be suboptimal. To overcome these limitations, we propose a meta learning-based framework MetaASSIST to adaptively learn the weighting parameter. Specifically, we propose three schemes with varying degrees of flexibility, ranging from slot-wise to both slot-wise and instance-wise, to convert the weighting parameter into learnable functions. These functions are trained in a meta-learning manner by taking the validation set as meta data. Experimental results demonstrate that all three schemes can achieve competitive performance. Most impressively, we achieve a state-of-the-art joint goal accuracy of 80.10% on MultiWOZ 2.4.

**Watch the Neighbors: A Unified K-Nearest Neighbor Contrastive Learning Framework for OOD Intent Discovery**
*Yutao Mou, Keqing He, Pei Wang, Yanan Wu, Jingang Wang, Wei Wu and Weiran Xu* 11:15-11:30 (Hall A, Room C)
Discovering out-of-domain (OOD) intent is important for developing new skills in task-oriented dialogue systems. The key challenges lie in how to transfer prior in-domain (IND) knowledge to OOD clustering, as well as jointly learn OOD representations and cluster assignments. Previous methods suffer from in-domain overfitting problem, and there is a natural gap between representation learning and clustering objectives. In this paper, we propose a unified K-nearest neighbor contrastive learning framework to discover OOD intents. Specifically, for IND pre-training stage, we propose a KCL objective to learn inter-class discriminative features, while maintaining intra-class diversity, which alleviates the in-domain overfitting problem. For OOD clustering stage, we propose a KCC method to form compact clusters by mining true hard negative samples, which bridges the gap between clustering and representation learning. Extensive experiments on three benchmark datasets show that our method achieves substantial improvements over the state-of-the-art methods.

**Counterfactual Data Augmentation via Perspective Transition for Open-Domain Dialogues**
*Jiao Ou, Jinchao Zhang, Yang Feng and Jie Zhou* 11:30-11:45 (Hall A, Room C)
The construction of open-domain dialogue systems requires high-quality dialogue datasets. The dialogue data admits a wide variety of responses for a given dialogue history, especially responses with different semantics. However, collecting high-quality such a dataset in most scenarios is labor-intensive and time-consuming. In this paper, we propose a data augmentation method to automatically augment high-quality responses with different semantics by counterfactual inference. Specifically, given an observed dialogue, our counterfactual generation model first infers semantically different responses by replacing the observed reply perspective with substituted ones. Furthermore, our data selection method filters out detrimental augmented responses. Experimental results show that our data augmentation method can augment high-quality responses with different semantics for a given dialogue history, and can outperform competitive baselines on multiple downstream tasks.

**There Is No Standard Answer: Knowledge-Grounded Dialogue Generation with Adversarial Activated Multi-Reference Learning**
*Xueliang Zhao, Tingchen Fu, Chongyang Tao and Rui Yan* 11:45-12:00 (Hall A, Room C)
Knowledge-grounded dialogue (KGC) shows excellent potential to deliver an engaging and informative response. However, existing approaches emphasize selecting one golden knowledge given a particular dialogue context, overlooking the one-to-many phenomenon in dialogue. As a result, existing paradigm limits the diversity of knowledge selection and generation. To this end, we establish a multi-reference KGC dataset and propose a series of metrics to systematically assess the one-to-many efficacy of existing KGC models. Furthermore, to extend the hypothesis space of knowledge selection to enhance the mapping relationship between multiple knowledge and multiple responses, we devise a span-based variational model and optimize the model in a wake-sleep style with an ameliorated evidence lower bound objective to learn the one-to-many generalization. Both automatic and human evaluations demonstrate the efficacy of our approach.

**D4: a Chinese Dialogue Dataset for Depression-Diagnosis-Oriented Chat**
*Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Mengyue Wu, Lu Chen, Zhen Wang and Kai Yu* 12:00-12:15 (Hall A, Room C)
In a depression-diagnosis-directed clinical session, doctors initiate a conversation with ample emotional support that guides the patients to expose their symptoms based on clinical diagnosis criteria. Such a dialogue system is distinguished from existing single-purpose human-machine dialog systems, as it combines task-oriented and chit-chats with uniqueness in dialogue topics and procedures. However, due to the social stigma associated with mental illness, the dialogue data related to depression consultation and diagnosis are rarely disclosed. Based on clinical depression diagnostic criteria ICD-11 and DSM-5, we designed a 3-phase procedure to construct $D^4$: a Chinese Dialogue Dataset for Depression-Diagnosis-Oriented Chat, which simulates the dialogue between doctors and patients during the diagnosis of depression, including diagnosis results and symptom summary given by professional psychiatrists for each conversation. Upon the newly-constructed dataset, four tasks mirroring the depression diagnosis process are established: response generation, topic prediction, dialog summary, and severity classification of depressive episode and suicide risk. Multi-scale evaluation results demonstrate that a more empathy-driven and diagnostic-accurate consultation dialogue system trained on our dataset can be achieved compared to rule-based bots.

**Navigating Connected Memories with a Task-oriented Dialog System**
*Satwik Kottur, Seungwhan Moon, Alborz Geramifard and Babak Damavandi* 12:15-12:30 (Hall A, Room C)
Recent years have seen an increasing trend in the volume of personal media captured by users, thanks to the advent of smartphones and smart glasses, resulting in large media collections. Despite conversation being an intuitive human-computer interface, current efforts focus mostly on single-shot natural language based media retrieval to aid users query their media and re-live their memories. This severely limits the search functionality as users can neither ask follow-up queries nor obtain information without first formulating a single-turn query.
In this work, we propose dialogs for connected memories as a powerful tool to empower users to search their media collection through a multi-turn, interactive conversation. Towards this, we collect a new task-oriented dialog dataset COMET, which contains 11.5k user-assistant dialogs (totalling 103k utterances), grounded in simulated personal memory graphs. We employ a resource-efficient, two-phase data collection pipeline that uses: (1) a novel multimodal dialog simulator that generates synthetic dialog flows grounded in memory graphs, and, (2) manual paraphrasing to obtain natural language utterances. We analyze COMET, formulate four main tasks to benchmark meaningful progress, and adopt state-of-the-art language models as strong baselines, in order to highlight the multimodal challenges captured by our dataset.

## Speech, Vision, Robotics, Multimodal Grounding 2 & TACL

11:00-12:30 (Hall A, Room D)

**SpeechUT: Bridging Speech and Text with Hidden-Unit for Encoder-Decoder Based Speech-Text Pre-training**
*Ziqiang Zhang, Long Zhou, Junyi Ao, Shujie Liu, Lirong Dai, Jinyu Li and Furu Wei* 11:00-11:15 (Hall A, Room D)
The rapid development of single-modal pre-training has prompted researchers to pay more attention to cross-modal pre-training methods. In this paper, we propose a unified-modal speech-unit-text pre-training model, SpeechUT, to connect the representations of a speech encoder and a text decoder with a shared unit encoder. Leveraging hidden-unit as an interface to align speech and text, we can decompose the

speech-to-text model into a speech-to-unit model and a unit-to-text model, which can be jointly pre-trained with unpaired speech and text data respectively. Our proposed SpeechUT is fine-tuned and evaluated on automatic speech recognition (ASR) and speech translation (ST) tasks. Experimental results show that SpeechUT gets substantial improvements over strong baselines, and achieves state-of-the-art performance on both the LibriSpeech ASR and MuST-C ST tasks. To better understand the proposed SpeechUT, detailed analyses are conducted. The code and pre-trained models are available at https://aka.ms/SpeechUT.

**Can Visual Context Improve Automatic Speech Recognition for an Embodied Agent?**
*Pradip Pramanick and Chayan Sarkar* 11:15-11:30 (Hall A, Room D)
The usage of automatic speech recognition (ASR) systems are becoming omnipresent ranging from personal assistant to chatbots, home, and industrial automation systems, etc. Modern robots are also equipped with ASR capabilities for interacting with humans as speech is the most natural interaction modality. However, ASR in robots faces additional challenges as compared to a personal assistant. Being an embodied agent, a robot must recognize the physical entities around it and therefore reliably recognize the speech containing the description of such entities. However, current ASR systems are often unable to do so due to limitations in ASR training, such as generic datasets and open-vocabulary modeling. Also, adverse conditions during inference, such as noise, accented, and far-field speech makes the transcription inaccurate. In this work, we present a method to incorporate a robot's visual information into an ASR system and improve the recognition of a spoken utterance containing a visible entity. Specifically, we propose a new decoder biasing technique to incorporate the visual context while ensuring the ASR output does not degrade for incorrect context. We achieve a 59% relative reduction in WER from an unmodified ASR system.

**Why is Winoground Hard? Investigating Failures in Visuolinguistic Compositionality**
*Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath and Kyle Mahowald* 11:30-11:45 (Hall A, Room D)
Recent visuolinguistic pre-trained models show promising progress on various end tasks such as image retrieval and video captioning. Yet, they fail miserably on the recently proposed Winoground dataset, which challenges models to match paired images and English captions, with items constructed to overlap lexically but differ in meaning (e.g., "there is a mug in some grass" vs. "there is some grass in a mug"). By annotating the dataset using new fine-grained tags, we show that solving the Winoground task requires not just compositional language understanding, but a host of other abilities like commonsense reasoning or locating small, out-of-focus objects in low-resolution images. In this paper, we identify the dataset's main challenges through a suite of experiments on related tasks (probing task, image retrieval task), data augmentation, and manual inspection of the dataset. Our analysis suggests that a main challenge in visuolinguistic models may lie in fusing visual and textual representations, rather than in compositional language understanding. We release our annotation and code at https://github.com/ajd12342/why-winoground-hard.

**Do Vision-and-Language Transformers Learn Grounded Predicate-Noun Dependencies?**
*Mitja Nikolaus, Emmanuelle Salin, Stephane Ayache, Abdellah Fourtassi and Benoit Favre* 11:45-12:00 (Hall A, Room D)
Recent advances in vision-and-language modeling have seen the development of Transformer architectures that achieve remarkable performance on multimodal reasoning tasks. Yet, the exact capabilities of these black-box models are still poorly understood. While much of previous work has focused on studying their ability to learn meaning at the word-level, their ability to track syntactic dependencies between words has received less attention. We take a first step in closing this gap by creating a new multimodal task targeted at evaluating understanding of predicate-noun dependencies in a controlled setup. We evaluate a range of state-of-the-art models and find that their performance on the task varies considerably, with some models performing relatively well and others at chance level. In an effort to explain this variability, our analyses indicate that the quality (and not only sheer quantity) of pretraining data is essential. Additionally, the best performing models leverage fine-grained multimodal pretraining objectives in addition to the standard image-text matching objectives. This study highlights that targeted and controlled evaluations are a crucial step for a precise and rigorous test of the multimodal knowledge of vision-and-language models.

[TACL] **Learning English with Peppa Pig**
*Mitja Nikolaus, Afra Alishahi and Grzegorz Chrupała* 12:00-12:15 (Hall A, Room D)
Recent computational models of the acquisition of spoken language via grounding in perception exploit associations between the spoken and visual modalities and learn to represent speech and visual data in a joint vector space. A major unresolved issue from the point of ecological validity is the training data, typically consisting of images or videos paired with spoken descriptions of what is depicted. Such a setup guarantees an unrealistically strong correlation between speech and the visual data. In the real world the coupling between the linguistic and the visual modality is loose, and often confounded by correlations with non-semantic aspects of the speech signal. Here we address this shortcoming by using a dataset based on the children's cartoon Peppa Pig . We train a simple bi-modal architecture on the portion of the data consisting of dialog between characters, and evaluate on segments containing descriptive narrations. Despite the weak and confounded signal in this training data our model succeeds at learning aspects of the visual semantics of spoken language.

**Learning a Grammar Inducer from Massive Uncurated Instructional Videos**
*Songyang Zhang, Linfeng Song, Lifeng Jin, Haitao Mi, Kun Xu, Dong Yu and Jiebo Luo* 12:15-12:30 (Hall A, Room D)
Video-aided grammar induction aims to leverage video information for finding more accurate syntactic grammars for accompanying text. While previous work focuses on building systems for inducing grammars on text that are well-aligned with video content, we investigate the scenario, in which text and video are only in loose correspondence. Such data can be found in abundance online, and the weak correspondence is similar to the indeterminacy problem studied in language acquisition. Furthermore, we build a new model that can better learn video-span correlation without manually designed features adopted by previous work. Experiments show that our model trained only on large-scale YouTube data with no text-video alignment reports strong and robust performances across three unseen datasets, despite domain shift and noisy label issues. Furthermore our model yields higher F1 scores than the previous state-of-the-art systems trained on in-domain data.

## Information Extraction 2

11:00-12:30 (Hall B)

**Entity Extraction in Low Resource Domains with Selective Pre-training of Large Language Models**
*Aniruddha Mahapatra, Sharmila Reddy Nangi, Aparna Garimella and Anandhavelu N* 11:00-11:15 (Hall B)
Transformer-based language models trained on large natural language corpora have been very useful in downstream entity extraction tasks. However, they often result in poor performances when applied to domains that are different from those they are pretrained on. Continued pretraining using unlabeled data from target domains can help improve the performances of these language models on the downstream tasks. However, using all of the available unlabeled data for pretraining can be time-intensive; also, it can be detrimental to the performance of the downstream tasks, if the unlabeled data is not aligned with the data distribution for the target tasks. Previous works employed external super-

vision in the form of ontologies for selecting appropriate data samples for pretraining, but external supervision can be quite hard to obtain in low-resource domains. In this paper, we introduce effective ways to select data from unlabeled corpora of target domains for language model pretraining to improve the performances in target entity extraction tasks. Our data selection strategies do not require any external supervision. We conduct extensive experiments for the task of named entity recognition (NER) on seven different domains and show that language models pretrained on target domain unlabeled data obtained using our data selection strategies achieve better performances compared to those using data selection strategies in previous works that use external supervision. We also show that these pretrained language models using our data selection strategies outperform those pretrained on all of the available unlabeled target domain data.

**Multilingual Relation Classification via Efficient and Effective Prompting**
*Yuxuan Chen, David Harbecke and Leonhard Hennig* 11:15-11:30 (Hall B)
Prompting pre-trained language models has achieved impressive performance on various NLP tasks, especially in low data regimes. Despite the success of prompting in monolingual settings, applying prompt-based methods in multilingual scenarios has been limited to a narrow set of tasks, due to the high cost of handcrafting multilingual prompts. In this paper, we present the first work on prompt-based multilingual relation classification (RC), by introducing an efficient and effective method that constructs prompts from relation triples and involves only minimal translation for the class labels. We evaluate its performance in fully supervised, few-shot and zero-shot scenarios, and analyze its effectiveness across 14 languages, prompt variants, and English-task training in cross-lingual settings. We find that in both fully supervised and few-shot scenarios, our prompt method beats competitive baselines: fine-tuning XLM-R_EM and null prompts. It also outperforms the random baseline by a large margin in zero-shot experiments. Our method requires little in-language knowledge and can be used as a strong baseline for similar multilingual classification tasks.

**Fine-grained Contrastive Learning for Relation Extraction**
*William Hogan, Jiacheng Li and Jingbo Shang* 11:30-11:45 (Hall B)
Recent relation extraction (RE) works have shown encouraging improvements by conducting contrastive learning on silver labels generated by distant supervision before fine-tuning on gold labels. Existing methods typically assume all these silver labels are accurate and treat them equally; however, distant supervision is inevitably noisy–some silver labels are more reliable than others. In this paper, we propose fine-grained contrastive learning (FineCL) for RE, which leverages fine-grained information about which silver labels are and are not noisy to improve the quality of learned relationship representations for RE. We first assess the quality of silver labels via a simple and automatic approach we call "learning order denoising," where we train a language model to learn these relations and record the order of learned training instances. We show that learning order largely corresponds to label accuracy–early-learned silver labels have, on average, more accurate labels than later-learned silver labels. Then, during pre-training, we increase the weights of accurate labels within a novel contrastive learning objective. Experiments on several RE benchmarks show that FineCL makes consistent and significant performance gains over state-of-the-art methods.

**SQUIRE: A Sequence-to-sequence Framework for Multi-hop Knowledge Graph Reasoning**
*Yushi Bai, Xin Lv, Juanzi Li, Lei Hou, Yincen Qu, zelin Dai and Feiyu Xiong* 11:45-12:00 (Hall B)
Multi-hop knowledge graph (KG) reasoning has been widely studied in recent years to provide interpretable predictions on missing links with evidential paths. Most previous works use reinforcement learning (RL) based methods that learn to navigate the path towards the target entity. However, these methods suffer from slow and poor convergence, and they may fail to infer a certain path when there is a missing edge along the path. Here we present SQUIRE, the first Sequence-to-sequence based multi-hop reasoning framework, which utilizes an encoder-decoder Transformer structure to translate the query to a path. Our framework brings about two benefits: (1) It can learn and predict in an end-to-end fashion, which gives better and faster convergence; (2) Our transformer model does not rely on existing edges to generate the path, and has the flexibility to complete missing edges along the path, especially in sparse KGs. Experiments on standard and sparse KGs show that our approach yields significant improvement over prior methods, while converging 4x-7x faster.

**Style Transfer as Data Augmentation: A Case Study on Named Entity Recognition**
*Shuguang Chen, Leonardo Neves and Thamar Solorio* 12:00-12:15 (Hall B)
In this work, we take the named entity recognition task in the English language as a case study and explore style transfer as a data augmentation method to increase the size and diversity of training data in low-resource scenarios. We propose a new method to effectively transform the text from a high-resource domain to a low-resource domain by changing its style-related attributes to generate synthetic data for training. Moreover, we design a constrained decoding algorithm along with a set of key ingredients for data selection to guarantee the generation of valid and coherent data. Experiments and analysis on five different domain pairs under different data regimes demonstrate that our approach can significantly improve results compared to current state-of-the-art data augmentation methods. Our approach is a practical solution to data scarcity, and we expect it to be applicable to other NLP tasks.

**Rescue Implicit and Long-tail Cases: Nearest Neighbor Relation Extraction**
*Zhen Wan, Qianying Liu, Zhuoyuan Mao, Fei Cheng, Sadao Kurohashi and Jiwei Li* 12:15-12:30 (Hall B)
Relation extraction (RE) has achieved remarkable progress with the help of pre-trained language models. However, existing RE models are usually incapable of handling two situations: implicit expressions and long-tail relation types, caused by language complexity and data sparsity. In this paper, we introduce a simple enhancement of RE using $k$ nearest neighbors ($k$NN-RE). $k$NN-RE allows the model to consult training relations at test time through a nearest-neighbor search and provides a simple yet effective means to tackle the two issues above. Additionally, we observe that $k$NN-RE serves as an effective way to leverage distant supervision (DS) data for RE. Experimental results show that the proposed $k$NN-RE achieves state-of-the-art performances on a variety of supervised RE datasets, i.e., ACE05, SciERC, and Wiki80, along with outperforming the best model to date on the i2b2 and Wiki80 datasets in the setting of allowing using DS. Our code and models are available at: https://github.com/YukinoWan/kNN-RE.

## CL & TACL 3

11:00-12:30 (Collaboratorium)

---

[TACL] **Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond**
*Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch and Diyi Yang* 11:00-11:15 (Collaboratorium)
A fundamental goal of scientific research is to learn about causal relationships. However, despite its critical role in the life and social sciences, causality has not had the same importance in Natural Language Processing (NLP), which has traditionally placed more emphasis on predictive tasks. This distinction is beginning to fade, with an emerging area of interdisciplinary research at the convergence of causal inference and language processing. Still, research on causality in NLP remains scattered across domains without unified definitions, benchmark datasets and

clear articulations of the challenges and opportunities in the application of causal inference to the textual domain, with its unique properties. In this survey, we consolidate research across academic areas and situate it in the broader NLP landscape. We introduce the statistical challenge of estimating causal effects with text, encompassing settings where text is used as an outcome, treatment, or to address confounding. In addition, we explore potential uses of causal inference to improve the robustness, fairness, and interpretability of NLP models. We thus provide a unified overview of causal inference for the NLP community.

[TACL] **Multi-task Active Learning for Pre-trained Transformer-based Models**
*Guy Rotman and Roi Reichart* 11:15-11:30 (Collaboratorium)
Multi-task learning, in which several tasks are jointly learned by a single model, allows NLP models to share information from multiple annotations and may facilitate better predictions when the tasks are inter-related. This technique, however, requires annotating the same text with multiple annotation schemes which may be costly and laborious. Active learning (AL) has been demonstrated to optimize annotation processes by iteratively selecting unlabeled examples whose annotation is most valuable for the NLP model. Yet, multi-task active learning (MT-AL) has not been applied to state-of-the-art pre-trained Transformer-based NLP models. This paper aims to close this gap. We explore various multi-task selection criteria in three realistic multi-task scenarios, reflecting different relations between the participating tasks, and demonstrate the effectiveness of multi-task compared to single-task selection. Our results suggest that MT-AL can be effectively used in order to minimize annotation efforts for multi-task NLP models.

[TACL] **Saturated Transformers are Constant-Depth Threshold Circuits**
*William Merrill, Ashish Sabharwal and Noah A. Smith* 11:30-11:45 (Collaboratorium)
Transformers have become a standard neural network architecture for many NLP problems, motivating theoretical analysis of their power in terms of formal languages. Recent work has shown that transformers with hard attention are quite limited in power (Hahn, 2020), as they can be simulated by constant-depth AND/OR circuits (Hao et al., 2022). However, hard attention is a strong assumption, which may complicate the relevance of these results in practice. In this work, we analyze the circuit complexity of transformers with saturated attention: a generalization of hard attention that more closely captures the attention patterns learnable in practical transformers. We first show that saturated transformers transcend the known limitations of hard-attention transformers. We then prove saturated transformers with floating-point values can be simulated by constant-depth threshold circuits, giving the class TC0 as an upper bound on the formal languages they recognize.

[TACL] **Unit Tests for Concepts in Neural Networks**
*Charles Lovering and Ellie Pavlick* 11:45-12:00 (Collaboratorium)
Many complex problems are naturally understood in terms of symbolic concepts. For example, our concept of "cat" is related to our concepts of "ears" and "whiskers" in a non-arbitrary way. Fodor (1998) proposes one theory of concepts, which emphasizes symbolic representations related via constituency structures. Whether neural networks are consistent with such a theory is open for debate. We propose unit tests for evaluating whether a system's behavior is consistent with several key aspects of Fodor's criteria. Using a simple visual concept learning task, we evaluate several modern neural architectures against this specification. We find that models succeed on tests of groundedness, modularity, and reusability of concepts, but that important questions about causality remain open. Resolving these will require new methods for analyzing models' internal states.

[TACL] **On the Role of Negative Precedent in Legal Outcome Prediction**
*Josef Valvoda, Ryan Cotterell and Simone Teufel* 12:00-12:15 (Collaboratorium)
Every legal case sets a precedent by developing the law in one of the following two ways. It either expands its scope, in which case it sets positive precedent, or it narrows it down, in which case it sets negative precedent. While legal outcome prediction, which is nothing other than the prediction of positive precedents, is an increasingly popular task in AI, we are the first to investigate negative precedent prediction by focusing on negative outcomes. We discover an asymmetry in existing models' ability to predict positive and negative outcomes. Where state-of-the-art outcome prediction models predicts positive outcomes at $75.06$ $F\_1$, they predicts negative outcomes at only $10.09$ $F\_1$, worse than a random baseline. To address this performance gap, we develop two new models inspired by the dynamics of a court process. Our first model significantly improves positive outcome prediction score to $77.15$ $F\_1$ and our second model more than doubles the negative outcome prediction performance to $24.01$ $F\_1$. Despite this improvement, shifting focus to negative outcomes reveals that there is still plenty of room to grow when it comes to modelling law.

[TACL] **Typical Decoding for Natural Language Generation**
*Clara Meister, Tiago Pimentel, Gian Wiher and Ryan Cotterell* 12:15-12:30 (Collaboratorium)
Despite achieving incredibly low perplexities on myriad natural language corpora, today's language models still often underperform when used to generate text. This dichotomy has puzzled the language generation community for the last few years. In this work, we posit that the abstraction of natural language as a communication channel (a la Shannon, 1948) can provide new insights into the behaviors of probabilistic language generators, e.g., why high-probability texts can be dull or repetitive. Humans use language as a means of communicating information, and do so in a simultaneously efficient and error-minimizing manner; they choose each word in a string with this (perhaps subconscious) goal in mind. We propose that generation from probabilistic models should mimic this behavior. Rather than always choosing words from the high-probability region of the distribution—which have a low Shannon information content—we sample from the set of words with information content close to the conditional entropy of our model, i.e., close to the expected information content. This decision criterion can be realized through a simple and efficient implementation, which we call typical sampling. Automatic and human evaluations show that, in comparison to nucleus and top-k sampling, typical sampling offers competitive performance in terms of quality while consistently reducing the number of degenerate repetitions.

## Poster Sessions 15 & 16

11:00-12:30 (Atrium)

---

**Continued Pretraining for Better Zero- and Few-Shot Promptability**
*Zhaofeng Wu, Robert L Logan IV, Pete Walsh, Akshita Bhagia, Dirk Groeneveld, Sameer Singh and Iz Beltagy* 11:00-12:30 (Atrium)
Recently introduced language model prompting methods can achieve high accuracy in zero- and few-shot settings while requiring few to no learned task-specific parameters. Nevertheless, these methods still often trail behind full model finetuning. In this work, we investigate if a dedicated continued pretraining stage could improve "promptability", i.e., zero-shot performance with natural language prompts or few-shot performance with prompt tuning. We reveal settings where existing continued pretraining methods lack promptability. We also identify current methodological gaps, which we fill with thorough large-scale experiments. We demonstrate that a simple recipe, continued pretraining that incorporates a trainable prompt during multi-task learning, leads to improved promptability in both zero- and few-shot settings compared to existing methods, up to 31% relative. On the other hand, we find that continued pretraining using MAML-style meta-learning, a method

that directly optimizes few-shot promptability, yields subpar performance. We validate our findings with two prompt tuning methods, and, based on our results, we provide concrete recommendations to optimize promptability for different use cases.

**Non-Autoregressive Neural Machine Translation: A Call for Clarity**
*Robin Schmidt, Telmo Pires, Stephan Peitz and Jonas Lööf* 11:00-12:30 (Atrium)
Non-autoregressive approaches aim to improve the inference speed of translation models by only requiring a single forward pass to generate the output sequence instead of iteratively producing each predicted token. Consequently, their translation quality still tends to be inferior to their autoregressive counterparts due to several issues involving output token interdependence. In this work, we take a step back and revisit several techniques that have been proposed for improving non-autoregressive translation models and compare their combined translation quality and speed implications under third-party testing environments. We provide novel insights for establishing strong baselines using length prediction or CTC-based architecture variants and contribute standardized BLEU, chrF++, and TER scores using sacreBLEU on four translation tasks, which crucially have been missing as inconsistencies in the use of tokenized BLEU lead to deviations of up to 1.7 BLEU points. Our open-sourced code is integrated into fairseq for reproducibility.

**Abstractive Summarization Guided by Latent Hierarchical Document Structure**
*Yifu Qiu and Shay B. Cohen* 11:00-12:30 (Atrium)
Sequential abstractive neural summarizers often do not use the underlying structure in the input article or dependencies between the input sentences. This structure is essential to integrate and consolidate information from different parts of the text. To address this shortcoming, we propose a hierarchy-aware graph neural network (HierGNN) which captures such dependencies through three main steps: 1) learning a hierarchical document structure through a latent structure tree learned by a sparse matrix-tree computation; 2) propagating sentence information over this structure using a novel message-passing node propagation mechanism to identify salient information; 3) using graph-level attention to concentrate the decoder on salient information. Experiments confirm HierGNN improves strong sequence models such as BART, with a 0.55 and 0.75 margin in average ROUGE-1/2/L for CNN/DM and XSum. Further human evaluation demonstrates that summaries produced by our model are more relevant and less redundant than the baselines, into which HierGNN is incorporated. We also find HierGNN synthesizes summaries by fusing multiple source sentences more, rather than compressing a single source sentence, and that it processes long inputs more effectively.

**PLOG: Table-to-Logic Pretraining for Logical Table-to-Text Generation**
*Ao Liu, Haoyu Dong, Naoaki Okazaki, Shi Han and Dongmei Zhang* 11:00-12:30 (Atrium)
Logical table-to-text generation is a task that involves generating logically faithful sentences from tables, which requires models to derive logical-level facts from table records via logical inference. It raises a new challenge on the logical-level content planning of table-to-text models. However, directly learning the logical inference knowledge from table-text pairs is very difficult for neural models because of the ambiguity of natural language and the scarcity of parallel data. Hence even large-scale pre-trained language models present low logical fidelity on logical table-to-text. In this work, we propose a Pretrained Logical Form Generator (PLOG) framework to improve generation fidelity. Specifically, PLOG is first pretrained on a table-to-logical-form generation (table-to-logic) task, then finetuned on downstream table-to-text tasks. The logical forms are formally defined with unambiguous semantics. Hence we can collect a large amount of accurate logical forms from tables without human annotation. In addition, PLOG can learn logical inference from table-logic pairs much more reliably than from table-text pairs. To evaluate our model, we further collect a controlled logical table-to-text dataset CONTLOG based on an existing dataset. On two benchmarks, LOGICNLG and CONTLOG, PLOG outperforms strong baselines by a large margin on the logical fidelity, demonstrating the effectiveness of table-to-logic pretraining.

**Training Language Models with Memory Augmentation**
*Zexuan Zhong, Tao Lei and Danqi Chen* 11:00-12:30 (Atrium)
Recent work has improved language models (LMs) remarkably by equipping them with a non-parametric memory component. However, most existing approaches only introduce mem-ories at testing time or represent them using a separately trained encoder, resulting in suboptimal training of the language model. In this work, we present TRIME, a novel yet simple training approach designed for training LMs with memory augmentation. Our approach uses a training objective that directly takes in-batch examples as accessible memory. We also present new methods for memory construction and data batching, which are used for adapting to different sets of memories—local, long-term, and external memory—at testing time. We evaluate TRIME on multiple language modeling and machine translation benchmarks and show that it is able to achieve significant improvements across all the settings. Concretely, TRIME reduces the perplexity from 18.70 to 15.37 on WIKITEXT-103, by effectively leveraging a large memory set from the training corpus. Compared to standard LM training, TRIME adds negligible computational overhead and is compatible with different neural architectures, making it a versatile solution for training memory-augmented LMs.

**Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation**
*Max Glockner, Yufang Hou and Iryna Gurevych* 11:00-12:30 (Atrium)
Misinformation emerges in times of uncertainty when credible information is limited. This is challenging for NLP-based fact-checking as it relies on counter-evidence, which may not yet be available. Despite increasing interest in automatic fact-checking, it is still unclear if automated approaches can realistically refute harmful real-world misinformation. Here, we contrast and compare NLP fact-checking with how professional fact-checkers combat misinformation in the absence of counter-evidence. In our analysis, we show that, by design, existing NLP task definitions for fact-checking cannot refute misinformation as professional fact-checkers do for the majority of claims. We then define two requirements that the evidence in datasets must fulfill for realistic fact-checking: It must be (1) sufficient to refute the claim and (2) not leaked from existing fact-checking articles. We survey existing fact-checking datasets and find that all of them fail to satisfy both criteria. Finally, we perform experiments to demonstrate that models trained on a large-scale fact-checking dataset rely on leaked evidence, which makes them unsuitable in real-world scenarios. Taken together, we show that current NLP fact-checking cannot realistically combat real-world misinformation because it depends on unrealistic assumptions about counter-evidence in the data.

**TemporalWiki: A Lifelong Benchmark for Training and Evaluating Ever-Evolving Language Models**
*Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim and Minjoon Seo*11:00-12:30 (Atrium)
Language Models (LMs) become outdated as the world changes; they often fail to perform tasks requiring recent factual information which was absent or different during training, a phenomenon called temporal misalignment. This is especially a challenging problem because the research community still lacks a coherent dataset for assessing the adaptability of LMs to frequently-updated knowledge corpus such as Wikipedia. To this end, we introduce TemporalWiki, a lifelong benchmark for ever-evolving LMs that utilizes the difference between consecutive snapshots of English Wikipedia and English Wikidata for training and evaluation, respectively. The benchmark hence allows researchers to periodically track an LM's ability to retain previous knowledge and acquire updated/new knowledge at each point in time. We also find that training an LM on the diff data through continual learning methods achieves similar or better perplexity than on the entire snapshot in our benchmark with 12 times less computational cost, which verifies that factual knowledge in LMs can be safely updated with minimal training data via continual learning.

**Video Question Answering: Datasets, Algorithms and Challenges**

*Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng and Tat-Seng Chua* 11:00-12:30 (Atrium)
This survey aims to sort out the recent advances in video question answering (VideoQA) and point towards future directions. We firstly categorize the datasets into 1) normal VideoQA, multi-modal VideoQA and knowledge-based VideoQA, according to the modalities invoked in the question-answer pairs, or 2) factoid VideoQA and inference VideoQA, according to the technical challenges in comprehending the questions and deriving the correct answers. We then summarize the VideoQA techniques, including those mainly designed for Factoid QA (e.g., the early spatio-temporal attention-based methods and the recently Transformer-based ones) and those targeted at explicit relation and logic inference (e.g., neural modular networks, neural symbolic methods, and graph-structured methods). Aside from the backbone techniques, we delve into the specific models and find out some common and useful insights either for video modeling, question answering, or for cross-modal correspondence learning. Finally, we point out the research trend of studying beyond factoid VideoQA to inference VideoQA, as well as towards the robustness and interpretability. Additionally, we maintain a repository, https://github.com/VRU-NExT/VideoQA, to keep trace of the latest VideoQA papers, datasets, and their open-source implementations if available. With these efforts, we strongly hope this survey could shed light on the follow-up VideoQA research.

**ADDMU: Detection of Far-Boundary Adversarial Examples with Data and Model Uncertainty Estimation**
*Fan Yin, Yao Li, Cho-Jui Hsieh and Kai-Wei Chang* 11:00-12:30 (Atrium)
Adversarial Examples Detection (AED) is a crucial defense technique against adversarial attacks and has drawn increasing attention from the Natural Language Processing (NLP) community. Despite the surge of new AED methods, our studies show that existing methods heavily rely on a shortcut to achieve good performance. In other words, current search-based adversarial attacks in NLP stop once model predictions change, and thus most adversarial examples generated by those attacks are located near model decision boundaries. To surpass this shortcut and fairly evaluate AED methods, we propose to test AED methods with **F**ar **B**oundary (**FB**) adversarial examples. Existing methods show worse than random guess performance under this scenario. To overcome this limitation, we propose a new technique, **ADDMU**, **a**dversary **d**etection with **d**ata and **m**odel **u**ncertainty, which combines two types of uncertainty estimation for both regular and FB adversarial example detection. Our new method outperforms previous methods by 3.6 and 6.0 *AUC* points under each scenario. Finally, our analysis shows that the two types of uncertainty provided by **ADDMU** can be leveraged to characterize adversarial examples and identify the ones that contribute most to model's robustness in adversarial training.

**Help me write a Poem - Instruction Tuning as a Vehicle for Collaborative Poetry Writing**
*Tuhin Chakrabarty, Vishakh Padmakumar and He He* 11:00-12:30 (Atrium)
Recent work in training large language models (LLMs) to follow natural language instructions has opened up exciting opportunities for natural language interface design. Building on the prior success of large language models in the realm of computer assisted creativity, in this work, we present *CoPoet*, a collaborative poetry writing system, with the goal of to study if LLM's actually improve the quality of the generated content. In contrast to auto-completing a user's text, CoPoet is controlled by user instructions that specify the attributes of the desired text, such as *Write a sentence about 'love'* or *Write a sentence ending in 'fly'*. The core component of our system is a language model fine-tuned on a diverse collection of instructions for poetry writing. Our model is not only competitive to publicly available LLMs trained on instructions (InstructGPT), but also capable of satisfying unseen compositional instructions. A study with 15 qualified crowdworkers shows that users successfully write poems with CoPoet on diverse topics ranging from *Monarchy* to *Climate change*, which are preferred by third-party evaluators over poems written without the system.

**Does Self-Rationalization Improve Robustness to Spurious Correlations?**
*Alexis Ross, Matthew Peters and Ana Marasovic* 11:00-12:30 (Atrium)
Rationalization is fundamental to human reasoning and learning. NLP models trained to produce rationales along with predictions, called self-rationalization models, have been investigated for their interpretability and utility to end-users. However, the extent to which training with human-written rationales facilitates learning remains an under-explored question. We ask whether training models to self-rationalize can aid in their learning to solve tasks for the right reasons. Specifically, we evaluate how training self-rationalization models with free-text rationales affects robustness to spurious correlations in fine-tuned encoder-decoder and decoder-only models of six different sizes. We evaluate robustness to spurious correlations by measuring performance on 1) manually annotated challenge datasets and 2) subsets of original test sets where reliance on spurious correlations would fail to produce correct answers. We find that while self-rationalization can improve robustness to spurious correlations in low-resource settings, it tends to hurt robustness in higher-resource settings. Furthermore, these effects depend on model family and size, as well as on rationale content. Together, our results suggest that explainability can come at the cost of robustness; thus, appropriate care should be taken when training self-rationalizing models with the goal of creating more trustworthy models.

**Efficient Pre-training of Masked Language Model via Concept-based Curriculum Masking**
*Mingyu Lee, Jun-Hyung Park, Junho Kim, Kang-Min Kim and SangKeun Lee* 11:00-12:30 (Atrium)
Self-supervised pre-training has achieved remarkable success in extensive natural language processing tasks. Masked language modeling (MLM) has been widely used for pre-training effective bidirectional representations but comes at a substantial training cost. In this paper, we propose a novel concept-based curriculum masking (CCM) method to efficiently pre-train a language model. CCM has two key differences from existing curriculum learning approaches to effectively reflect the nature of MLM. First, we introduce a novel curriculum that evaluates the MLM difficulty of each token based on a carefully-designed linguistic difficulty criterion. Second, we construct a curriculum that masks easy words and phrases first and gradually masks related ones to the previously masked ones based on a knowledge graph. Experimental results show that CCM significantly improves pre-training efficiency. Specifically, the model trained with CCM shows comparative performance with the original BERT on the General Language Understanding Evaluation benchmark at half of the training cost.

**Differentiable Data Augmentation for Contrastive Sentence Representation Learning**
*Tianduo Wang and Wei Lu* 11:00-12:30 (Atrium)
Fine-tuning a pre-trained language model via the contrastive learning framework with a large amount of unlabeled sentences or labeled sentence pairs is a common way to obtain high-quality sentence representations. Although the contrastive learning framework has shown its superiority on sentence representation learning over previous methods, the potential of such a framework is under-explored so far due to the simple method it used to construct positive pairs. Motivated by this, we propose a method that makes hard positives from the original training examples. A pivotal ingredient of our approach is the use of prefix that attached to a pre-trained language model, which allows for differentiable data augmentation during contrastive learning. Our method can be summarized in two steps: supervised prefix-tuning followed by joint contrastive fine-tuning with unlabeled or labeled examples. Our experiments confirm the effectiveness of our data augmentation approach. The proposed method yields significant improvements over existing methods under both semi-supervised and supervised settings. Our experiments under a low labeled data setting also show that our method is more label-efficient than the state-of-the-art contrastive learning methods.

**QASem Parsing: Text-to-text Modeling of QA-based Semantics**
*Ayal Klein, Eran Hirsch, Ron Eliav, Valentina Pyatkin, Avi Caciularu and Ido Dagan* 11:00-12:30 (Atrium)
Various works suggest the appeals of incorporating explicit semantic representations when addressing challenging realistic NLP scenarios. Common approaches offer either comprehensive linguistically-based formalisms, like AMR, or alternatively Open-IE, which provides a shallow and partial representation. More recently, an appealing trend introduces semi-structured natural-language structures as an intermediate

meaning-capturing representation, often in the form of questions and answers.
In this work, we further promote this line of research by considering three prior QA-based semantic representations. These cover verbal, nominalized and discourse-based predications, regarded as jointly providing a comprehensive representation of textual information — termed QASem. To facilitate this perspective, we investigate how to best utilize pre-trained sequence-to-sequence language models, which seem particularly promising for generating representations that consist of natural language expressions (questions and answers). In particular, we examine and analyze input and output linearization strategies, as well as data augmentation and multitask learning for a scarce training data setup. Consequently, we release the first unified QASem parsing tool, easily applicable for downstream tasks that can benefit from an explicit semi-structured account of information units in text.

**When does Parameter-Efficient Transfer Learning Work for Machine Translation?**
*Ahmet Üstün and Asa Cooper Stickland* 11:00-12:30 (Atrium)
Parameter-efficient fine-tuning methods (PEFTs) offer the promise of adapting large pre-trained models while only tuning a small number of parameters. They have been shown to be competitive with full model fine-tuning for many downstream tasks. However, prior work indicates that PEFTs may not work as well for machine translation (MT), and there is no comprehensive study showing when PEFTs work for MT. We conduct a comprehensive empirical study of PEFTs for MT, considering (1) various parameter budgets, (2) a diverse set of language-pairs, and (3) different pre-trained models. We find that 'adapters', in which small feed-forward networks are added after every layer, are indeed on par with full model fine-tuning when the parameter budget corresponds to 10% of total model parameters. Nevertheless, as the number of tuned parameters decreases, the performance of PEFTs decreases. The magnitude of this decrease depends on the language pair, with PEFTs particularly struggling for distantly related language-pairs. We find that using PEFTs with a larger pre-trained model outperforms full fine-tuning with a smaller model, and for smaller training data sizes, PEFTs outperform full fine-tuning for the same pre-trained model.

**Towards Pragmatic Production Strategies for Natural Language Generation Tasks**
*Mario Giulianelli* 11:00-12:30 (Atrium)
This position paper proposes a conceptual framework for the design of Natural Language Generation (NLG) systems that follow efficient and effective production strategies in order to achieve complex communicative goals. In this general framework, efficiency is characterised as the parsimonious regulation of production and comprehension costs while effectiveness is measured with respect to task-oriented and contextually grounded communicative goals. We provide concrete suggestions for the estimation of goals, costs, and utility via modern statistical methods, demonstrating applications of our framework to the classic pragmatic task of visually grounded referential games and to abstractive text summarisation, two popular generation tasks with real-world applications. In sum, we advocate for the development of NLG systems that learn to make pragmatic production decisions from experience, by reasoning about goals, costs, and utility in a human-like way.

**Hierarchical Phrase-Based Sequence-to-Sequence Learning**
*Bailin Wang, Ivan Titov, Jacob Andreas and Yoon Kim* 11:00-12:30 (Atrium)
This paper describes a neural transducer that maintains the flexibility of standard sequence-to-sequence (seq2seq) models while incorporating hierarchical phrases as a source of inductive bias during training and as explicit constraints during inference. Our approach trains two models: a discriminative parser based on a bracketing transduction grammar whose derivation tree hierarchically aligns source and target phrases, and a neural seq2seq model that learns to translate the aligned phrases one-by-one. We use the same seq2seq model to translate at all phrase scales, which results in two inference modes: one mode in which the parser is discarded and only the seq2seq component is used at the sequence-level, and another in which the parser is combined with the seq2seq model. Decoding in the latter mode is done with the cube-pruned CKY algorithm, which is more involved but can make use of new translation rules during inference. We formalize our model as a source-conditioned synchronous grammar and develop an efficient variational inference algorithm for training. When applied on top of both randomly initialized and pretrained seq2seq models, we find that it performs well compared to baselines on small scale machine translation benchmarks.

**SMaLL-100: Introducing Shallow Multilingual Machine Translation Model for Low-Resource Languages**
*Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson and Laurent Besacier* 11:00-12:30 (Atrium)
In recent years, multilingual machine translation models have achieved promising performance on low-resource language pairs by sharing information between similar languages, thus enabling zero-shot translation. To overcome the "curse of multilinguality", these models often opt for scaling up the number of parameters, which makes their use in resource-constrained environments challenging. We introduce SMaLL-100, a distilled version of the M2M-100(12B) model, a massively multilingual machine translation model covering 100 languages. We train SMaLL-100 with uniform sampling across all language pairs and therefore focus on preserving the performance of low-resource languages. We evaluate SMaLL-100 on different low-resource benchmarks: FLORES-101, Tatoeba, and TICO-19 and demonstrate that it outperforms previous massively multilingual models of comparable sizes (200-600M) while improving inference latency and memory usage. Additionally, our model achieves comparable results to M2M-100 (1.2B), while being 3.6x smaller and 4.3x faster at inference.

**Better Hit the Nail on the Head than Beat around the Bush: Removing Protected Attributes with a Single Projection**
*Pantea Haghighatkhah, Antske Fokkens, Pia Sommerauer, Bettina Speckmann and Kevin Verbeek* 11:00-12:30 (Atrium)
Bias elimination and recent probing studies attempt to remove specific information from embedding spaces. Here it is important to remove as much of the target information as possible, while preserving any other information present. INLP is a popular recent method which removes specific information through iterative nullspace projections. Multiple iterations, however, increase the risk that information other than the target is negatively affected. We introduce two methods that find a single targeted projection: Mean Projection (MP, more efficient) and Tukey Median Projection (TMP, with theoretical guarantees). Our comparison between MP and INLP shows that (1) one MP projection removes linear separability based on the target and (2) MP has less impact on the overall space. Further analysis shows that applying random projections after MP leads to the same overall effects on the embedding space as the multiple projections of INLP. Applying one targeted (MP) projection hence is methodologically cleaner than applying multiple (INLP) projections that introduce random effects.

**Revisiting DocRED - Addressing the False Negative Problem in Relation Extraction**
*Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng and Sharifah Mahani Aljunied* 11:00-12:30 (Atrium)
The DocRED dataset is one of the most popular and widely used benchmarks for document-level relation extraction (RE). It adopts a recommend-revise annotation scheme so as to have a large-scale annotated dataset. However, we find that the annotation of DocRED is incomplete, i.e., false negative samples are prevalent. We analyze the causes and effects of the overwhelming false negative problem in the DocRED dataset. To address the shortcoming, we re-annotate 4,053 documents in the DocRED dataset by adding the missed relation triples back to the original DocRED. We name our revised DocRED dataset Re-DocRED. We conduct extensive experiments with state-of-the-art neural models on both datasets, and the experimental results show that the models trained and evaluated on our Re-DocRED achieve performance improvements of around 13 F1 points. Moreover, we conduct a comprehensive analysis to identify the potential areas for further improvement.

**Beyond prompting: Making Pre-trained Language Models Better Zero-shot Learners by Clustering Representations**
*Yu Fei, Zhao Meng, Ping Nie, Roger Wattenhofer and Mrinmaya Sachan* 11:00-12:30 (Atrium)

Recent work has demonstrated that pre-trained language models (PLMs) are zero-shot learners. However, most existing zero-shot methods involve heavy human engineering or complicated self-training pipelines, hindering their application to new situations. In this work, we show that zero-shot text classification can be improved simply by clustering texts in the embedding spaces of PLMs. Specifically, we fit the unlabeled texts with a Bayesian Gaussian Mixture Model after initializing cluster positions and shapes using class names. Despite its simplicity, this approach achieves superior or comparable performance on both topic and sentiment classification datasets and outperforms prior works significantly on unbalanced datasets. We further explore the applicability of our clustering approach by evaluating it on 14 datasets with more diverse topics, text lengths, and numbers of classes. Our approach achieves an average of 20% absolute improvement over prompt-based zero-shot learning. Finally, we compare different PLM embedding spaces and find that texts are well-clustered by topics even if the PLM is not explicitly pre-trained to generate meaningful sentence embeddings. This work indicates that PLM embeddings can categorize texts without task-specific fine-tuning, thus providing a new way to analyze and utilize their knowledge and zero-shot learning ability.

**AfriCLIRMatrix: Enabling Cross-Lingual Information Retrieval for African Languages**
*Odunayo Ogundepo, Xinyu Zhang, Shuo Sun, Kevin Duh and Jimmy Lin* 11:00-12:30 (Atrium)
Language diversity in NLP is critical in enabling the development of tools for a wide range of users. However, there are limited resources for building such tools for many languages, particularly those spoken in Africa. For search, most existing datasets feature few or no African languages, directly impacting researchers' ability to build and improve information access capabilities in those languages. Motivated by this, we created AfriCLIRMatrix, a test collection for cross-lingual information retrieval research in 15 diverse African languages. In total, our dataset contains 6 million queries in English and 23 million relevance judgments automatically mined from Wikipedia inter-language links, covering many more African languages than any existing information retrieval test collection. In addition, we release BM25, dense retrieval, and sparse–dense hybrid baselines to provide a starting point for the development of future systems. We hope that these efforts can spur additional work in search for African languages. AfriCLIRMatrix can be downloaded at https://github.com/castorini/africlirmatrix.

**ArtELingo: A Million Emotion Annotations of WikiArt with Emphasis on Diversity over Language and Culture**
*Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Church and Mohamed Elhoseiny*11:00-12:30 (Atrium)
This paper introduces ArtELingo, a new benchmark and dataset, designed to encourage work on diversity across languages and cultures. Following ArtEmis, a collection of 80k artworks from WikiArt with 0.45M emotion labels and English-only captions, ArtELingo adds another 0.79M annotations in Arabic and Chinese, plus 4.8K in Spanish to evaluate "cultural-transfer" performance. 51K artworks have 5 annotations or more in 3 languages. This diversity makes it possible to study similarities and differences across languages and cultures. Further, we investigate captioning tasks, and find diversity improves the performance of baseline models. ArtELingo is publicly available at 'www.artelingo.org' with standard splits and baseline models. We hope our work will help ease future research on multilinguality and culturally-aware AI.

**Are representations built from the ground up? An empirical examination of local composition in language models**
*Emmy Liu and Graham Neubig* 11:00-12:30 (Atrium)
Compositionality, the phenomenon where the meaning of a phrase can be derived from its constituent parts, is a hallmark of human language. At the same time, many phrases are non-compositional, carrying a meaning beyond that of each part in isolation. Representing both of these types of phrases is critical for language understanding, but it is an open question whether modern language models (LMs) learn to do so; in this work we examine this question. We first formulate a problem of predicting the LM-internal representations of longer phrases given those of their constituents. We find that the representation of a parent phrase can be predicted with some accuracy given an affine transformation of its children. While we would expect the predictive accuracy to correlate with human judgments of semantic compositionality, we find this is largely not the case, indicating that LMs may not accurately distinguish between compositional and non-compositional phrases. We perform a variety of analyses, shedding light on when different varieties of LMs do and do not generate compositional representations, and discuss implications for future modeling work.

**Late Fusion with Triplet Margin Objective for Multimodal Ideology Prediction and Analysis**
*Changyuan Qiu, Winston Wu, Xinliang Frederick Zhang and Lu Wang* 11:00-12:30 (Atrium)
Prior work on ideology prediction has largely focused on single modalities, i.e., text or images. In this work, we introduce the task of multimodal ideology prediction, where a model predicts binary or five-point scale ideological leanings, given a text-image pair with political content. We first collect five new large-scale datasets with English documents and images along with their ideological leanings, covering news articles from a wide range of mainstream media in US and social media posts from Reddit and Twitter. We conduct in-depth analyses on news articles and reveal differences in image content and usage across the political spectrum. Furthermore, we perform extensive experiments and ablation studies, demonstrating the effectiveness of targeted pretraining objectives on different model components. Our best-performing model, a late-fusion architecture pretrained with a triplet objective over multimodal content, outperforms the state-of-the-art text-only model by almost 4% and a strong multimodal baseline with no pretraining by over 3%.

**Meta-Learning Fast Weight Language Models**
*Kevin Clark, Kelvin Guu, Ming-Wei Chang, Panupong Pasupat, Geoffrey Hinton and Mohammad Norouzi* 11:00-12:30 (Atrium)
Dynamic evaluation of language models (LMs) adapts model parameters at test time using gradient information from previous tokens and substantially improves LM performance. However, it requires over 3x more compute than standard inference. We present Fast Weight Layers (FWLs), a neural component that provides the benefits of dynamic evaluation much more efficiently by expressing gradient updates as linear attention. A key improvement over dynamic evaluation is that FWLs can also be applied at training time, so the model learns to make good use of gradient updates. FWLs can easily be added on top of existing transformer models, require relatively little extra compute or memory to run, and significantly improve language modeling perplexity.

**Large Dual Encoders Are Generalizable Retrievers**
*Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang and Yinfei Yang* 11:00-12:30 (Atrium)
It has been shown that dual encoders trained on one domain often fail to generalize to other domains for retrieval tasks. One widespread belief is that the bottleneck layer of a dual encoder, where the final score is simply a dot-product between a query vector and a passage vector, is too limited compared to models with fine-grained interactions between the query and the passage. In this paper, we challenge this belief by scaling up the size of the dual encoder model *while keeping the bottleneck layer as a single dot-product with a fixed size.* With multi-stage training, scaling up the model size brings significant improvement on a variety of retrieval tasks, especially for out-of-domain generalization. We further analyze the impact of the bottleneck layer and demonstrate diminishing improvement when scaling up the embedding size. Experimental results show that our dual encoders, **G**eneralizable **T**5-based dense **R**etrievers (GTR), outperform previous sparse and dense retrievers on the BEIR dataset significantly. Most surprisingly, our ablation study finds that GTR is very data efficient, as it only needs 10

**Specializing Multi-domain NMT via Penalizing Low Mutual Information**
*Jiyoung Lee, Hantae Kim, Hyunchang Cho, Edward Choi and Cheonbok Park* 11:00-12:30 (Atrium)
Multi-domain Neural Machine Translation (NMT) trains a single model with multiple domains. It is appealing because of its efficacy in

handling multiple domains within one model. An ideal multi-domain NMT learns distinctive domain characteristics simultaneously, however, grasping the domain peculiarity is a non-trivial task. In this paper, we investigate domain-specific information through the lens of mutual information (MI) and propose a new objective that penalizes low MI to become higher.Our method achieved the state-of-the-art performance among the current competitive multi-domain NMT models. Also, we show our objective promotes low MI to be higher resulting in domain-specialized multi-domain NMT.

**A Dataset for Hyper-Relational Extraction and a Cube-Filling Approach**
*Yew Ken Chia, Lidong Bing, Sharifah Mahani Aljunied, Luo Si and Soujanya Poria* 11:00-12:30 (Atrium)
Relation extraction has the potential for large-scale knowledge graph construction, but current methods do not consider the qualifier attributes for each relation triplet, such as time, quantity or location. The qualifiers form hyper-relational facts which better capture the rich and complex knowledge graph structure. For example, the relation triplet (Leonard Parker, Educated At, Harvard University) can be factually enriched by including the qualifier (End Time, 1967). Hence, we propose the task of hyper-relational extraction to extract more specific and complete facts from text. To support the task, we construct HyperRED, a large-scale and general-purpose dataset. Existing models cannot perform hyper-relational extraction as it requires a model to consider the interaction between three entities. Hence, we propose CubeRE, a cube-filling model inspired by table-filling approaches and explicitly considers the interaction between relation triplets and qualifiers. To improve model scalability and reduce negative class imbalance, we further propose a cube-pruning method. Our experiments show that CubeRE outperforms strong baselines and reveal possible directions for future research. Our code and data are available at github.com/declare-lab/HyperRED.

**SentBS: Sentence-level Beam Search for Controllable Summarization**
*Chenhui Shen, Liying Cheng, Lidong Bing, Yang You and Luo Si* 11:00-12:30 (Atrium)
A wide range of control perspectives have been explored in controllable text generation. Structure-controlled summarization is recently proposed as a useful and interesting research direction. However, current structure-controlling methods have limited effectiveness in enforcing the desired structure. To address this limitation, we propose a sentence-level beam search generation method (SentBS), where evaluation is conducted throughout the generation process to select suitable sentences for subsequent generations. We experiment with different combinations of decoding methods to be used as sub-components by SentBS and evaluate results on the structure-controlled dataset MReD. Experiments show that all explored combinations for SentBS can improve the agreement between the generated text and the desired structure, with the best method significantly reducing the structural discrepancies suffered by the existing model, by approximately 68%.

**A Fine-grained Chinese Software Privacy Policy Dataset for Sequence Labeling and Regulation Compliant Identification**
*Kaifa Zhao, Le Yu, Shiyao Zhou, Jing Li, Xiapu Luo, Yat Fei Aemon Chiu and Yutong Liu* 11:00-12:30 (Atrium)
Privacy protection raises great attention on both legal levels and user awareness. To protect user privacy, countries enact laws and regulations requiring software privacy policies to regulate their behavior. However, privacy policies are written in professional languages with many legal terms and software jargon that prevent users from understanding and even reading them. It is necessary and urgent to use NLP techniques to analyze privacy policies. However, existing datasets ignore law requirements and are limited to English. In this paper, we construct the first Chinese privacy policy dataset, namely CA4P-483, to facilitate the sequence labeling tasks and regulation compliance identification between privacy policies and software. Our dataset includes 483 Chinese Android application privacy policies, over 11K sentences, and 52K fine-grained annotations. We evaluate families of robust and representative baseline models on our dataset. Based on baseline performance, we provide findings and potential research directions on our dataset. Finally, we investigate the potential applications of CA4P-483 combing regulation requirements and program analysis.

**Variational Autoencoder with Disentanglement Priors for Low-Resource Task-Specific Natural Language Generation**
*Zhuang Li, Lizhen Qu, Qiongkai Xu, Tongtong Wu, Tianyang Zhan and Gholamreza Haffari* 11:00-12:30 (Atrium)
In this paper, we propose a variational autoencoder with disentanglement priors, VAE-Dprior, for task-specific natural language generation with none or a handful of task-specific labeled examples. In order to tackle compositional generalization across tasks, our model performs disentangled representation learning by introducing a conditional prior for the latent content space and another conditional prior for the latent label space. Both types of priors satisfy a novel property called $\epsilon$-disentangled. We show both empirically and theoretically that the novel priors can disentangle representations even without specific regularizations as in the prior work. The content prior enables directly sampling diverse content representations from the content space learned from the seen tasks, and fuse them with the representations of novel tasks for generating semantically diverse texts in the low-resource settings. Our extensive experiments demonstrate the superior performance of our model over competitive baselines in terms of i) data augmentation in continuous zero/few-shot learning, and ii) text style transfer in the few-shot setting.

**The better your Syntax, the better your Semantics? Probing Pretrained Language Models for the English Comparative Correlative**
*Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal and Hinrich Schütze* 11:00-12:30 (Atrium)
Construction Grammar (CxG) is a paradigm from cognitive linguistics emphasising the connection between syntax and semantics. Rather than rules that operate on lexical items, it posits constructions as the central building blocks of language, i.e., linguistic units of different granularity that combine syntax and semantics. As a first step towards assessing the compatibility of CxG with the syntactic and semantic knowledge demonstrated by state-of-the-art pretrained language models (PLMs), we present an investigation of their capability to classify and understand one of the most commonly studied constructions, the English comparative correlative (CC). We conduct experiments examining the classification accuracy of a syntactic probe on the one hand and the models' behaviour in a semantic application task on the other, with BERT, RoBERTa, and DeBERTa as the example PLMs. Our results show that all three investigated PLMs are able to recognise the structure of the CC but fail to use its meaning. While human-like performance of PLMs on many NLP tasks has been alleged, this indicates that PLMs still suffer from substantial shortcomings in central domains of linguistic knowledge.

**Federated Meta-Learning for Emotion and Sentiment Aware Multi-modal Complaint Identification**
*Apoorva Singh, C Siddarth, Sriparna Saha and Tanmay Sen* 11:00-12:30 (Atrium)

**Balancing out Bias: Achieving Fairness Through Balanced Training**
*Xudong Han, Timothy Baldwin and Trevor Cohn* 11:00-12:30 (Atrium)
Group bias in natural language processing tasks manifests as disparities in system error rates across texts authorized by different demographic groups, typically disadvantaging minority groups. Dataset balancing has been shown to be effective at mitigating bias, however existing approaches do not directly account for correlations between author demographics and linguistic variables, limiting their effectiveness. To achieve Equal Opportunity fairness, such as equal job opportunity without regard to demographics, this paper introduces a simple, but highly effective, objective for countering bias using balanced training. We extend the method in the form of a gated model, which incorporates protected attributes as input, and show that it is effective at reducing bias in predictions through demographic input perturbation, outperforming all other bias mitigation techniques when combined with balanced training.

**Prompting ELECTRA: Few-Shot Learning with Discriminative Pre-Trained Models**

*Mengzhou Xia, Mikel Artetxe, Jingfei Du, Danqi Chen and Veselin Stoyanov* 11:00-12:30 (Atrium)
Pre-trained masked language models successfully perform few-shot learning by formulating downstream tasks as text infilling. How- ever, as a strong alternative in full-shot settings, discriminative pre-trained models like ELECTRA do not fit into the paradigm. In this work, we adapt prompt-based few-shot learning to ELECTRA and show that it outperforms masked language models in a wide range of tasks. ELECTRA is pre-trained to distinguish if a token is generated or original. We naturally extend that to prompt-based few-shot learning by training to score the originality of the target options without introducing new parameters. Our method can be easily adapted to tasks involving multi-token predictions without extra computation overhead. Analysis shows that ELECTRA learns distributions that align better with downstream tasks.

**CDialog: A Multi-turn Covid-19 Conversation Dataset for Entity-Aware Dialog Generation**
*Deeksha Varshney, Aizan Zafar, Niranshu Behera and Asif Ekbal* 11:00-12:30 (Atrium)

**PCL: Peer-Contrastive Learning with Diverse Augmentations for Unsupervised Sentence Embeddings**
*Qiyu Wu, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng and Daxin Jiang* 11:00-12:30 (Atrium)
Learning sentence embeddings in an unsupervised manner is fundamental in natural language processing. Recent common practice is to couple pre-trained language models with unsupervised contrastive learning, whose success relies on augmenting a sentence with a semantically-close positive instance to construct contrastive pairs. Nonetheless, existing approaches usually depend on a mono-augmenting strategy, which causes learning shortcuts towards the augmenting biases and thus corrupts the quality of sentence embeddings. A straightforward solution is resorting to more diverse positives from a multi-augmenting strategy, while an open question remains about how to unsupervisedly learn from the diverse positives but with uneven augmenting qualities in the text field. As one answer, we propose a novel Peer-Contrastive Learning (PCL) with diverse augmentations. PCL constructs diverse contrastive positives and negatives at the group level for unsupervised sentence embeddings. PCL performs peer-positive contrast as well as peer-network cooperation, which offers an inherent anti-bias ability and an effective way to learn from diverse augmentations. Experiments on STS benchmarks verify the effectiveness of PCL against its competitors in unsupervised sentence embeddings.

**Exploring the Secrets Behind the Learning Difficulty of Meaning Representations for Semantic Parsing**
*Zhenwen Li, Jiaqi Guo, Qian Liu, Jian-Guang LOU and Tao Xie* 11:00-12:30 (Atrium)
Previous research has shown that the design of Meaning Representation (MR) greatly influences the final model performance of a neural semantic parser. Therefore, designing a good MR is a long-term goal for semantic parsing. However, it is still an art as there is no quantitative indicator that can tell us which MR among a set of candidates may have the best final model performance. In practice, in order to select an MR design, researchers often have to go through the whole training-testing process for all design candidates, and the process often costs a lot. In this paper, we propose a data-aware metric called ISS (denoting incremental structural stability) of MRs, and demonstrate that ISS is highly correlated with the final performance. The finding shows that ISS can be used as an indicator for MR design to avoid the costly training-testing process.

**Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs**
*Maarten Sap, Ronan Le Bras, Daniel Fried and Yejin Choi* 11:00-12:30 (Atrium)
Social intelligence and Theory of Mind (TOM), i.e., the ability to reason about the different mental states, intents, and reactions of all people involved, allows humans to effectively navigate and understand everyday social interactions. As NLP systems are used in increasingly complex social situations, their ability to grasp social dynamics becomes crucial.

In this work, we examine the open question of social intelligence and Theory of Mind in modern NLP systems from an empirical and theorybased perspective. We show that one of today's largest language models (GPT-3; Brown et al., 2020) lacks this kind of social intelligence out-of-the box, using two tasks: SocialIQa (Sap et al., 2019), which measure models' ability to understand intents and reactions of participants of social interactions, and ToMi (Le, Boureau, and Nickel, 2019), which measures whether models can infer mental states and realities of participants of situations.

Our results show that models struggle substantially at these Theory of Mind tasks, with well-below-human accuracies of 55% and 60% on SocialIQa and ToMi, respectively. To conclude, we draw on theories from pragmatics to contextualize this shortcoming of large language models, by examining the limitations stemming from their data, neural architecture, and training paradigms. Challenging the prevalent narrative that only scale is needed, we posit that person-centric NLP approaches might be more effective towards neural Theory of Mind.

**Summarizing Community-based Question-Answer Pairs**
*Ting-Yao Hsu, Yoshi Suhara and Xiaolan Wang* 11:00-12:30 (Atrium)
Community-based Question Answering (CQA), which allows users to acquire their desired information, has increasingly become an essential component of online services in various domains such as E-commerce, travel, and dining. However, an overwhelming number of CQA pairs makes it difficult for users without particular intent to find useful information spread over CQA pairs. To help users quickly digest the key information, we propose the novel CQA summarization task that aims to create a concise summary from CQA pairs. To this end, we first design a multi-stage data annotation process and create a benchmark dataset, COQASUM, based on the Amazon QA corpus. We then compare a collection of extractive and abstractive summarization methods and establish a strong baseline approach DedupLED for the CQA summarization task. Our experiment further confirms two key challenges, sentence-type transfer and deduplication removal, towards the CQA summarization task. Our data and code are publicly available.

[INDUSTRY] **A Comprehensive Evaluation of Biomedical Entity-centric Search**
*Elena Tutubalina, Zulfat Miftahutdinov, Vladimir Muravlev and Anastasia Shneyderman* 11:00-12:30 (Atrium)
Biomedical information retrieval has often been studied as a task of detecting whether a system correctly detects entity spans and links these entities to concepts from a given terminology. Most academic research has focused on evaluation of named entity recognition (NER) and entity linking (EL) models which are key components to recognizing diseases and genes in PubMed abstracts. In this work, we perform a fine-grained evaluation intended to understand the efficiency of state-of-the-art BERT-based information extraction (IE) architecture as a biomedical search engine. We present a novel manually annotated dataset of abstracts for disease and gene search. The dataset contains 23K query-abstract pairs, where 152 queries are selected from logs of our target discovery platform and PubMed abstracts annotated with relevance judgments. Specifically, the query list also includes a subset of concepts with at least one ambiguous concept name. As a baseline, we use off-she-shelf Elasticsearch with BM25. Our experiments on NER, EL, and retrieval in a zero-shot setup show the neural IE architecture shows superior performance for both disease and gene concept queries.

[INDUSTRY] **Domain Adaptation of Machine Translation with Crowdworkers**
*Makoto Morishita, Jun Suzuki and Masaaki Nagata* 11:00-12:30 (Atrium)
Although a machine translation model trained with a large in-domain parallel corpus achieves remarkable results, it still works poorly when no in-domain data are available. This situation restricts the applicability of machine translation when the target domain's data are limited. However, there is great demand for high-quality domain-specific machine translation models for many domains. We propose a framework that efficiently and effectively collects parallel sentences in a target domain from the web with the help of crowdworkers. With the collected parallel data, we can quickly adapt a machine translation model to the target domain. Our experiments show that the proposed method can

collect target-domain parallel data over a few days at a reasonable cost. We tested it with five domains, and the domain-adapted model improved the BLEU scores to +19.7 by an average of +7.8 points compared to a general-purpose translation model.

[INDUSTRY] **Biomedical NER for the Enterprise with Distillated BERN2 and the Kazu Framework**
*Wonjin Yoon, Richard G. Jackson, Elliot Ford, Vladimir Poroshin and Jaewoo Kang* 11:00-12:30 (Atrium)
In order to assist the drug discovery/development process, pharmaceutical companies often apply biomedical NER and linking techniques over internal and public corpora. Decades of study of the field of BioNLP has produced a plethora of algorithms, systems and datasets. However, our experience has been that no single open source system meets all the requirements of a modern pharmaceutical company. In this work, we describe these requirements according to our experience of the industry, and present Kazu, a highly extensible, scalable open source framework designed to support BioNLP for the pharmaceutical sector. Kazu is a built around a computationally efficient version of the BERN2 NER model (TinyBERN2), and subsequently wraps several other BioNLP technologies into one coherent system.

[INDUSTRY] **Topic Modeling by Clustering Language Model Embeddings: Human Validation on an Industry Dataset**
*Anton Eklund and Mona Forsman* 11:00-12:30 (Atrium)
Topic models are powerful tools to get an overview of large collections of text data, a situation that is prevalent in industry applications. A rising trend within topic modeling is to directly cluster dimension-reduced embeddings created with pretrained language models. It is difficult to evaluate these models because there is no ground truth and automatic measurements may not mimic human judgment. To address this problem, we created a tool called STELLAR for interactive topic browsing which we used for human evaluation of topics created from a real-world dataset used in industry. Embeddings created with BERT were used together with UMAP and HDBSCAN to model the topics. The human evaluation found that our topic model creates coherent topics. The following discussion revolves around the requirements of industry and what research is needed for production-ready systems.

## Demo Session 6

11:00-12:30 (Atrium)

[DEMO] **KeywordScape: Visual Document Exploration using Contextualized Keyword Embeddings**
*Henrik Voigt, Monique Meuschke, Sina Zarrieß and Kai Lawonn* 11:00-12:30 (Atrium)
Although contextualized word embeddings have led to great improvements in automatic language understanding, their potential for practical applications in document exploration and visualization has been little explored. Common visualization techniques used for, e.g., model analysis usually provide simple scatter plots of token-level embeddings that do not provide insight into their contextual use. In this work, we propose KeywordScape, a visual exploration tool that allows to overview, summarize, and explore the semantic content of documents based on their keywords. While existing keyword-based exploration tools assume that keywords have static meanings, our tool represents keywords in terms of their contextualized embeddings. Our application visualizes these embeddings in a semantic landscape that represents keywords as islands on a spherical map. This keeps keywords with similar context close to each other, allowing for a more precise search and comparison of documents.

[DEMO] **Arabic Word-level Readability Visualization for Assisted Text Simplification**
*Reem Hazim, Hind Saddiki, Bashar Alhafni, Muhamed Al Khalil and Nizar Habash* 11:00-12:30 (Atrium)
This demo paper presents a Google Docs add-on for automatic Arabic word-level readability visualization. The add-on includes a lemmatization component that is connected to a five-level readability lexicon and Arabic WordNet-based substitution suggestions. The add-on can be used for assessing the reading difficulty of a text and identifying difficult words as part of the task of manual text simplification. We make our add-on and its code publicly available.

[DEMO] **LogiTorch: A PyTorch-based library for logical reasoning on natural language**
*Chadi Helwe, Chloé Clavel and Fabian Suchanek* 11:00-12:30 (Atrium)
Logical reasoning on natural language is one of the most challenging tasks for deep learning models. There has been an increasing interest in developing new benchmarks to evaluate the reasoning capabilities of language models such as BERT. In parallel, new models based on transformers have emerged to achieve ever better performance on these datasets. However, there is currently no library for logical reasoning that includes such benchmarks and models.This paper introduces LogiTorch, a PyTorch-based library that includes different logical reasoning benchmarks, different models, as well as utility functions such as co-reference resolution. This makes it easy to directly use the preprocessed datasets, to run the models, or to finetune them with different hyperparameters. LogiTorch is open source and can be found on GitHub.

[DEMO] **Paraphrastic Representations at Scale**
*John Wieting, Kevin Gimpel, Graham Neubig and Taylor Berg-Kirkpatrick* 11:00-12:30 (Atrium)
We present a system that allows users to train their own state-of-the-art paraphrastic sentence representations in a variety of languages. We release trained models for English, Arabic, German, Spanish, French, Russian, Turkish, and Chinese. We train these models on large amounts of data, achieving significantly improved performance from our original papers on a suite of monolingual semantic similarity, cross-lingual semantic similarity, and bitext mining tasks. Moreover, the resulting models surpass all prior work on efficient unsupervised semantic textual similarity, even significantly outperforming supervised BERT-based models like Sentence-BERT (Reimers and Gurevych, 2019). Most importantly, our models are orders of magnitude faster than other strong similarity models and can be used on CPU with little difference in inference speed (even improved speed over GPU when using more CPU cores), making these models an attractive choice for users without access to GPUs or for use on embedded devices. Finally, we add significantly increased functionality to the code bases for training paraphrastic sentence models, easing their use for both inference and for training them for any desired language with parallel data. We also include code to automatically download and preprocess training data.

[DEMO] **KGI: An Integrated Framework for Knowledge Intensive Language Tasks**
*Md Faisal Mahbub Chowdhury, Michael Glass, Gaetano Rossiello, Alfio Gliozzo and Nandana Mihindukulasooriya* 11:00-12:30 (Atrium)
In this paper, we present a system to showcase the capabilities of the latest state-of-the-art retrieval augmented generation models trained on knowledge-intensive language tasks, such as slot filling, open domain question answering, dialogue, and fact-checking. Moreover, given a user query, we show how the output from these different models can be combined to cross-examine the outputs of each other. Particularly, we show how accuracy in dialogue can be improved using the question answering model. We are also releasing all models used in the demo as a contribution of this paper. A short video demonstrating the system is available at https://ibm.box.com/v/emnlp2022-demo.

# Session 14 - 15:30-17:00

## Virtual Portal 13

15:30-17:00 (Hall A, Room A)

**SpanProto: A Two-stage Span-based Prototypical Network for Few-shot Named Entity Recognition**
*Jianing Wang, Chengyu Wang, Chuanqi Tan, Minghui Qiu, Songfang Huang, jun huang and Ming Gao* 15:30-17:00 (Hall A, Room A)
Few-shot Named Entity Recognition (NER) aims to identify named entities with very little annotated data. Previous methods solve this problem based on token-wise classification, which ignores the information of entity boundaries, and inevitably the performance is affected by the massive non-entity tokens. To this end, we propose a seminal span-based prototypical network (SpanProto) that tackles few-shot NER via a two-stage approach, including span extraction and mention classification. In the span extraction stage, we transform the sequential tags into a global boundary matrix, enabling the model to focus on the explicit boundary information. For mention classification, we leverage prototypical learning to capture the semantic representations for each labeled span and make the model better adapt to novel-class entities. To further improve the model performance, we split out the false positives generated by the span extractor but not labeled in the current episode set, and then present a margin-based loss to separate them from each prototype region. Experiments over multiple benchmarks demonstrate that our model outperforms strong baselines by a large margin.

**RelU-Net: Syntax-aware Graph U-Net for Relational Triple Extraction**
*Yunqi Zhang, Yubo Chen and Yongfeng Huang* 15:30-17:00 (Hall A, Room A)
Relational triple extraction is a critical task for natural language processing. Existing methods mainly focused on capturing semantic information, but suffered from ignoring the syntactic structure of the sentence, which is proved in the relation classification task to contain rich relational information. This is due to the absence of entity locations, which is the prerequisite for pruning noisy edges from the dependency tree, when extracting relational triples. In this paper, we propose a unified framework to tackle this challenge and incorporate syntactic information for relational triple extraction. First, we propose to automatically contract the dependency tree into a core relational topology and eliminate redundant information with graph pooling operations. Then, we propose a symmetrical expanding path with graph unpooling operations to fuse the contracted core syntactic interactions with the original sentence context. We also propose a bipartite graph matching objective function to capture the reflections between the core topology and golden relational facts. Since our model shares similar contracting and expanding paths with encoder-decoder models like U-Net, we name our model as Relation U-Net (RelU-Net). We conduct experiments on several datasets and the results prove the effectiveness of our method.

**Concadia: Towards Image-Based Text Generation with a Purpose**
*Elisa Kreiss, Fei Fang, Noah Goodman and Christopher Potts* 15:30-17:00 (Hall A, Room A)
Current deep learning models often achieve excellent results on benchmark image-to-text datasets but fail to generate texts that are useful in practice. We argue that to close this gap, it is vital to distinguish descriptions from captions based on their distinct communicative roles. Descriptions focus on visual features and are meant to replace an image (often to increase accessibility), whereas captions appear alongside an image to supply additional information. To motivate this distinction and help people put it into practice, we introduce the publicly available Wikipedia-based dataset Concadia consisting of 96,918 images with corresponding English-language descriptions, captions, and surrounding context. Using insights from Concadia, models trained on it, and a preregistered human-subjects experiment with human- and model-generated texts, we characterize the commonalities and differences between descriptions and captions. In addition, we show that, for generating both descriptions and captions, it is useful to augment image-to-text models with representations of the textual context in which the image appeared.

**Wider & Closer: Mixture of Short-channel Distillers for Zero-shot Cross-lingual Named Entity Recognition**
*Jun-Yu Ma, Beiduo Chen, Jia-Chen Gu, Zhenhua Ling, Wu Guo, Quan Liu, Zhigang Chen and Cong Liu* 15:30-17:00 (Hall A, Room A)
Zero-shot cross-lingual named entity recognition (NER) aims at transferring knowledge from annotated and rich-resource data in source languages to unlabeled and lean-resource data in target languages. Existing mainstream methods based on the teacher-student distillation framework ignore the rich and complementary information lying in the intermediate layers of pre-trained language models, and domain-invariant information is easily lost during transfer. In this study, a mixture of short-channel distillers (MSD) method is proposed to fully interact the rich hierarchical information in the teacher model and to transfer knowledge to the student model sufficiently and efficiently. Concretely, a multi-channel distillation framework is designed for sufficient information transfer by aggregating multiple distillers as a mixture. Besides, an unsupervised method adopting parallel domain adaptation is proposed to shorten the channels between the teacher and student models to preserve domain-invariant features. Experiments on four datasets across nine languages demonstrate that the proposed method achieves new state-of-the-art performance on zero-shot cross-lingual NER and shows great generalization and compatibility across languages and fields.

**Learning Robust Representations for Continual Relation Extraction via Adversarial Class Augmentation**
*Peiyi Wang, Yifan Song, Tianyu Liu, Binghuai Lin, Yunbo Cao, Sujian Li and Zhifang Sui* 15:30-17:00 (Hall A, Room A)
Continual relation extraction (CRE) aims to continually learn new relations from a class-incremental data stream. CRE model usually suffers from catastrophic forgetting problem, i.e., the performance of old relations seriously degrades when the model learns new relations. Most previous work attributes catastrophic forgetting to the corruption of the learned representations as new relations come, with an implicit assumption that the CRE models have adequately learned the old relations. In this paper, through empirical studies we argue that this assumption may not hold, and an important reason for catastrophic forgetting is that the learned representations do not have good robustness against the appearance of analogous relations in the subsequent learning process. To address this issue, we encourage the model to learn more precise and robust representations through a simple yet effective adversarial class augmentation mechanism (ACA), which is easy to implement and model-agnostic. Experimental results show that ACA can consistently improve the performance of state-of-the-art CRE models on two popular benchmarks.

**UniRel: Unified Representation and Interaction for Joint Relational Triple Extraction**
*Wei Tang, Benfeng Xu, Yuyue Zhao, Zhendong Mao, Yifeng Liu, Yong Liao and Haiyong Xie* 15:30-17:00 (Hall A, Room A)
Relational triple extraction is challenging for its difficulty in capturing rich correlations between entities and relations. Existing works suffer from 1) heterogeneous representations of entities and relations, and 2) heterogeneous modeling of entity-entity interactions and entity-relation interactions. Therefore, the rich correlations are not fully exploited by existing works. In this paper, we propose UniRel to address these challenges. Specifically, we unify the representations of entities and relations by jointly encoding them within a concatenated natural language sequence, and unify the modeling of interactions with a proposed Interaction Map, which is built upon the off-the-shelf self-attention mechanism within any Transformer block. With comprehensive experiments on two popular relational triple extraction datasets, we demonstrate that UniRel is more effective and computationally efficient. The source code is available at https://github.com/wtangdev/UniRel.

**MetaTKG: Learning Evolutionary Meta-Knowledge for Temporal Knowledge Graph Reasoning**
*Yuwei Xia, Mengqi Zhang, Qiang Liu, Shu Wu and Xiao-Yu Zhang* 15:30-17:00 (Hall A, Room A)
Reasoning over Temporal Knowledge Graphs (TKGs) aims to predict future facts based on given history. One of the key challenges for prediction is to learn the evolution of facts. Most existing works focus on exploring evolutionary information in history to obtain effective temporal embeddings for entities and relations, but they ignore the variation in evolution patterns of facts, which makes them struggle to adapt to future data with different evolution patterns. Moreover, new entities continue to emerge along with the evolution of facts over time. Since existing models highly rely on historical information to learn embeddings for entities, they perform poorly on such entities with little historical information. To tackle these issues, we propose a novel Temporal Meta-learning framework for TKG reasoning, MetaTKG for brevity. Specifically, our method regards TKG prediction as many temporal meta-tasks, and utilizes the designed Temporal Meta-learner to learn evolutionary meta-knowledge from these meta-tasks. The proposed method aims to guide the backbones to learn to adapt quickly to future data and deal with entities with little historical information by the learned meta-knowledge. Specially, in temporal meta-learner, we design a Gating Integration module to adaptively establish temporal correlations between meta-tasks. Extensive experiments on four widely-used datasets and three backbones demonstrate that our method can greatly improve the performance.

**Q-TOD: A Query-driven Task-oriented Dialogue System**
*Xin Tian, Yingzhan Lin, Mengfei Song, Siqi Bao, Fan Wang, Huang He, Shuqi SUN and Hua Wu* 15:30-17:00 (Hall A, Room A)
Existing pipelined task-oriented dialogue systems usually have difficulties adapting to unseen domains, whereas end-to-end systems are plagued by large-scale knowledge bases in practice. In this paper, we introduce a novel query-driven task-oriented dialogue system, namely Q-TOD. The essential information from the dialogue context is extracted into a query, which is further employed to retrieve relevant knowledge records for response generation. Firstly, as the query is in the form of natural language and not confined to the schema of the knowledge base, the issue of domain adaption is alleviated remarkably in Q-TOD. Secondly, as the query enables the decoupling of knowledge retrieval from the generation, Q-TOD gets rid of the issue of knowledge base scalability. To evaluate the effectiveness of the proposed Q-TOD, we collect query annotations for three publicly available task-oriented dialogue datasets. Comprehensive experiments verify that Q-TOD outperforms strong baselines and establishes a new state-of-the-art performance on these datasets.

**Dial2vec: Self-Guided Contrastive Learning of Unsupervised Dialogue Embeddings**
*Che Liu, Rui Wang, Junfeng Jiang, Yongbin Li and Fei Huang* 15:30-17:00 (Hall A, Room A)
In this paper, we introduce the task of learning unsupervised dialogue embeddings. Trivial approaches such as combining pre-trained word or sentence embeddings and encoding through pre-trained language models (PLMs) have been shown to be feasible for this task. However, these approaches typically ignore the conversational interactions between interlocutors, resulting in poor performance. To address this issue, we proposed a self-guided contrastive learning approach named dial2vec. Dial2vec considers a dialogue as an information exchange process. It captures the interaction patterns between interlocutors and leverages them to guide the learning of the embeddings corresponding to each interlocutor. Then the dialogue embedding is obtained by an aggregation of the embeddings from all interlocutors. To verify our approach, we establish a comprehensive benchmark consisting of six widely-used dialogue datasets. We consider three evaluation tasks: domain categorization, semantic relatedness, and dialogue retrieval. Dial2vec achieves on average 8.7, 9.0, and 13.8 points absolute improvements in terms of purity, Spearman's correlation, and mean average precision (MAP) over the strongest baseline on the three tasks respectively. Further analysis shows that dial2vec obtains informative and discriminative embeddings for both interlocutors under the guidance of the conversational interactions and achieves the best performance when aggregating them through the interlocutor-level pooling strategy. All codes and data are publicly available at https://github.com/AlibabaResearch/DAMO-ConvAI/tree/main/dial2vec.

**Graph Hawkes Transformer for Extrapolated Reasoning on Temporal Knowledge Graphs**
*Haohai Sun, Shangyi Geng, Jialun Zhong, Han Hu and Kun He* 15:30-17:00 (Hall A, Room A)
Temporal Knowledge Graph (TKG) reasoning has attracted increasing attention due to its enormous potential value, and the critical issue is how to model the complex temporal structure information effectively. Recent studies use the method of encoding graph snapshots into hidden vector space and then performing heuristic deductions, which perform well on the task of entity prediction. However, these approaches cannot predict when an event will occur and have the following limitations: 1) there are many facts not related to the query that can confuse the model; 2) there exists information forgetting caused by long-term evolutionary processes. To this end, we propose a Graph Hawkes Transformer (GHT) for both TKG entity prediction and time prediction tasks in the future time. In GHT, there are two variants of Transformer, which capture the instantaneous structural information and temporal evolution information, respectively, and a new relational continuous-time encoding function to facilitate feature evolution with the Hawkes process. Extensive experiments on four public datasets demonstrate its superior performance, especially on long-term evolutionary tasks.

**Enhancing Joint Multiple Intent Detection and Slot Filling with Global Intent-Slot Co-occurrence**
*Mengxiao Song, Bowen Yu, Li Quangang, Wang Yubin, Tingwen Liu and Hongbo Xu* 15:30-17:00 (Hall A, Room A)
Multi-intent detection and slot filling joint model attracts more and more attention since it can handle multi-intent utterances, which is closer to complex real-world scenarios. Most existing joint models rely entirely on the training procedure to obtain the implicit correlation between intents and slots. However, they ignore the fact that leveraging the rich global knowledge in the corpus can determine the intuitive and explicit correlation between intents and slots. In this paper, we aim to make full use of the statistical co-occurrence frequency between intents and slots as prior knowledge to enhance joint multiple intent detection and slot filling. To be specific, an intent-slot co-occurrence graph is constructed based on the entire training corpus to globally discover correlation between intents and slots. Based on the global intent-slot co-occurrence, we propose a novel graph neural network to model the interaction between the two subtasks. Experimental results on two public multi-intent datasets demonstrate that our approach outperforms the state-of-the-art models.

**IELM: An Open Information Extraction Benchmark for Pre-Trained Language Models**
*Chenguang Wang, Xiao Liu and Dawn Song* 15:30-17:00 (Hall A, Room A)
We introduce a new open information extraction (OIE) benchmark for pre-trained language models (LM). Recent studies have demonstrated that pre-trained LMs, such as BERT and GPT, may store linguistic and relational knowledge. In particular, LMs are able to answer "fill-in-the-blank" questions when given a pre-defined relation category. Instead of focusing on pre-defined relations, we create an OIE benchmark aiming to fully examine the open relational information present in the pre-trained LMs. We accomplish this by turning pre-trained LMs into zero-shot OIE systems. Surprisingly, pre-trained LMs are able to obtain competitive performance on both standard OIE datasets (CaRB and Re-OIE2016) and two new large-scale factual OIE datasets (TAC KBP-OIE and Wikidata-OIE) that we establish via distant supervision. For instance, the zero-shot pre-trained LMs outperform the F1 score of the state-of-the-art supervised OIE methods on our factual OIE datasets without needing to use any training sets.

**ACENet: Attention Guided Commonsense Reasoning on Hybrid Knowledge Graph**
*Chuzhan Hao, Minghui Xie and Peng Zhang* 15:30-17:00 (Hall A, Room A)
Augmenting pre-trained language models (PLMs) with knowledge graphs (KGs) has demonstrated superior performance on commonsense reasoning. Given a commonsense based QA context (question and multiple choices), existing approaches usually estimate the plausibility of

candidate choices separately based on their respective retrieved KGs, without considering the interference among different choices. In this paper, we propose an Attention guided Commonsense rEasoning Network (ACENet)[5] to endow the neural network with the capability of integrating hybrid knowledge. Specifically, our model applies the multi-layer interaction of answer choices to continually strengthen correct choice information and guide the message passing of GNN. In addition, we also design a mix attention mechanism of nodes and edges to iteratively select supporting evidence on hybrid knowledge graph. Experimental results demonstrate the effectiveness of our proposed model through considerable performance gains across CommonsenseQA and OpenbookQA datasets.

**IRRGN: An Implicit Relational Reasoning Graph Network for Multi-turn Response Selection**
*Jingcheng Deng, Hengwei Dai, Xuewei Guo, Yuanchen Ju and Wei Peng* 15:30-17:00 (Hall A, Room A)
The task of response selection in multi-turn dialogue is to find the best option from all candidates. In order to improve the reasoning ability of the model, previous studies pay more attention to using explicit algorithms to model the dependencies between utterances, which are deterministic, limited and inflexible. In addition, few studies consider differences between the options before and after reasoning. In this paper, we propose an Implicit Relational Reasoning Graph Network to address these issues, which consists of the Utterance Relational Reasoner (URR) and the Option Dual Comparator (ODC). URR aims to implicitly extract dependencies between utterances, as well as utterances and options, and make reasoning with relational graph convolutional networks. ODC focuses on perceiving the difference between the options through dual comparison, which can eliminate the interference of the noise options. Experimental results on two multi-turn dialogue reasoning benchmark datasets MuTual and MuTualplus show that our method significantly improves the baseline of four pre-trained language models and achieves state-of-the-art performance. The model surpasses human performance for the first time on the MuTual dataset.

**Predicting Prerequisite Relations for Unseen Concepts**
*Yaxin Zhu and Hamed Zamani* 15:30-17:00 (Hall A, Room A)
Concept prerequisite learning (CPL) plays a key role in developing technologies that assist people to learn a new complex topic or concept. Previous work commonly assumes that all concepts are given at training time and solely focuses on predicting the unseen prerequisite relationships between them. However, many real-world scenarios deal with concepts that are left undiscovered at training time, which is relatively unexplored. This paper studies this problem and proposes a novel alternating knowledge distillation approach to take advantage of both content- and graph-based models for this task. Extensive experiments on three public benchmarks demonstrate up to 10% improvements in terms of F1 score.

**Boosting Document-Level Relation Extraction by Mining and Injecting Logical Rules**
*Shengda Fan, Shasha Mo and Jianwei Niu* 15:30-17:00 (Hall A, Room A)
Document-level relation extraction (DocRE) aims at extracting relations of all entity pairs in a document. A key challenge to DocRE lies in the complex interdependency between the relations of entity pairs. Unlike most prior efforts focusing on implicitly powerful representations, the recently proposed LogiRE (Ru et al., 2021) explicitly captures the interdependency by learning logical rules. However, LogiRE requires extra parameterized modules to reason merely after training backbones, and this disjointed optimization of backbones and extra modules may lead to sub-optimal results. In this paper, we propose MILR, a logic enhanced framework that boosts DocRE by Mining and Injecting Logical Rules. MILR first mines logical rules from annotations based on frequencies. Then in training, consistency regularization is leveraged as an auxiliary loss to penalize instances that violate mined rules. Finally, MILR infers from a global perspective based on integer programming. Compared with LogiRE, MILR does not introduce extra parameters and injects logical rules during both training and inference. Extensive experiments on two benchmarks demonstrate that MILR not only improves the relation extraction performance (1.1%-3.8% F1) but also makes predictions more logically consistent (over 4.5% Logic). More importantly, MILR also consistently outperforms LogiRE on both counts. Code is available at https://github.com/XingYing-stack/MILR.

**Towards relation extraction from speech**
*Tongtong Wu, Guitao Wang, Jinming Zhao, Zhaoran Liu, Guilin Qi, Yuan-Fang Li and Gholamreza Haffari* 15:30-17:00 (Hall A, Room A)
Relation extraction typically aims to extract semantic relationships between entities from the unstructured text. One of the most essential data sources for relation extraction is the spoken language, such as interviews and dialogues. However, the error propagation introduced in automatic speech recognition (ASR) has been ignored in relation extraction, and the end-to-end speech-based relation extraction method has been rarely explored. In this paper, we propose a new listening information extraction task, i.e., speech relation extraction. We construct the training dataset for speech relation extraction via text-to-speech systems, and we construct the testing dataset via crowd-sourcing with native English speakers. We explore speech relation extraction via two approaches: the pipeline approach conducting text-based extraction with a pretrained ASR module, and the end2end approach via a new proposed encoder-decoder model, or what we called SpeechRE. We conduct comprehensive experiments to distinguish the challenges in speech relation extraction, which may shed light on future explorations. We share the code and data on https://github.com/wutong8023/SpeechRE.

## Virtual Portal 14

15:30-17:00 (Hall A, Room B)

**ROSE: Robust Selective Fine-tuning for Pre-trained Language Models**
*Lan Jiang, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou and Rui Jiang* 15:30-17:00 (Hall A, Room B)
Even though the large-scale language models have achieved excellent performances, they suffer from various adversarial attacks. A large body of defense methods has been proposed. However, they are still limited due to redundant attack search spaces and the inability to defend against various types of attacks. In this work, we present a novel fine-tuning approach called **RO**bust **SE**letive fine-tuning (**ROSE**) to address this issue. ROSE conducts selective updates when adapting pre-trained models to downstream tasks, filtering out invaluable and unrobust updates of parameters. Specifically, we propose two strategies: the first-order and second-order ROSE for selecting target robust parameters. The experimental results show that ROSE achieves significant improvements in adversarial robustness on various downstream NLP tasks, and the ensemble method even surpasses both variants above. Furthermore, ROSE can be easily incorporated into existing fine-tuning methods to improve their adversarial robustness further. The empirical analysis confirms that ROSE eliminates unrobust spurious updates during fine-tuning, leading to solutions corresponding to flatter and wider optima than the conventional method. Code is available at `https://github.com/jiangllan/ROSE`.

**CodeRetriever: A Large Scale Contrastive Pre-Training Method for Code Search**
*Xiaonan Li, Yeyun Gong, Yelong Shen, Xipeng Qiu, Hang Zhang, Bolun Yao, Weizhen Qi, Daxin Jiang, Weizhu Chen and Nan Duan* 15:30-

[5] `https://github.com/HAOChuzhan/ACENet`.

17:00 (Hall A, Room B)
In this paper, we propose the CodeRetriever model, which learns the function-level code semantic representations through large-scale code-text contrastive pre-training. We adopt two contrastive learning schemes in CodeRetriever: unimodal contrastive learning and bimodal contrastive learning. For unimodal contrastive learning, we design an unsupervised learning approach to build semantic-related code pairs based on the documentation and function name. For bimodal contrastive learning, we leverage the documentation and in-line comments of code to build code-text pairs. Both contrastive objectives can fully leverage large-scale code corpus for pre-training. Extensive experimental results show that CodeRetriever achieves new state-of-the-art with significant improvement over existing code pre-trained models, on eleven domain/language-specific code search tasks with six programming languages in different code granularity (function-level, snippet-level and statement-level). These results demonstrate the effectiveness and robustness of CodeRetriever. The codes and resources are available at `https://github.com/microsoft/AR2/tree/main/CodeRetriever`.

**Candidate Soups: Fusing Candidate Results Improves Translation Quality for Non-Autoregressive Translation**
*Huanran Zheng, Wei Zhu, Pengfei Wang and Xiaoling Wang* 15:30-17:00 (Hall A, Room B)
Non-autoregressive translation (NAT) model achieves a much faster inference speed than the autoregressive translation (AT) model because it can simultaneously predict all tokens during inference. However, its translation quality suffers from degradation compared to AT. And existing NAT methods only focus on improving the NAT model's performance but do not fully utilize it. In this paper, we propose a simple but effective method called "Candidate Soups," which can obtain high-quality translations while maintaining the inference speed of NAT models. Unlike previous approaches that pick the individual result and discard the remainders, Candidate Soups (CDS) can fully use the valuable information in the different candidate translations through model uncertainty. Extensive experiments on two benchmarks (WMT'14 EN–DE and WMT'16 EN–RO) demonstrate the effectiveness and generality of our proposed method, which can significantly improve the translation quality of various base models. More notably, our best variant outperforms the AT model on three translation tasks with 7.6× speedup.

**Exploring Representation-level Augmentation for Code Search**
*Haochen Li, Chunyan Miao, Cyril Leung, Yanxian Huang, Yuan Huang, Hongyu Zhang and Yanlin Wang* 15:30-17:00 (Hall A, Room B)
Code search, which aims at retrieving the most relevant code fragment for a given natural language query, is a common activity in software development practice. Recently, contrastive learning is widely used in code search research, where many data augmentation approaches for source code (e.g., semantic-preserving program transformation) are proposed to learn better representations. However, these augmentations are at the raw-data level, which requires additional code analysis in the preprocessing stage and additional training cost in the training stage. In this paper, we explore augmentation methods that augment data (both code and query) at representation level which does not require additional data processing and training, and based on this we propose a general format of representation-level augmentation that unifies existing methods. Then, we propose three new augmentation methods (linear extrapolation, binary interpolation, and Gaussian scaling) based on the general format. Furthermore, we theoretically analyze the advantages of the proposed augmentation methods over traditional contrastive learning methods on code search. We experimentally evaluate the proposed representation-level augmentation methods with state-of-the-art code search models on a large-scale public dataset consisting of six programming languages. The experimental results show that our approach can consistently boost the performance of the studied code search models.

**COPEN: Probing Conceptual Knowledge in Pre-trained Language Models**
*Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu and Qun Liu* 15:30-17:00 (Hall A, Room B)
Conceptual knowledge is fundamental to human cognition and knowledge bases. However, existing knowledge probing works only focus on evaluating factual knowledge of pre-trained language models (PLMs) and ignore conceptual knowledge. Since conceptual knowledge often appears as implicit commonsense behind texts, designing probes for conceptual knowledge is hard. Inspired by knowledge representation schemata, we comprehensively evaluate conceptual knowledge of PLMs by designing three tasks to probe whether PLMs organize entities by conceptual similarities, learn conceptual properties, and conceptualize entities in contexts, respectively. For the tasks, we collect and annotate 24k data instances covering 393 concepts, which is COPEN, a COnceptual knowledge Probing bENchmark. Extensive experiments on different sizes and types of PLMs show that existing PLMs systematically lack conceptual knowledge and suffer from various spurious correlations. We believe this is a critical bottleneck for realizing human-like cognition in PLMs. COPEN and our codes are publicly released at https://github.com/THU-KEG/COPEN.

**Towards Robust k-Nearest-Neighbor Machine Translation**
*Hui Jiang, Ziyao Lu, Fandong Meng, Chulun Zhou, Jie Zhou, Degen Huang and Jinsong Su* 15:30-17:00 (Hall A, Room B)
k-Nearest-Neighbor Machine Translation (kNN-MT) becomes an important research direction of NMT in recent years. Its main idea is to retrieve useful key-value pairs from an additional datastore to modify translations without updating the NMT model. However, the underlying retrieved noisy pairs will dramatically deteriorate the model performance. In this paper, we conduct a preliminary study and find that this problem results from not fully exploiting the prediction of the NMT model. To alleviate the impact of noise, we propose a confidence-enhanced kNN-MT model with robust training. Concretely, we introduce the NMT confidence to refine the modeling of two important components of kNN-MT: kNN distribution and the interpolation weight. Meanwhile we inject two types of perturbations into the retrieved pairs for robust training. Experimental results on four benchmark datasets demonstrate that our model not only achieves significant improvements over current kNN-MT models, but also exhibits better robustness. Our code is available at https://github.com/DeepLearnXMU/Robust-knn-mt.

**A Survey of Active Learning for Natural Language Processing**
*Zhisong Zhang, Emma Strubell and Eduard Hovy* 15:30-17:00 (Hall A, Room B)
In this work, we provide a literature review of active learning (AL) for its applications in natural language processing (NLP). In addition to a fine-grained categorization of query strategies, we also investigate several other important aspects of applying AL to NLP problems. These include AL for structured prediction tasks, annotation cost, model learning (especially with deep neural models), and starting and stopping AL. Finally, we conclude with a discussion of related topics and future directions.

**Unifying the Convergences in Multilingual Neural Machine Translation**
*yichong huang, Xiaocheng Feng, Xinwei Geng and Bing Qin* 15:30-17:00 (Hall A, Room B)
Although all-in-one-model multilingual neural machine translation (MNMT) has achieved remarkable progress, the convergence inconsistency in the joint training is ignored, i.e., different language pairs reaching convergence in different epochs. This leads to the trained MNMT model over-fitting low-resource language translations while under-fitting high-resource ones. In this paper, we propose a novel training strategy named LSSD (LanguageSpecific Self-Distillation), which can alleviate the convergence inconsistency and help MNMT models achieve the best performance on each language pair simultaneously. Specifically, LSSD picks up language-specific best checkpoints for each language pair to teach the current model on the fly. Furthermore, we systematically explore three sample-level manipulations of knowledge transferring. Experimental results on three datasets show that LSSD obtains consistent improvements towards all language pairs and achieves the state-of-the-art.

**The Devil in Linear Transformer**
*Zhen Qin, Xiaodong Han, Weixuan Sun, Dongxu Li, Lingpeng Kong, Nick Barnes and Yiran Zhong* 15:30-17:00 (Hall A, Room B)

Linear transformers aim to reduce the quadratic space-time complexity of vanilla transformers. However, they usually suffer from degraded performances on various tasks and corpus. In this paper, we examine existing kernel-based linear transformers and identify two key issues that lead to such performance gaps: 1) unbounded gradients in the attention computation adversely impact the convergence of linear transformer models; 2) attention dilution which trivially distributes attention scores over long sequences while neglecting neighbouring structures. To address these issues, we first identify that the scaling of attention matrices is the devil in unbounded gradients, which turns out unnecessary in linear attention as we show theoretically and empirically. To this end, we propose a new linear attention that replaces the scaling operation with a normalization to stabilize gradients. For the issue of attention dilution, we leverage a diagonal attention to confine attention to only neighbouring tokens in early layers. Benefiting from the stable gradients and improved attention, our new linear transformer model, transNormer, demonstrates superior performance on text classification and language modeling tasks, as well as on the challenging Long-Range Arena benchmark, surpassing vanilla transformer and existing linear variants by a clear margin while being significantly more space-time efficient. The code is available at https://github.com/OpenNLPLab/Transnormer .

**Zero-Shot Learners for Natural Language Understanding via a Unified Multiple Choice Perspective**
*Ping Yang, Junjie Wang, Ruyi Gan, Xinyu Zhu, Lin Zhang, Ziwei Wu, Xinyu Gao, Jiaxing Zhang and Tetsuya Sakai* 15:30-17:00 (Hall A, Room B)
We propose a new paradigm for zero-shot learners that is format agnostic, i.e., it is compatible with any format and applicable to a list of language tasks, such as text classification, commonsense reasoning, coreference resolution, and sentiment analysis. Zero-shot learning aims to train a model on a given task such that it can address new learning tasks without any additional training. Our approach converts zero-shot learning into multiple-choice tasks, avoiding problems in commonly used large-scale generative models such as FLAN. It not only adds generalization ability to models but also significantly reduces the number of parameters. Our method shares the merits of efficient training and deployment. Our approach shows state-of-the-art performance on several benchmarks and produces satisfactory results on tasks such as natural language inference and text classification. Our model achieves this success with only 235M parameters, which is substantially smaller than state-of-the-art models with billions of parameters. The code and pre-trained models are available at https://github.com/IDEA-CCNL/Fengshenbang-LM/tree/main/fengshen/examples/unimc .

**Hypoformer: Hybrid Decomposition Transformer for Edge-friendly Neural Machine Translation**
*sunzhu li, Peng Zhang, Guobing Gan, Xiuqing Lv, Benyou Wang, Junqiu Wei and Xin Jiang* 15:30-17:00 (Hall A, Room B)
Transformer has been demonstrated effective in Neural Machine Translation (NMT). However, it is memory-consuming and time-consuming in edge devices, resulting in some difficulties for real-time feedback. To compress and accelerate Transformer, we propose a Hybrid Tensor-Train (HTT) decomposition, which retains full rank and meanwhile reduces operations and parameters. A Transformer using HTT, named Hypoformer, consistently and notably outperforms the recent light-weight SOTA methods on three standard translation tasks under different parameter and speed scales. In extreme low resource scenarios, Hypoformer has 7.1 points absolute improvement in BLEU and 1.27 X speedup than vanilla Transformer on IWSLT'14 De-En task.

**STGN: an Implicit Regularization Method for Learning with Noisy Labels in Natural Language Processing**
*Tingting Wu, Xiao Ding, Minji Tang, Hao Zhang, Bing Qin and Ting Liu* 15:30-17:00 (Hall A, Room B)
Noisy labels are ubiquitous in natural language processing (NLP) tasks. Existing work, namely learning with noisy labels in NLP, is often limited to dedicated tasks or specific training procedures, making it hard to be widely used. To address this issue, SGD noise has been explored to provide a more general way to alleviate the effect of noisy labels by involving benign noise in the process of stochastic gradient descent. However, previous studies exert identical perturbation for all samples, which may cause overfitting on incorrect ones or optimizing correct ones inadequately. To facilitate this, we propose a novel stochastic tailor-made gradient noise (STGN), mitigating the effect of inherent label noise by introducing tailor-made benign noise for each sample. Specifically, we investigate multiple principles to precisely and stably discriminate correct samples from incorrect ones and thus apply different intensities of perturbation to them. A detailed theoretical analysis shows that STGN has good properties, beneficial for model generalization. Experiments on three different NLP tasks demonstrate the effectiveness and versatility of STGN. Also, STGN can boost existing robust training methods.

**Character-level White-Box Adversarial Attacks against Transformers via Attachable Subwords Substitution**
*Aiwei Liu, Honghai Yu, Xuming Hu, Shu'ang Li, Li Lin, Fukun Ma, Yawen Yang and Lijie Wen* 15:30-17:00 (Hall A, Room B)
We propose the first character-level white-box adversarial attack method against transformer models. The intuition of our method comes from the observation that words are split into subtokens before being fed into the transformer models and the substitution between two close subtokens has a similar effect with the character modification. Our method mainly contains three steps. First, a gradient-based method is adopted to find the most vulnerable words in the sentence. Then we split the selected words into subtokens to replace the origin tokenization result from the transformer tokenizer. Finally, we utilize an adversarial loss to guide the substitution of attachable subtokens in which the Gumbel-softmax trick is introduced to ensure gradient propagation. Meanwhile, we introduce the visual and length constraint in the optimization process to achieve minimum character modifications. Extensive experiments on both sentence-level and token-level tasks demonstrate that our method could outperform the previous attack methods in terms of success rate and edit distance. Furthermore, human evaluation verifies our adversarial examples could preserve their origin labels.

**Cross-Linguistic Syntactic Difference in Multilingual BERT: How Good is It and How Does It Affect Transfer?**
*Ningyu Xu, Tao Gui, Ruotian Ma, Qi Zhang, Jingting Ye, Menghan Zhang and Xuanjing Huang* 15:30-17:00 (Hall A, Room B)
Multilingual BERT (mBERT) has demonstrated considerable cross-lingual syntactic ability, whereby it enables effective zero-shot cross-lingual transfer of syntactic knowledge. The transfer is more successful between some languages, but it is not well understood what leads to this variation and whether it fairly reflects difference between languages. In this work, we investigate the distributions of grammatical relations induced from mBERT in the context of 24 typologically different languages. We demonstrate that the distance between the distributions of different languages is highly consistent with the syntactic difference in terms of linguistic formalisms. Such difference learnt via self-supervision plays a crucial role in the zero-shot transfer performance and can be predicted by variation in morphosyntactic properties between languages. These results suggest that mBERT properly encodes languages in a way consistent with linguistic diversity and provide insights into the mechanism of cross-lingual transfer.

**Learning to Explain Selectively: A Case Study on Question Answering**
*Shi Feng and Jordan Boyd-Graber* 15:30-17:00 (Hall A, Room B)
Explanations promise to bridge the gap between humans and AI, yet it remains difficult to achieve consistent improvement in AI-augmented human decision making. The usefulness of AI explanations depends on many factors, and always showing the same type of explanation in all cases is suboptimal—so is relying on heuristics to adapt explanations for each scenario. We propose learning to explain "selectively": for each decision that the user makes, we use a model to choose the best explanation from a set of candidates and update this model with feedback to optimize human performance. We experiment on a question answering task, Quizbowl, and show that selective explanations improve human performance for both experts and crowdworkers.

**Reorder and then Parse, Fast and Accurate Discontinuous Constituency Parsing**

*Kailai Sun, Zuchao Li and Hai Zhao* 15:30-17:00 (Hall A, Room B)
Discontinuous constituency parsing is still kept developing for its efficiency and accuracy are far behind its continuous counterparts. Motivated by the observation that a discontinuous constituent tree can be simply transformed into a pseudo-continuous one by artificially reordering words in the sentence, we propose a novel reordering method, thereby construct fast and accurate discontinuous constituency parsing systems working in continuous way. Specifically, we model the relative position changes of words as a list of actions. By parsing and performing this actions, the corresponding pseudo-continuous sequence is derived. Discontinuous parse tree can be further inferred via integrating a high-performance pseudo-continuous constituency parser. Our systems are evaluated on three classical discontinuous constituency treebanks, achieving new state-of-the-art on two treebanks and showing a distinct advantage in speed.

**XPrompt: Exploring the Extreme of Prompt Tuning**
*Fang Ma, Chen Zhang, Lei Ren, Jingang Wang, Qifan Wang, Wei Wu, Xiaojun Quan and Dawei Song* 15:30-17:00 (Hall A, Room B)
Prompt tuning learns soft prompts to condition the frozen Pre-trained Language Models (PLMs) for performing downstream tasks in a parameter-efficient manner. While prompt tuning has gradually reached the performance level of fine-tuning as the model scale increases, there is still a large performance gap between prompt tuning and fine-tuning for models of moderate and small scales (typically less than 11B parameters). In this paper, we empirically show that the trained prompt tokens can have a negative impact on a downstream task and thus degrade its performance. To bridge the gap, we propose a novel Prompt tuning model with an eXtremely small scale (XPrompt) under the regime of lottery tickets hypothesis. Specifically, XPrompt eliminates the negative prompt tokens at different granularity levels through a hierarchical structured pruning, yielding a more parameter-efficient prompt yet with a competitive performance. Comprehensive experiments are carried out on the SuperGLUE tasks, and the results indicate that XPrompt is able to close the performance gap at smaller model scales.

**Instance Regularization for Discriminative Language Model Pre-training**
*Zhuosheng Zhang, Hai Zhao and Ming Zhou* 15:30-17:00 (Hall A, Room B)
Discriminative pre-trained language models (PrLMs) can be generalized as denoising auto-encoders that work with two procedures, ennoising and denoising. First, an ennoising process corrupts texts with arbitrary noising functions to construct training instances. Then, a denoising language model is trained to restore the corrupted tokens. Existing studies have made progress by optimizing independent strategies of either ennoising or denosing. They treat training instances equally throughout the training process, with little attention on the individual contribution of those instances. To model explicit signals of instance contribution, this work proposes to estimate the complexity of restoring the original sentences from corrupted ones in language model pre-training. The estimations involve the corruption degree in the ennoising data construction process and the prediction confidence in the denoising counterpart. Experimental results on natural language understanding and reading comprehension benchmarks show that our approach improves pre-training efficiency, effectiveness, and robustness. Code is publicly available at https://github.com/cooelf/InstanceReg.

[INDUSTRY] **A Stacking-based Efficient Method for Toxic Language Detection on Live Streaming Chat**
*Yuto Oikawa, Yuki Nakayama and Koji Murakami* 15:30-17:00 (Hall A, Room B)
In a live streaming chat on a video streaming service, it is crucial to filter out toxic comments with online processing to prevent users from reading comments in real-time. However, recent toxic language detection methods rely on deep learning methods, which can not be scalable considering inference speed. Also, these methods do not consider constraints of computational resources expected depending on a deployed system (e.g., no GPU resource). This paper presents an efficient method for toxic language detection that is aware of real-world scenarios. Our proposed architecture is based on partial stacking that feeds initial results with low confidence to meta-classifier. Experimental results show that our method achieves a much faster inference speed than BERT-based models with comparable performance.

[INDUSTRY] **Consultation Checklists: Standardising the Human Evaluation of Medical Note Generation**
*Aleksandar Savkov, Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Anya Belz and Ehud Reiter* 15:30-17:00 (Hall A, Room B)
Evaluating automatically generated text is generally hard due to the inherently subjective nature of many aspects of the output quality. This difficulty is compounded in automatic consultation note generation by differing opinions between medical experts both about which patient statements should be included in generated notes and about their respective importance in arriving at a diagnosis. Previous real-world evaluations of note-generation systems saw substantial disagreement between expert evaluators. In this paper we propose a protocol that aims to increase objectivity by grounding evaluations in Consultation Checklists, which are created in a preliminary step and then used as a common point of reference during quality assessment. We observed good levels of inter-annotator agreement in a first evaluation study using the protocol; further, using Consultation Checklists produced in the study as reference for automatic metrics such as ROUGE or BERTScore improves their correlation with human judgements compared to using the original human note.

[INDUSTRY] **Controlled Language Generation for Language Learning Items**
*Kevin Stowe, Debanjan Ghosh and Mengxuan Zhao* 15:30-17:00 (Hall A, Room B)
This work aims to employ natural language generation (NLG) to rapidly generate items for English language learning applications: this requires both language models capable of generating fluent, high-quality English, and to control the output of the generation to match the requirements of the relevant items. We experiment with deep pretrained models for this task, developing novel methods for controlling items for factors relevant in language learning: diverse sentences for different proficiency levels and argument structure to test grammar. Human evaluation demonstrates high grammatically scores for all models (3.4 and above out of 4), and higher length (24%) and complexity (9%) over the baseline for the advanced proficiency model. Our results show that we can achieve strong performance while adding additional control to ensure diverse, tailored content for individual users.

[INDUSTRY] **Fact Checking Machine Generated Text with Dependency Trees**
*Alex Estes, Nikhita Vedula, Marcus Collins, Matt Cecil and Oleg Rokhlenko* 15:30-17:00 (Hall A, Room B)
Factual and logical errors made by Natural Language Generation (NLG) systems limit their applicability in many settings. We study this problem in a conversational search and recommendation setting, and observe that we can often make two simplifying assumptions in this domain: (i) there exists a body of structured knowledge we can use for verifying factuality of generated text; and (ii) the text to be factually assessed typically has a well-defined structure and style. Grounded in these assumptions, we propose a fast, unsupervised and explainable technique, DepChecker, that assesses factuality of input text based on rules derived from structured knowledge patterns and dependency relations with respect to the input text. We show that DepChecker outperforms state-of-the-art, general purpose fact-checking techniques in this special, but important case.

[INDUSTRY] **Distinguish Sense from Nonsense: Out-of-Scope Detection for Virtual Assistants**
*Cheng Qian, Haode Qi, Gengyu Wang, Ladislav Kunc and Saloni Potdar* 15:30-17:00 (Hall A, Room B)
Out of Scope (OOS) detection in Conversational AI solutions enables a chatbot to handle a conversation gracefully when it is unable to make sense of the end-user query. Accurately tagging a query as out-of-domain is particularly hard in scenarios when the chatbot is not equipped to handle a topic which has semantic overlap with an existing topic it is trained on. We propose a simple yet effective OOS detection method that outperforms standard OOS detection methods in a real-world deployment of virtual assistants. We discuss the various design and deploy-

ment considerations for a cloud platform solution to train virtual assistants and deploy them at scale. Additionally, we propose a collection of datasets that replicates real-world scenarios and show comprehensive results in various settings using both offline and online evaluation metrics.

[INDUSTRY] **PLATO-Ad: A Unified Advertisement Text Generation Framework with Multi-Task Prompt Learning**
*Zeyang Lei, Chao Zhang, Xinchao Xu, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, Yi Yang and Shuanglong Li*15:30-17:00 (Hall A, Room B)
Online advertisement text generation aims at generating attractive and persuasive text ads to appeal to users clicking ads or purchasing products. While pretraining-based models have achieved remarkable success in generating high-quality text ads, some challenges still remain, such as ad generation in low-resource scenarios and training efficiency for multiple ad tasks. In this paper, we propose a novel unified text ad generation framework with multi-task prompt learning, called PLATO-Ad, to tackle these problems. Specifically, we design a three-phase transfer learning mechanism to tackle the low-resource ad generation problem. Furthermore, we present a novel multi-task prompt learning mechanism to efficiently utilize a single lightweight model to solve multiple ad generation tasks without loss of performance compared to training a separate model for each task. Finally, we conduct offline and online evaluations and experiment results show that PLATO-Ad significantly outperforms the state-of-the-art on both offline and online metrics. PLATO-Ad has been deployed in a leading advertising platform with 3.5% CTR improvement on search ad descriptions and 10.4% CTR improvement on feed ad titles.

[DEMO] **stopes - Modular Machine Translation Pipelines**
*Pierre Andrews, Guillaume Wenzek, Kevin Heffernan, Onur Çelebi, Anna Sun, Ammar Kamran, Yingzhe Guo, Alexandre Mourachko, Holger Schwenk and Angela Fan* 15:30-17:00 (Hall A, Room B)
Neural machine translation, as other natural language deep learning applications, is hungry for data. As research evolves, the data pipelines supporting that research evolve too, oftentimes re-implementing the same core components. Despite the potential of modular codebases, researchers have but little time to put code structure and reusability first. Unfortunately, this makes it very hard to publish clean, reproducible code to benefit a wider audience. In this paper, we motivate and describe stopes , a framework that addresses these issues while empowering scalability and versatility for research use cases. This library was a key enabler of the No Language Left Behind project, establishing new state of the art performance for a multilingual machine translation model covering 200 languages. stopes and the pipelines described are released under the MIT license at https://github.com/facebookresearch/stopes.

## Virtual Portal 15

15:30-17:00 (Hall A, Room C)

**Open-Topic False Information Detection on Social Networks with Contrastive Adversarial Learning**
*Guanghui Ma, Chunming Hu, Ling Ge and Hong Zhang* 15:30-17:00 (Hall A, Room C)
Current works about false information detection based on conversation graphs on social networks focus primarily on two research streams from the standpoint of topic distribution: in-topic and cross-topic techniques, which assume that the data topic distribution is identical or cross, respectively. This signifies that all test data topics are seen or unseen by the model. However, these assumptions are too harsh for actual social networks that contain both seen and unseen topics simultaneously, hence restricting their practical application. In light of this, this paper develops a novel open-topic scenario that is better suited to actual social networks. In this open-topic scenario, we empirically find that the existing models suffer from impairment in the detection performance for seen or unseen topic data, resulting in poor overall model performance. To address this issue, we propose a novel Contrastive Adversarial Learning Network, CALN, that employs an unsupervised topic clustering method to capture topic-specific features to enhance the model's performance for seen topics and an unsupervised adversarial learning method to align data representation distributions to enhance the model's generalisation to unseen topics. Experiments on two benchmark datasets and a variety of graph neural networks demonstrate the effectiveness of our approach.

**Empowering Dual-Encoder with Query Generator for Cross-Lingual Dense Retrieval**
*Houxing Ren, Linjun Shou, Ning Wu, Ming Gong and Daxin Jiang* 15:30-17:00 (Hall A, Room C)
In monolingual dense retrieval, lots of works focus on how to distill knowledge from cross-encoder re-ranker to dual-encoder retriever and these methods achieve better performance due to the effectiveness of cross-encoder re-ranker. However, we find that the performance of the cross-encoder re-ranker is heavily influenced by the number of training samples and the quality of negative samples, which is hard to obtain in the cross-lingual setting. In this paper, we propose to use a query generator as the teacher in the cross-lingual setting, which is less dependent on enough training samples and high-quality negative samples. In addition to traditional knowledge distillation, we further propose a novel enhancement method, which uses the query generator to help the dual-encoder align queries from different languages, but does not need any additional parallel sentences. The experimental results show that our method outperforms the state-of-the-art methods on two benchmark datasets.

**A Joint Learning Framework for Restaurant Survival Prediction and Explanation**
*Xin Li, Xiaojie Zhang, Peng JiaHao, Rui Mao, Mingyang Zhou, Xing Xie and Hao Liao* 15:30-17:00 (Hall A, Room C)
The bloom of the Internet and the recent breakthroughs in deep learning techniques open a new door to AI for E-commence, with a trend of evolving from using a few financial factors such as liquidity and profitability to using more advanced AI techniques to process complex and multi-modal data. In this paper, we tackle the practical problem of restaurant survival prediction. We argue that traditional methods ignore two essential respects, which are very helpful for the task: 1) modeling customer reviews and 2) jointly considering status prediction and result explanation. Thus, we propose a novel joint learning framework for explainable restaurant survival prediction based on the multi-modal data of user-restaurant interactions and users' textual reviews. Moreover, we design a graph neural network to capture the high-order interactions and design a co-attention mechanism to capture the most informative and meaningful signal from noisy textual reviews. Our results on two datasets show a significant and consistent improvement over the SOTA techniques (average 6.8% improvement in prediction and 45.3% improvement in explanation).

**Towards Multi-Modal Sarcasm Detection via Hierarchical Congruity Modeling with Knowledge Enhancement**
*Hui Liu, Wenya Wang and Haoliang Li* 15:30-17:00 (Hall A, Room C)
Sarcasm is a linguistic phenomenon indicating a discrepancy between literal meanings and implied intentions. Due to its sophisticated nature, it is usually difficult to be detected from the text itself. As a result, multi-modal sarcasm detection has received more and more attention in both academia and industries. However, most existing techniques only modeled the atomic-level inconsistencies between the text input and its accompanying image, ignoring more complex compositions for both modalities. Moreover, they neglected the rich information contained in external knowledge, e.g., image captions. In this paper, we propose a novel hierarchical framework for sarcasm detection by exploring both the atomic-level congruity based on multi-head cross attentions and the composition-level congruity based on graph neural networks, where

a post with low congruity can be identified as sarcasm. In addition, we exploit the effect of various knowledge resources for sarcasm detection. Evaluation results on a public multi-modal sarcasm detection dataset based on Twitter demonstrate the superiority of our proposed model.

**MetaFill: Text Infilling for Meta-Path Generation on Heterogeneous Information Networks**
*Zequn Liu, Kefei Duan, Junwei Yang, Hanwen Xu, Ming Zhang and Sheng Wang* 15:30-17:00 (Hall A, Room C)
Heterogeneous information network (HIN) is essential to study complicated networks containing multiple edge types and node types. Meta-path, a sequence of node types and edge types, is the core technique to embed HINs. Since manually curating meta-paths is time-consuming, there is a pressing need to develop automated meta-path generation approaches. Existing meta-path generation approaches cannot fully exploit the rich textual information in HINs, such as node names and edge type names. To address this problem, we propose MetaFill, a text-infilling-based approach for meta-path generation. The key idea of MetaFill is to formulate meta-path identification problem as a word sequence infilling problem, which can be advanced by pretrained language models (PLMs). We observed the superior performance of MetaFill against existing meta-path generation methods and graph embedding methods that do not leverage meta-paths in both link prediction and node classification on two real-world HIN datasets. We further demonstrated how MetaFill can accurately classify edges in the zero-shot setting, where existing approaches cannot generate any meta-paths. MetaFill exploits PLMs to generate meta-paths for graph embedding, opening up new avenues for language model applications in graph analysis.

**TeleMelody: Lyric-to-Melody Generation with a Template-Based Two-Stage Method**
*Zeqian Ju, Peiling Lu, Xu Tan, Rui Wang, Chen Zhang, Songruoyao Wu, Kejun Zhang, Xiang-Yang Li, Tao Qin and Tie-Yan Liu* 15:30-17:00 (Hall A, Room C)
Lyric-to-melody generation is an important task in automatic songwriting. Previous lyric-to-melody generation systems usually adopt end-to-end models that directly generate melodies from lyrics, which suffer from several issues: 1) lack of paired lyric-melody training data; 2) lack of control on generated melodies. In this paper, we develop TeleMelody, a two-stage lyric-to-melody generation system with music template (e.g., tonality, chord progression, rhythm pattern, and cadence) to bridge the gap between lyrics and melodies (i.e., the system consists of a lyric-to-template module and a template-to-melody module). TeleMelody has two advantages. First, it is data efficient. The template-to-melody module is trained in a self-supervised way (i.e., the source template is extracted from the target melody) that does not need any lyric-melody paired data. The lyric-to-template module is made up of some rules and a lyric-to-rhythm model, which is trained with paired lyric-rhythm data that is easier to obtain than paired lyric-melody data. Second, it is controllable. The design of the template ensures that the generated melodies can be controlled by adjusting the musical elements in the template. Both subjective and objective experimental evaluations demonstrate that TeleMelody generates melodies with higher quality, better controllability, and less requirement on paired lyric-melody data than previous generation systems.

**Toward the Limitation of Code-Switching in Cross-Lingual Transfer**
*Yukun Feng, Feng Li and Philipp Koehn* 15:30-17:00 (Hall A, Room C)
Multilingual pretrained models have shown strong cross-lingual transfer ability. Some works used code-switching sentences, which consist of tokens from multiple languages, to enhance the cross-lingual representation further, and have shown success in many zero-shot cross-lingual tasks. However, code-switched tokens are likely to cause grammatical incoherence in newly substituted sentences, and negatively affect the performance on token-sensitive tasks, such as Part-of-Speech (POS) tagging and Named-Entity-Recognition (NER). This paper mitigates the limitation of the code-switching method by not only making the token replacement but considering the similarity between the context and the switched tokens so that the newly substituted sentences are grammatically consistent during both training and inference. We conduct experiments on cross-lingual POS and NER over 30+ languages, and demonstrate the effectiveness of our method by outperforming the mBERT by 0.95 and original code-switching method by 1.67 on F1 scores.

**Federated Model Decomposition with Private Vocabulary for Text Classification**
*Zhuo Zhang, Xiangjing Hu, Lizhen Qu, Qifan Wang and Zenglin Xu* 15:30-17:00 (Hall A, Room C)
With the necessity of privacy protection, it becomes increasingly vital to train deep neural models in a federated learning manner for natural language processing (NLP) tasks. However, recent studies show eavesdroppers (i.e., dishonest servers) can still reconstruct the private input in federated learning (FL). Such a data reconstruction attack relies on the mappings between vocabulary and associated word embedding in NLP tasks, which are unfortunately less studied in current FL methods. In this paper, we propose a fedrated model decomposition method that protects the privacy of vocabularies, shorted as FEDEVOCAB. In FEDEVOCAB, each participant keeps the local embedding layer in the local device and detaches the local embedding parameters from federated aggregation. However, it is challenging to train an accurate NLP model when the private mappings are unknown and vary across participants in a cross-device FL setting. To address this problem, we further propose an adaptive updating technique to improve the performance of local models. Experimental results show that FEDEVOCAB maintains competitive performance and provides better privacy-preserving capacity compared to status quo methods.

**Tiny-Attention Adapter: Contexts Are More Important Than the Number of Parameters**
*Hongyu Zhao, Hao Tan and Hongyuan Mei* 15:30-17:00 (Hall A, Room C)
Adapter-tuning is a paradigm that transfers a pretrained language model to downstream tasks by adding and tuning a small number of new parameters. Previously proposed adapter architectures are all feed-forward neural networks. In this paper, we investigate the effectiveness of using tiny-attention—i.e., attention with extremely small per-head dimensionality—as adapters. Our tiny-attention adapter learns to modify the hidden states at each position directly conditioned on the hidden states at all the other positions, which is missed by the previously proposed adapters. Moreover, we view its multiple attention heads as a mixture of experts and propose to average their weights during deployment, which further reduces its inference computation cost. On the GLUE benchmark, our tiny-attention adapter outperforms the other parameter-efficient transfer learning methods as well as full fine-tuning while only updating 0.05% of the parameters. On the FewGLUE benchmark, its performance is comparable to that of GPT-3 and PET.

**Enhancing Multilingual Language Model with Massive Multilingual Knowledge Triples**
*Linlin Liu, Xin Li, Ruidan He, Lidong Bing, Shafiq Joty and Luo Si* 15:30-17:00 (Hall A, Room C)
Knowledge-enhanced language representation learning has shown promising results across various knowledge-intensive NLP tasks. However, prior methods are limited in efficient utilization of multilingual knowledge graph (KG) data for language model (LM) pretraining. They often train LMs with KGs in indirect ways, relying on extra entity/relation embeddings to facilitate knowledge injection. In this work, we explore methods to make better use of the multilingual annotation and language agnostic property of KG triples, and present novel knowledge based multilingual language models (KMLMs) trained directly on the knowledge triples. We first generate a large amount of multilingual synthetic sentences using the Wikidata KG triples. Then based on the intra- and inter-sentence structures of the generated data, we design pretraining tasks to enable the LMs to not only memorize the factual knowledge but also learn useful logical patterns. Our pretrained KMLMs demonstrate significant performance improvements on a wide range of knowledge-intensive cross-lingual tasks, including named entity recognition (NER), factual knowledge retrieval, relation classification, and a newly designed logical reasoning task.

**"It's Not Just Hate": A Multi-Dimensional Perspective on Detecting Harmful Speech Online**
*Federico Bianchi, Stefanie HIlls, Patricia Rossini, Dirk Hovy, Rebekah Tromble and Nava Tintarev* 15:30-17:00 (Hall A, Room C)

Well-annotated data is a prerequisite for good Natural Language Processing models. Too often, though, annotation decisions are governed by optimizing time or annotator agreement. We make a case for nuanced efforts in an interdisciplinary setting for annotating offensive online speech. Detecting offensive content is rapidly becoming one of the most important real-world NLP tasks. However, most datasets use a single binary label, e.g., for hate or incivility, even though each concept is multi-faceted. This modeling choice severely limits nuanced insights, but also performance. We show that a more fine-grained multi-label approach to predicting incivility and hateful or intolerant content addresses both conceptual and performance issues. We release a novel dataset of over 40,000 tweets about immigration from the US and UK, annotated with six labels for different aspects of incivility and intolerance. Our dataset not only allows for a more nuanced understanding of harmful speech online, models trained on it also outperform or match performance on benchmark datasets

**Semantic Novelty Detection and Characterization in Factual Text Involving Named Entities**
*Nianzu Ma, Sahisnu Mazumder, Alexander Politowicz, Bing Liu, Eric Robertson and Scott Grigsby* 15:30-17:00 (Hall A, Room C)
Much of the existing work on text novelty detection has been studied at the topic level, i.e., identifying whether the topic of a document or a sentence is novel or not. Little work has been done at the fine-grained semantic level (or contextual level). For example, given that we know Elon Musk is the CEO of a technology company, the sentence "Elon Musk acted in the sitcom The Big Bang Theory" is novel and surprising because normally a CEO would not be an actor. Existing topic-based novelty detection methods work poorly on this problem because they do not perform semantic reasoning involving relations between named entities in the text and their background knowledge. This paper proposes an effective model (called PAT-SND) to solve the problem, which can also characterize the novelty. An annotated dataset is also created. Evaluation shows that PAT-SND outperforms 10 baselines by large margins.

**AdapterShare: Task Correlation Modeling with Adapter Differentiation**
*Zhi Chen, Bei Chen, Lu Chen, Kai Yu and Jian-Guang LOU* 15:30-17:00 (Hall A, Room C)
Thanks to the development of pre-trained language models, multitask learning (MTL) methods achieve a great success in natural language understanding area. However, current MTL methods pay more attention to task selection or model design to fuse as much knowledge as possible, while intrinsic task correlation is often neglected. It is important to learn sharing strategy among multiple tasks rather than sharing everything. %The MTL model is directly shared among all the tasks. %For example, in traditional MTL methods, the last classification layers or the decoder layers are manually separated. More deeply, In this paper, we propose AdapterShare, an adapter differentiation method to explicitly model the task correlation among multiple tasks. AdapterShare is automatically learned based on the gradients on tiny held-out validation data. Compared to single-task learning and fully shared MTL methods, our proposed method obtains obvious performance improvement. Compared to the existing MTL method AdapterFusion, AdapterShare achieves absolute 1.90 average points improvement on five dialogue understanding tasks and 2.33 points gain on NLU tasks.

**Rethinking Task-Specific Knowledge Distillation: Contextualized Corpus as Better Textbook**
*Chang Liu, Chongyang Tao, Jianxin Liang, Tao Shen, Jiazhan Feng, Quzhe Huang and Dongyan Zhao* 15:30-17:00 (Hall A, Room C)
Knowledge distillation has been proven effective when customizing small language models for specific tasks. Here, a corpus as 'textbook' plays an indispensable role, only through which the teacher can teach the student. Prevailing methods adopt a two-stage distillation paradigm: general distillation first with task-agnostic general corpus and task-specific distillation next with augmented task-specific corpus. We argue that such a paradigm may not be optimal. In general distillation, it's extravagant to let the diverse but desultory general knowledge overwhelms the limited model capacity of the student. While in task-specific distillation, the task corpus is usually limited and narrow, preventing the student from learning enough knowledge. To mitigate the issues in the two gapped corpora, we present a better textbook for the student to learn: contextualized corpus that contextualizes task corpus with large-scale general corpus through relevance-based text retrieval. Experimental results on GLUE benchmark demonstrate that contextualized corpus is the better textbook compared with jointly using general corpus and augmented task-specific corpus. Surprisingly, it enables task-specific distillation from scratch without general distillation while maintaining comparable performance, making it more flexible to customize the student model with desired model size under various computation constraints.

**SHARE: a System for Hierarchical Assistive Recipe Editing**
*Shuyang Li, Yufei Li, Jianmo Ni and Julian McAuley* 15:30-17:00 (Hall A, Room C)
The large population of home cooks with dietary restrictions is under-served by existing cooking resources and recipe generation models. To help them, we propose the task of controllable recipe editing: adapt a base recipe to satisfy a user-specified dietary constraint. This task is challenging, and cannot be adequately solved with human-written ingredient substitution rules or existing end-to-end recipe generation models. We tackle this problem with SHARE: a System for Hierarchical Assistive Recipe Editing, which performs simultaneous ingredient substitution before generating natural-language steps using the edited ingredients. By decoupling ingredient and step editing, our step generator can explicitly integrate the available ingredients. Experiments on the novel RecipePairs dataset—83K pairs of similar recipes where each recipe satisfies one of seven dietary constraints—demonstrate that SHARE produces convincing, coherent recipes that are appropriate for a target dietary constraint. We further show through human evaluations and real-world cooking trials that recipes edited by SHARE can be easily followed by home cooks to create appealing dishes.

**Zero-shot Cross-lingual Transfer of Prompt-based Tuning with a Unified Multilingual Prompt**
*Lianzhe Huang, Shuming Ma, Dongdong Zhang, Furu Wei and Houfeng Wang* 15:30-17:00 (Hall A, Room C)
Prompt-based tuning has been proven effective for pretrained language models (PLMs). While most of the existing work focuses on the monolingual prompts, we study the multilingual prompts for multilingual PLMs, especially in the zero-shot cross-lingual setting. To alleviate the effort of designing different prompts for multiple languages, we propose a novel model that uses a unified prompt for all languages, called UniPrompt. Different from the discrete prompts and soft prompts, the unified prompt is model-based and language-agnostic. Specifically, the unified prompt is initialized by a multilingual PLM to produce language-independent representation, after which is fused with the text input. During inference, the prompts can be pre-computed so that no extra computation cost is needed. To collocate with the unified prompt, we propose a new initialization method for the target label word to further improve the model's transferability across languages. Extensive experiments show that our proposed methods can significantly outperform the strong baselines across different languages. We release data and code to facilitate future research.

**A Federated Approach to Predicting Emojis in Hindi Tweets**
*Deep Gandhi, Jash Mehta, Nirali Parekh, Karan Waghela, Lynette D'Mello and Zeerak Talat* 15:30-17:00 (Hall A, Room C)
The use of emojis affords a visual modality to, often private, textual communication. The task of predicting emojis however provides a challenge for machine learning as emoji use tends to cluster into the frequently used and the rarely used emojis. Much of the machine learning research on emoji use has focused on high resource languages and has conceptualised the task of predicting emojis around traditional server-side machine learning approaches. However, traditional machine learning approaches for private communication can introduce privacy concerns, as these approaches require all data to be transmitted to a central storage. In this paper, we seek to address the dual concerns of emphasising high resource languages for emoji prediction and risking the privacy of people's data. We introduce a new dataset of 118k tweets (augmented from 25k unique tweets) for emoji prediction in Hindi, and propose a modification to the federated learning algorithm, CausalFedGSD, which aims to strike a balance between model performance and user privacy. We show that our approach obtains comparative

scores with more complex centralised models while reducing the amount of data required to optimise the models and minimising risks to user privacy.

**PAR: Political Actor Representation Learning with Social Context and Expert Knowledge**
*Shangbin Feng, Zhaoxuan Tan, Zilong Chen, Ningnan Wang, Peisheng Yu, Qinghua Zheng, Xiaojun Chang and Minnan Luo* 15:30-17:00 (Hall A, Room C)
Modeling the ideological perspectives of political actors is an essential task in computational political science with applications in many downstream tasks. Existing approaches are generally limited to textual data and voting records, while they neglect the rich social context and valuable expert knowledge for holistic ideological analysis. In this paper, we propose PAR, a Political Actor Representation learning framework that jointly leverages social context and expert knowledge. Specifically, we retrieve and extract factual statements about legislators to leverage social context information. We then construct a heterogeneous information network to incorporate social context and use relational graph neural networks to learn legislator representations. Finally, we train PAR with three objectives to align representation learning with expert knowledge, model ideological stance consistency, and simulate the echo chamber phenomenon. Extensive experiments demonstrate that PAR is better at augmenting political text understanding and successfully advances the state-of-the-art in political perspective detection and roll call vote prediction. Further analysis proves that PAR learns representations that reflect the political reality and provide new insights into political behavior.

## Virtual Portal 16

15:30-17:00 (Hall A, Room D)

**MUSIED: A Benchmark for Event Detection from Multi-Source Heterogeneous Informal Texts**
*Xiangyu Xi, Jianwei Lv, Shuaipeng Liu, Wei Ye, Fan Yang and Guanglu Wan* 15:30-17:00 (Hall A, Room D)
Event detection (ED) identifies and classifies event triggers from unstructured texts, serving as a fundamental task for information extraction. Despite the remarkable progress achieved in the past several years, most research efforts focus on detecting events from formal texts (e.g., news articles, Wikipedia documents, financial announcements). Moreover, the texts in each dataset are either from a single source or multiple yet relatively homogeneous sources. With massive amounts of user-generated text accumulating on the Web and inside enterprises, identifying meaningful events in these informal texts, usually from multiple heterogeneous sources, has become a problem of significant practical value. As a pioneering exploration that expands event detection to the scenarios involving informal and heterogeneous texts, we propose a new large-scale Chinese event detection dataset based on user reviews, text conversations, and phone conversations in a leading e-commerce platform for food service. We carefully investigate the proposed dataset's textual informality and multi-domain heterogeneity characteristics by inspecting data samples quantitatively and qualitatively. Extensive experiments with state-of-the-art event detection methods verify the unique challenges posed by these characteristics, indicating that multi-domain informal event detection remains an open problem and requires further efforts. Our benchmark and code are released at https://github.com/myeclipse/MUSIED.

**Reproducibility Issues for BERT-based Evaluation Metrics**
*Yanran Chen, Jonas Belouadi and Steffen Eger* 15:30-17:00 (Hall A, Room D)
Reproducibility is of utmost concern in machine learning and natural language processing (NLP). In the field of natural language generation (especially machine translation), the seminal paper of Post (2018) has pointed out problems of reproducibility of the dominant metric, BLEU, at the time of publication. Nowadays, BERT-based evaluation metrics considerably outperform BLEU. In this paper, we ask whether results and claims from four recent BERT-based metrics can be reproduced. We find that reproduction of claims and results often fails because of (i) heavy undocumented preprocessing involved in the metrics, (ii) missing code and (iii) reporting weaker results for the baseline metrics. (iv) In one case, the problem stems from correlating not to human scores but to a wrong column in the csv file, inflating scores by 5 points. Motivated by the impact of preprocessing, we then conduct a second study where we examine its effects more closely (for one of the metrics). We find that preprocessing can have large effects, especially for highly inflectional languages. In this case, the effect of preprocessing may be larger than the effect of the aggregation mechanism (e.g., greedy alignment vs. Word Mover Distance).

**KECP: Knowledge Enhanced Contrastive Prompting for Few-shot Extractive Question Answering**
*Jianing Wang, Chengyu Wang, Minghui Qiu, Qiuhui Shi, Hongbin Wang, jun huang and Ming Gao* 15:30-17:00 (Hall A, Room D)
Extractive Question Answering (EQA) is one of the most essential tasks in Machine Reading Comprehension (MRC), which can be solved by fine-tuning the span selecting heads of Pre-trained Language Models (PLMs). However, most existing approaches for MRC may perform poorly in the few-shot learning scenario. To solve this issue, we propose a novel framework named Knowledge Enhanced Contrastive Prompt-tuning (KECP). Instead of adding pointer heads to PLMs, we introduce a seminal paradigm for EQA that transforms the task into a non-autoregressive Masked Language Modeling (MLM) generation problem. Simultaneously, rich semantics from the external knowledge base (KB) and the passage context support enhancing the query's representations. In addition, to boost the performance of PLMs, we jointly train the model by the MLM and contrastive learning objectives. Experiments on multiple benchmarks demonstrate that our method consistently outperforms state-of-the-art approaches in few-shot settings by a large margin.

**DisCup: Discriminator Cooperative Unlikelihood Prompt-tuning for Controllable Text Generation**
*Hanqing Zhang and Dawei Song* 15:30-17:00 (Hall A, Room D)
Prompt learning with immensely large Casual Language Models (CLMs) has been shown promising for attribute-controllable text generation (CTG). However, vanilla prompt tuning tends to imitate training corpus characteristics beyond the control attributes, resulting in a poor generalization ability. Moreover, it is less able to capture the relationship between different attributes, further limiting the control performance. In this paper, we propose a new CTG approach, namely DisCup, which incorporates the attribute knowledge of discriminator to optimize the control-prompts, steering a frozen CLM to produce attribute-specific texts. Specifically, the frozen CLM model, capable of producing multitudinous texts, is first used to generate the next-token candidates based on the context, so as to ensure the diversity of tokens to be predicted. Then, we leverage an attribute-discriminator to select desired/undesired tokens from those candidates, providing the inter-attribute knowledge. Finally, we bridge the above two traits by an unlikelihood objective for prompt-tuning. Extensive experimental results show that DisCup can achieve a new state-of-the-art control performance while maintaining an efficient and high-quality text generation, only relying on around 10 virtual tokens.

**Context Matters for Image Descriptions for Accessibility: Challenges for Referenceless Evaluation Metrics**
*Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris and Christopher Potts* 15:30-17:00 (Hall A, Room D)
Few images on the Web receive alt-text descriptions that would make them accessible to blind and low vision (BLV) users. Image-based NLG systems have progressed to the point where they can begin to address this persistent societal problem, but these systems will not be fully

successful unless we evaluate them on metrics that guide their development correctly. Here, we argue against current referenceless metrics – those that don't rely on human-generated ground-truth descriptions – on the grounds that they do not align with the needs of BLV users. The fundamental shortcoming of these metrics is that they do not take context into account, whereas contextual information is highly valued by BLV users. To substantiate these claims, we present a study with BLV participants who rated descriptions along a variety of dimensions. An in-depth analysis reveals that the lack of context-awareness makes current referenceless metrics inadequate for advancing image accessibility. As a proof-of-concept, we provide a contextual version of the referenceless metric CLIPScore which begins to address the disconnect to the BLV data.

**Evaluating Parameter Efficient Learning for Generation**
*Peng Xu, Mostofa Patwary, Shrimai Prabhumoye, Virginia Adams, Ryan Prenger, Wei Ping, Nayeon Lee, Mohammad Shoeybi and Bryan Catanzaro* 15:30-17:00 (Hall A, Room D)
Parameter efficient learning methods (PERMs) have recently gained significant attention as they provide an efficient way for pre-trained language models (PLMs) to adapt to a downstream task. However, these conclusions are mostly drawn from in-domain evaluations over the full training set. In this paper, we present comparisons between PERMs and finetuning from three new perspectives: (1) the effect of sample and model size to in-domain evaluations, (2) generalization to unseen domains and new datasets, and (3) the faithfulness of generations. Our results show that for in-domain settings (a) there is a cross point of sample size for which PERMs will perform better than finetuning when training with fewer samples, and (b) larger PLMs have larger cross points. For cross-domain and cross-dataset cases, we show that (a) Adapter (Houlsby et al., 2019) performs the best amongst all the PERMs studied here, and (b) it outperforms finetuning if the task dataset is below a certain size. We also compare the faithfulness of generations and show that PERMs can achieve better faithfulness score than finetuning, especially for small training set, by as much as 6%. Finally, we apply Adapter to MT-NLG 530b (Smith et al., 2022) and achieve new state-of-the-art results on Xsum (Narayan et al., 2018) for all ROUGE scores (ROUGE-1 49.17, ROUGE-2 27.20, ROUGE-L 40.98).

**Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning**
*Roshanak Mirzaee and Parisa Kordjamshidi* 15:30-17:00 (Hall A, Room D)
Recent research shows synthetic data as a source of supervision helps pretrained language models (PLM) transfer learning to new target tasks/domains. However, this idea is less explored for spatial language. We provide two new data resources on multiple spatial language processing tasks. The first dataset is synthesized for transfer learning on spatial question answering (SQA) and spatial role labeling (SpRL). Compared to previous SQA datasets, we include a larger variety of spatial relation types and spatial expressions. Our data generation process is easily extendable with new spatial expression lexicons. The second one is a real-world SQA dataset with human-generated questions built on an existing corpus with SPRL annotations. This dataset can be used to evaluate spatial language processing models in realistic situations. We show pretraining with automatically generated data significantly improves the SOTA results on several SQA and SPRL benchmarks, particularly when the training data in the target domain is small.

**Revisiting Grammatical Error Correction Evaluation and Beyond**
*Peiyuan Gong, Xuebo Liu, Heyan Huang and Min Zhang* 15:30-17:00 (Hall A, Room D)
Pretraining-based (PT-based) automatic evaluation metrics (e.g., BERTScore and BARTScore) have been widely used in several sentence generation tasks (e.g., machine translation and text summarization) due to their better correlation with human judgments over traditional overlap-based methods. Although PT-based methods have become the de facto standard for training grammatical error correction (GEC) systems, GEC evaluation still does not benefit from pretrained knowledge. This paper takes the first step towards understanding and improving GEC evaluation with pretraining. We first find that arbitrarily applying PT-based metrics to GEC evaluation brings unsatisfactory correlation results because of the excessive attention to inessential systems outputs (e.g., unchanged parts). To alleviate the limitation, we propose a novel GEC evaluation metric to achieve the best of both worlds, namely PT-M2 which only uses PT-based metrics to score those corrected parts. Experimental results on the CoNLL14 evaluation task show that PT-M2 significantly outperforms existing methods, achieving a new state-of-the-art result of 0.949 Pearson correlation. Further analysis reveals that PT-M2 is robust to evaluate competitive GEC systems. Source code and scripts are freely available at https://github.com/pygongnlp/PT-M2.

**RLET: A Reinforcement Learning Based Approach for Explainable QA with Entailment Trees**
*Tengxiao Liu, Qipeng Guo, Xiangkun Hu, Yue Zhang, Xipeng Qiu and Zheng Zhang* 15:30-17:00 (Hall A, Room D)
Interpreting the reasoning process from questions to answers poses a challenge in approaching explainable QA. A recently proposed structured reasoning format, entailment tree, manages to offer explicit logical deductions with entailment steps in a tree structure. To generate entailment trees, prior single pass sequence-to-sequence models lack visible internal decision probability, while stepwise approaches are supervised with extracted single step data and cannot model the tree as a whole. In this work, we propose RLET, a Reinforcement Learning based Entailment Tree generation framework, which is trained utilising the cumulative signals across the whole tree. RLET iteratively performs single step reasoning with sentence selection and deduction generation modules, from which the training signal is accumulated across the tree with elaborately designed aligned reward function that is consistent with the evaluation. To the best of our knowledge, we are the first to introduce RL into the entailment tree generation task. Experiments on three settings of the EntailmentBank dataset demonstrate the strength of using RL framework.

**Counterfactual Recipe Generation: Exploring Compositional Generalization in a Realistic Scenario**
*Xiao Liu, Yansong Feng, Jizhi Tang, Chengang Hu and Dongyan Zhao* 15:30-17:00 (Hall A, Room D)
People can acquire knowledge in an unsupervised manner by reading, and compose the knowledge to make novel combinations. In this paper, we investigate whether pretrained language models can perform compositional generalization in a realistic setting: recipe generation. We design the counterfactual recipe generation task, which asks models to modify a base recipe according to the change of an ingredient. This task requires compositional generalization at two levels: the surface level of incorporating the new ingredient into the base recipe, and the deeper level of adjusting actions related to the changing ingredient. We collect a large-scale recipe dataset in Chinese for models to learn culinary knowledge, and a subset of action-level fine-grained annotations for evaluation. We finetune pretrained language models on the recipe corpus, and use unsupervised counterfactual generation methods to generate modified recipes. Results show that existing models have difficulties in modifying the ingredients while preserving the original text style, and often miss actions that need to be adjusted. Although pretrained language models can generate fluent recipe texts, they fail to truly learn and use the culinary knowledge in a compositional way. Code and data are available at https://github.com/xxxiaol/counterfactual-recipe-generation.

**UniRPG: Unified Discrete Reasoning over Table and Text as Program Generation**
*Yongwei Zhou, Junwei Bao, Chaoqun Duan, Youzheng Wu, Xiaodong He and Tiejun Zhao* 15:30-17:00 (Hall A, Room D)
Question answering requiring discrete reasoning, e.g., arithmetic computing, comparison, and counting, over knowledge is a challenging task. In this paper, we propose UniRPG, a semantic-parsing-based approach advanced in interpretability and scalability, to perform Unified discrete Reasoning over heterogeneous knowledge resources, i.e., table and text, as Program Generation. Concretely, UniRPG consists of a neural programmer and a symbolic program executor, where a program is the composition of a set of pre-defined general atomic and higher-order operations and arguments extracted from table and text. First, the programmer parses a question into a program by generating operations and copying arguments, and then, the executor derives answers from table and text based on the program. To alleviate the costly program

annotation issue, we design a distant supervision approach for programmer learning, where pseudo programs are automatically constructed without annotated derivations. Extensive experiments on the TAT-QA dataset show that UniRPG achieves tremendous improvements and enhances interpretability and scalability compared with previous state-of-the-art methods, even without derivation annotation. Moreover, it achieves promising performance on the textual dataset DROP without derivation annotation.

**Long Text Generation with Topic-aware Discrete Latent Variable Model**
*Erguang Yang, Mingtong Liu, Deyi Xiong, YUJIE ZHANG, Yufeng Chen and Jinan Xu* 15:30-17:00 (Hall A, Room D)
Generating coherent long texts is an important yet challenging task, particularly for the open-ended generation. Prior work based on discrete latent codes focuses on the modeling of discourse relation, resulting in discrete codes only learning shallow semantics (Ji and Huang, 2021). A natural text always revolves around several related topics and the transition across them is natural and smooth. In this work, we investigate whether discrete latent codes can learn information of topics. To this end, we build a topic-aware latent code-guided text generation model. To encourage discrete codes to model information about topics, we propose a span-level bag-of-words training objective for the model. Automatic and manual evaluation experiments show that our method can generate more topic-relevant and coherent texts.

[DEMO] **LM-Debugger: An Interactive Tool for Inspection and Intervention in Transformer-Based Language Models**
*Mor Geva, Avi Caciularu, Guy Dar, Paul Roit, Shoval Sadde, Micah Shlain, Bar Tamir and Yoav Goldberg* 15:30-17:00 (Hall A, Room D)
The opaque nature and unexplained behavior of transformer-based language models (LMs) have spurred a wide interest in interpreting their predictions. However, current interpretation methods mostly focus on probing models from outside, executing behavioral tests, and analyzing salience input features, while the internal prediction construction process is largely not understood. In this work, we introduce LM-Debugger, an interactive debugger tool for transformer-based LMs, which provides a fine-grained interpretation of the model's internal prediction process, as well as a powerful framework for intervening in LM behavior. For its backbone, LM-Debugger relies on a recent method that interprets the inner token representations and their updates by the feed-forward layers in the vocabulary space. We demonstrate the utility of LM-Debugger for single-prediction debugging, by inspecting the internal disambiguation process done by GPT2. Moreover, we show how easily LM-Debugger allows to shift model behavior in a direction of the user's choice, by identifying a few vectors in the network and inducing effective interventions to the prediction process. We release LM-Debugger as an open-source tool and a demo over GPT2 models.

**arXivEdits: Understanding the Human Revision Process in Scientific Writing**
*Chao Jiang, Wei Xu and Samuel Stevens* 15:30-17:00 (Hall A, Room D)
Scientific publications are the primary means to communicate research discoveries, where the writing quality is of crucial importance. However, prior work studying the human editing process in this domain mainly focused on the abstract or introduction sections, resulting in an incomplete picture. In this work, we provide a complete computational framework for studying text revision in scientific writing. We first introduce arXivEdits, a new annotated corpus of 751 full papers from arXiv with gold sentence alignment across their multiple versions of revision, as well as fine-grained span-level edits and their underlying intentions for 1,000 sentence pairs. It supports our data-driven analysis to unveil the common strategies practiced by researchers for revising their papers. To scale up the analysis, we also develop automatic methods to extract revision at document-, sentence-, and word-levels. A neural CRF sentence alignment model trained on our corpus achieves 93.8 F1, enabling the reliable matching of sentences between different versions. We formulate the edit extraction task as a span alignment problem, and our proposed method extracts more fine-grained and explainable edits, compared to the commonly used diff algorithm. An intention classifier trained on our dataset achieves 78.9 F1 on the fine-grained intent classification task. Our data and system are released at tiny.one/arxivedits.

**MEE: A Novel Multilingual Event Extraction Dataset**
*Amir Pouran Ben Veyseh, Javid Ebrahimi, Franck Dernoncourt and Thien Nguyen* 15:30-17:00 (Hall A, Room D)
Event Extraction (EE) is one of the fundamental tasks in Information Extraction (IE) that aims to recognize event mentions and their arguments (i.e., participants) from text. Due to its importance, extensive methods and resources have been developed for Event Extraction. However, one limitation of current research for EE involves the under-exploration for non-English languages in which the lack of high-quality multilingual EE datasets for model training and evaluation has been the main hindrance. To address this limitation, we propose a novel Multilingual Event Extraction dataset (MEE) that provides annotation for more than 50K event mentions in 8 typologically different languages. MEE comprehensively annotates data for entity mentions, event triggers and event arguments. We conduct extensive experiments on the proposed dataset to reveal challenges and opportunities for multilingual EE. To foster future research in this direction, our dataset will be publicly available.

**Evade the Trap of Mediocrity: Promoting Diversity and Novelty in Text Generation via Concentrating Attention**
*Wenhao Li, Xiaoyuan Yi, Jinyi Hu, Maosong Sun and Xing Xie* 15:30-17:00 (Hall A, Room D)
Recently, powerful Transformer architectures have proven superior in generating high-quality sentences. Nevertheless, these models tend to produce dull high-frequency phrases, severely hurting the diversity and novelty of generated text. In this work, we dig into the intrinsic mechanism of this problem and found that sparser attention values in Transformer could improve diversity. To understand such a phenomenon, we first conduct both empirical and theoretical analysis and then attribute it to representation degeneration caused by the attentive mixture of the hidden states during training. We term this process the Trap of Mediocrity. To escape from such a trap, we introduce a novel attention regularization loss to control the sharpness of the attention distribution, which is transparent to model structures and can be easily implemented within 20 lines of python code. We prove that this method could be mathematically regarded as learning a Bayesian approximation of posterior attention. Experiments show that our method improved the diversity and novelty of the generated text while maintaining comparable quality on a variety of conditional and unconditional generation tasks.

**JDDC 2.1: A Multimodal Chinese Dialogue Dataset with Joint Tasks of Query Rewriting, Response Generation, Discourse Parsing, and Summarization**
*Nan Zhao, Haoran Li, Youzheng Wu and Xiaodong He* 15:30-17:00 (Hall A, Room D)
The popularity of multimodal dialogue has stimulated the need for a new generation of dialogue agents with multimodal interactivity. When users communicate with customer service, they may express their requirements by means of text, images, or even videos. Visual information usually acts as discriminators for product models, or indicators of product failures, which play an important role in the E-commerce scenario. On the other hand, detailed information provided by the images is limited, and typically, customer service systems cannot understand the intent of users without the input text. Thus, bridging the gap between the image and text is crucial for communicating with customers. In this paper, we construct JDDC 2.1, a large-scale multimodal multi-turn dialogue dataset collected from a mainstream Chinese E-commerce platform, containing about 246K dialogue sessions, 3M utterances, and 507K images, along with product knowledge bases and image category annotations. Over our dataset, we jointly define four tasks: the multimodal dialogue response generation task, the multimodal query rewriting task, the multimodal dialogue discourse parsing task, and the multimodal dialogue summarization task. JDDC 2.1 is the first corpus with annotations for all the above tasks over the same dialogue sessions, which facilitates the comprehensive research around the dialogue. In addition, we present several text-only and multimodal baselines and show the importance of visual information for these tasks. Our dataset and implements will be publicly available.

## Virtual Portal 17

15:30-17:00 (Hall B)

**R2F: A General Retrieval, Reading and Fusion Framework for Document-level Natural Language Inference**
*Hao Wang, Yixin Cao, Yangguang Li, Zhen Huang, Kun Wang and Jing Shao* 15:30-17:00 (Hall B)
Document-level natural language inference (DOCNLI) is a new challenging task in natural language processing, aiming at judging the entailment relationship between a pair of hypothesis and premise documents. Current datasets and baselines largely follow sentence-level settings, but fail to address the issues raised by longer documents. In this paper, we establish a general solution, named Retrieval, Reading and Fusion (R2F) framework, and a new setting, by analyzing the main challenges of DOCNLI: interpretability, long-range dependency, and cross-sentence inference. The basic idea of the framework is to simplify document-level task into a set of sentence-level tasks, and improve both performance and interpretability with the power of evidence. For each hypothesis sentence, the framework retrieves evidence sentences from the premise, and reads to estimate its credibility. Then the sentence-level results are fused to judge the relationship between the documents. For the setting, we contribute complementary evidence and entailment label annotation on hypothesis sentences, for interpretability study. Our experimental results show that R2F framework can obtain state-of-the-art performance and is robust for diverse evidence retrieval methods. Moreover, it can give more interpretable prediction results. Our model and code are released at https://github.com/phoenixsecularbird/R2F.

**RASAT: Integrating Relational Structures into Pretrained Seq2Seq Model for Text-to-SQL**
*Jiexing Qi, Jingyao Tang, Ziwei He, Xiangpeng Wan, Yu Cheng, Chenghu Zhou, Xinbing Wang, Quanshi Zhang and Zhouhan Lin*15:30-17:00 (Hall B)
Relational structures such as schema linking and schema encoding have been validated as a key component to qualitatively translating natural language into SQL queries. However, introducing these structural relations comes with prices: they often result in a specialized model structure, which largely prohibits using large pretrained models in text-to-SQL. To address this problem, we propose RASAT: a Transformer seq2seq architecture augmented with relation-aware self-attention that could leverage a variety of relational structures while inheriting the pretrained parameters from the T5 model effectively. Our model can incorporate almost all types of existing relations in the literature, and in addition, we propose introducing co-reference relations for the multi-turn scenario. Experimental results on three widely used text-to-SQL datasets, covering both single-turn and multi-turn scenarios, have shown that RASAT could achieve competitive results in all three benchmarks, achieving state-of-the-art execution accuracy (75.5% EX on Spider, 52.6% IEX on SParC, and 37.4% IEX on CoSQL).

**COM-MRC: A COntext-Masked Machine Reading Comprehension Framework for Aspect Sentiment Triplet Extraction**
*Zepeng Zhai, Hao Chen, Fangxiang Feng, Ruifan Li and Xiaojie WANG* 15:30-17:00 (Hall B)
Aspect Sentiment Triplet Extraction (ASTE) aims to extract sentiment triplets from sentences, which was recently formalized as an effective machine reading comprehension (MRC) based framework. However, when facing multiple aspect terms, the MRC-based methods could fail due to the interference from other aspect terms. In this paper, we propose a novel *COntext-Masked MRC* (COM-MRC) framework for ASTE. Our COM-MRC framework comprises three closely-related components: a context augmentation strategy, a discriminative model, and an inference method. Specifically, a context augmentation strategy is designed by enumerating all masked contexts for each aspect term. The discriminative model comprises four modules, i.e., aspect and opinion extraction modules, sentiment classification and aspect detection modules. In addition, a two-stage inference method first extracts all aspects and then identifies their opinions and sentiment through iteratively masking the aspects. Extensive experimental results on benchmark datasets show the effectiveness of our proposed COM-MRC framework, which outperforms state-of-the-art methods consistently.

**CEM: Machine-Human Chatting Handoff via Causal-Enhance Module**
*Shanshan Zhong, Jinghui Qin, Zhongzhan Huang and Daifeng Li* 15:30-17:00 (Hall B)
Aiming to ensure chatbot quality by predicting chatbot failure and enabling human-agent collaboration, Machine-Human Chatting Handoff (MHCH) has attracted lots of attention from both industry and academia in recent years. However, most existing methods mainly focus on the dialogue context or assist with global satisfaction prediction based on multi-task learning, which ignore the grounded relationships among the causal variables, like the user state and labor cost. These variables are significantly associated with handoff decisions, resulting in prediction bias and cost increasement. Therefore, we propose Causal-Enhance Module (CEM) by establishing the causal graph of MHCH based on these two variables, which is a simple yet effective module and can be easy to plug into the existing MHCH methods. For the impact of users, we use the user state to correct the prediction bias according to the causal relationship of multi-task. For the labor cost, we train an auxiliary cost simulator to calculate unbiased labor cost through counterfactual learning so that a model becomes cost-aware. Extensive experiments conducted on four real-world benchmarks demonstrate the effectiveness of CEM in generally improving the performance of existing MHCH methods without any elaborated model crafting.

**Face-Sensitive Image-to-Emotional-Text Cross-modal Translation for Multimodal Aspect-based Sentiment Analysis**
*Hao Yang, Yanyan Zhao and Bing Qin* 15:30-17:00 (Hall B)
Aspect-level multimodal sentiment analysis, which aims to identify the sentiment of the target aspect from multimodal data, recently has attracted extensive attention in the community of multimedia and natural language processing. Despite the recent success in textual aspect-based sentiment analysis, existing models mainly focused on utilizing the object-level semantic information in the image but ignore explicitly using the visual emotional cues, especially the facial emotions. How to distill visual emotional cues and align them with the textual content remains a key challenge to solve the problem. In this work, we introduce a face-sensitive image-to-emotional-text translation (FITE) method, which focuses on capturing visual sentiment cues through facial expressions and selectively matching and fusing with the target aspect in textual modality. To the best of our knowledge, we are the first that explicitly utilize the emotional information from images in the multimodal aspect-based sentiment analysis task. Experiment results show that our method achieves state-of-the-art results on the Twitter-2015 and Twitter-2017 datasets. The improvement demonstrates the superiority of our model in capturing aspect-level sentiment in multimodal data with facial expressions.

**Generating Literal and Implied Subquestions to Fact-check Complex Claims**
*Jifan Chen, Aniruddh Sriram, Eunsol Choi and Greg Durrett* 15:30-17:00 (Hall B)
Verifying political claims is a challenging task, as politicians can use various tactics to subtly misrepresent the facts for their agenda. Existing automatic fact-checking systems fall short here, and their predictions like "half-true" are not very useful in isolation, since it is unclear which parts of a claim are true and which are not. In this work, we focus on decomposing a complex claim into a comprehensive set of yes-no subquestions whose answers influence the veracity of the claim. We present CLAIMDECOMP, a dataset of decompositions for over 1000 claims. Given a claim and its verification paragraph written by fact-checkers, our trained annotators write subquestions covering both explicit propositions of the original claim and its implicit facets, such as asking about additional political context that changes our view of the claim's veracity. We study whether state-of-the-art models can generate such subquestions, showing that these models generate reasonable questions to ask, but predicting the comprehensive set of subquestions from the original claim without evidence remains challenging. We further show that these subquestions can help identify relevant evidence to fact-check the full claim and derive the veracity through their answers, suggest-

ing that they can be useful pieces of a fact-checking pipeline.

**Understanding Jargon: Combining Extraction and Generation for Definition Modeling**
*Jie Huang, Hanyin Shao, Kevin Chen-Chuan Chang, Jinjun Xiong and Wen-mei Hwu* 15:30-17:00 (Hall B)
Can machines know what twin prime is? From the composition of this phrase, machines may guess twin prime is a certain kind of prime, but it is still difficult to deduce exactly what twin stands for without additional knowledge. Here, twin prime is a jargon - a specialized term used by experts in a particular field. Explaining jargon is challenging since it usually requires domain knowledge to understand. Recently, there is an increasing interest in extracting and generating definitions of words automatically. However, existing approaches, either extraction or generation, perform poorly on jargon. In this paper, we propose to combine extraction and generation for jargon definition modeling: first extract self- and correlative definitional information of target jargon from the Web and then generate the final definitions by incorporating the extracted definitional information. Our framework is remarkably simple but effective: experiments demonstrate our method can generate high-quality definitions for jargon and outperform state-of-the-art models significantly, e.g., BLEU score from 8.76 to 22.66 and human-annotated score from 2.34 to 4.04.

**Exploiting Global and Local Hierarchies for Hierarchical Text Classification**
*Ting Jiang, Deqing Wang, Leilei Sun, Zhongzhi Chen, Fuzhen Zhuang and Qinghong Yang* 15:30-17:00 (Hall B)
Hierarchical text classification aims to leverage label hierarchy in multi-label text classification. Existing methods encode label hierarchy in a global view, where label hierarchy is treated as the static hierarchical structure containing all labels. Since global hierarchy is static and irrelevant to text samples, it makes these methods hard to exploit hierarchical information. Contrary to global hierarchy, local hierarchy as a structured labels hierarchy corresponding to each text sample. It is dynamic and relevant to text samples, which is ignored in previous methods. To exploit global and local hierarchies, we propose Hierarchy-guided BERT with Global and Local hierarchies (HBGL), which utilizes the large-scale parameters and prior language knowledge of BERT to model both global and local hierarchies. Moreover, HBGL avoids the intentional fusion of semantic and hierarchical modules by directly modeling semantic and hierarchical information with BERT. Compared with the state-of-the-art method HGCLR, our method achieves significant improvement on three benchmark datasets.

**A Span-level Bidirectional Network for Aspect Sentiment Triplet Extraction**
*Yuqi Chen, Chen Keming, Xian Sun and Zequn Zhang* 15:30-17:00 (Hall B)
Aspect Sentiment Triplet Extraction (ASTE) is a new fine-grained sentiment analysis task that aims to extract triplets of aspect terms, sentiments, and opinion terms from review sentences. Recently, span-level models achieve gratifying results on ASTE task by taking advantage of the predictions of all possible spans. Since all possible spans significantly increases the number of potential aspect and opinion candidates, it is crucial and challenging to efficiently extract the triplet elements among them. In this paper, we present a span-level bidirectional network which utilizes all possible spans as input and extracts triplets from spans bidirectionally. Specifically, we devise both the aspect decoder and opinion decoder to decode the span representations and extract triples from aspect-to-opinion and opinion-to-aspect directions. With these two decoders complementing with each other, the whole network can extract triplets from spans more comprehensively. Moreover, considering that mutual exclusion cannot be guaranteed between the spans, we design a similar span separation loss to facilitate the downstream task of distinguishing the correct span by expanding the KL divergence of similar spans during the training process; in the inference process, we adopt an inference strategy to remove conflicting triplets from the results base on their confidence scores. Experimental results show that our framework not only significantly outperforms state-of-the-art methods, but achieves better performance in predicting triplets with multi-token entities and extracting triplets in sentences contain multi-triplets.

**Learning Semantic Textual Similarity via Topic-informed Discrete Latent Variables**
*Erxin Yu, Lan Du, YUAN JIN, Zhepei Wei and Yi Chang* 15:30-17:00 (Hall B)
Recently, discrete latent variable models have received a surge of interest in both Natural Language Processing (NLP) and Computer Vision (CV), attributed to their comparable performance to the continuous counterparts in representation learning, while being more interpretable in their predictions. In this paper, we develop a topic-informed discrete latent variable model for semantic textual similarity, which learns a shared latent space for sentence-pair representation via vector quantization. Compared with previous models limited to local semantic contexts, our model can explore richer semantic information via topic modeling. We further boost the performance of semantic similarity by injecting the quantized representation into a transformer-based language model with a well-designed semantic-driven attention mechanism. We demonstrate, through extensive experiments across various English language datasets, that our model is able to surpass several strong neural baselines in semantic textual similarity tasks.

**Just Fine-tune Twice: Selective Differential Privacy for Large Language Models**
*Weiyan Shi, Ryan Shea, Si Chen, Chiyuan Zhang, Ruoxi Jia and Zhou Yu* 15:30-17:00 (Hall B)
Protecting large language models from privacy leakage is becoming increasingly crucial with their wide adoption in real-world products. Yet applying *differential privacy* (DP), a canonical notion with provable privacy guarantees for machine learning models, to those models remains challenging due to the trade-off between model utility and privacy loss. Utilizing the fact that sensitive information in language data tends to be sparse, Shi et al. (2021) formalized a DP notion extension called *Selective Differential Privacy* (SDP) to protect only the sensitive tokens defined by a policy function. However, their algorithm only works for RNN-based models. In this paper, we develop a novel framework, *Just Fine-tune Twice* (JFT), that achieves SDP for state-of-the-art large transformer-based models. Our method is easy to implement: it first fine-tunes the model with *redacted* in-domain data, and then fine-tunes it again with the *original* in-domain data using a private training mechanism. Furthermore, we study the scenario of imperfect implementation of policy functions that misses sensitive tokens and develop systematic methods to handle it. Experiments show that our method achieves strong utility compared to previous baselines. We also analyze the SDP privacy guarantee empirically with the canary insertion attack.

**Retrofitting Multilingual Sentence Embeddings with Abstract Meaning Representation**
*Deng Cai, Xin Li, Jackie Chun-Sing Ho, Lidong Bing and Wai Lam* 15:30-17:00 (Hall B)
We introduce a new method to improve existing multilingual sentence embeddings with Abstract Meaning Representation (AMR). Compared with the original textual input, AMR is a structured semantic representation that presents the core concepts and relations in a sentence explicitly and unambiguously. It also helps reduce the surface variations across different expressions and languages. Unlike most prior work that only evaluates the ability to measure semantic similarity, we present a thorough evaluation of existing multilingual sentence embeddings and our improved versions, which include a collection of five transfer tasks in different downstream applications. Experiment results show that retrofitting multilingual sentence embeddings with AMR leads to better state-of-the-art performance on both semantic textual similarity and transfer tasks.

**DEER: Descriptive Knowledge Graph for Explaining Entity Relationships**
*Jie Huang, Kerui Zhu, Kevin Chen-Chuan Chang, Jinjun Xiong and Wen-mei Hwu* 15:30-17:00 (Hall B)
We propose DEER (Descriptive Knowledge Graph for Explaining Entity Relationships) - an open and informative form of modeling entity relationships. In DEER, relationships between entities are represented by free-text relation descriptions. For instance, the relationship between entities of machine learning and algorithm can be represented as "Machine learning explores the study and construction of algorithms that can

learn from and make predictions on data." To construct DEER, we propose a self-supervised learning method to extract relation descriptions with the analysis of dependency patterns and generate relation descriptions with a transformer-based relation description synthesizing model, where no human labeling is required. Experiments demonstrate that our system can extract and generate high-quality relation descriptions for explaining entity relationships. The results suggest that we can build an open and informative knowledge graph without human annotation.

**Generative Data Augmentation with Contrastive Learning for Zero-Shot Stance Detection**
*Yang Li and Jiawei Yuan* 15:30-17:00 (Hall B)
Stance detection aims to identify whether the author of an opinionated text is in favor of, against, or neutral towards a given target. Remarkable success has been achieved when sufficient labeled training data is available. However, it is labor-intensive to annotate sufficient data and train the model for every new target. Therefore, zero-shot stance detection, aiming at identifying stances of unseen targets with seen targets, has gradually attracted attention. Among them, one of the important challenges is to reduce the domain transfer between seen and unseen targets. To tackle this problem, we propose a generative data augmentation approach to generate training samples containing targets and stances for testing data, and map the real samples and generated synthetic samples into the same embedding space with contrastive learning, then perform the final classification based on the augmented data. We evaluate our proposed model on two benchmark datasets. Experimental results show that our approach achieves state-of-the-art performance on most topics in the task of zero-shot stance detection.

**Text Style Transferring via Adversarial Masking and Styled Filling**
*Jiarui Wang, Richong Zhang, Junfan Chen, Jaein Kim and Yongyi Mao* 15:30-17:00 (Hall B)
Text style transfer is an important task in natural language processing with broad applications. Existing models following the masking and filling scheme suffer two challenges: the word masking procedure may mistakenly remove unexpected words and the selected words in the word filling procedure may lack diversity and semantic consistency. To tackle both challenges, in this study, we propose a style transfer model, with an adversarial masking approach and a styled filling technique (AMSF). Specifically, AMSF first trains a mask predictor by adversarial training without manual configuration. Then two additional losses, i.e. an entropy maximization loss and a consistency regularization loss, are introduced in training the word filling module to guarantee the diversity and semantic consistency of the transferred texts. Experimental results and analysis on two benchmark text style transfer data sets demonstrate the effectiveness of the proposed approaches.

**Generative Entity-to-Entity Stance Detection with Knowledge Graph Augmentation**
*Xinliang Frederick Zhang, Nick Beauchamp and Lu Wang* 15:30-17:00 (Hall B)
Stance detection is typically framed as predicting the sentiment in a given text towards a target entity. However, this setup overlooks the importance of the source entity, i.e., who is expressing the opinion. In this paper, we emphasize the imperative need for studying interactions among entities when inferring stances. We first introduce a new task, entity-to-entity (E2E) stance detection, which primes models to identify entities in their canonical names and discern stances jointly. To support this study, we curate a new dataset with 10,641 annotations labeled at the sentence level from news articles of different ideological leanings. We present a novel generative framework to allow the generation of canonical names for entities as well as stances among them. We further enhance the model with a graph encoder to summarize entity activities and external knowledge surrounding the entities. Experiments show that our model outperforms strong comparisons by large margins. Further analyses demonstrate the usefulness of E2E stance detection for understanding media quotation and stance landscape as well as inferring entity ideology.

**A Simple Contrastive Learning Framework for Interactive Argument Pair Identification via Argument-Context Extraction**
*Lida Shi, fausto giunchiglia, Rui Song, daqian Shi, Tongtong Liu, Xiaolei Diao and Hao Xu* 15:30-17:00 (Hall B)
Interactive argument pair identification is an emerging research task for argument mining, aiming to identify whether two arguments are interactively related. It is pointed out that the context of the argument is essential to improve identification performance. However, current context-based methods achieve limited improvements since the entire context typically contains much irrelevant information. In this paper, we propose a simple contrastive learning framework to solve this problem by extracting valuable information from the context. This framework can construct hard argument-context samples and obtain a robust and uniform representation by introducing contrastive learning. We also propose an argument-context extraction module to enhance information extraction by discarding irrelevant blocks. The experimental results show that our method achieves the state-of-the-art performance on the benchmark dataset. Further analysis demonstrates the effectiveness of our proposed modules and visually displays more compact semantic representations.

[DEMO] **CogKTR: A Knowledge-Enhanced Text Representation Toolkit for Natural Language Understanding**
*Zhuoran Jin, Tianyi Men, Hongbang Yuan, Yuyang Zhou, Pengfei Cao, Yubo Chen, Zhipeng Xue, Kang Liu and Jun Zhao* 15:30-17:00 (Hall B)
As the first step of modern natural language processing, text representation encodes discrete texts as continuous embeddings. Pre-trained language models (PLMs) have demonstrated strong ability in text representation and significantly promoted the development of natural language understanding (NLU). However, existing PLMs represent a text solely by its context, which is not enough to support knowledge-intensive NLU tasks. Knowledge is power, and fusing external knowledge explicitly into PLMs can provide knowledgeable text representations. Since previous knowledge-enhanced methods differ in many aspects, making it difficult for us to reproduce previous methods, implement new methods, and transfer between different methods. It is highly desirable to have a unified paradigm to encompass all kinds of methods in one framework. In this paper, we propose CogKTR, a knowledge-enhanced text representation toolkit for natural language understanding. According to our proposed Unified Knowledge-Enhanced Paradigm (UniKEP), CogKTR consists of four key stages, including knowledge acquisition, knowledge representation, knowledge injection, and knowledge application. CogKTR currently supports easy-to-use knowledge acquisition interfaces, multi-source knowledge embeddings, diverse knowledge-enhanced models, and various knowledge-intensive NLU tasks. Our unified, knowledgeable and modular toolkit is publicly available at GitHub, with an online system and a short instruction video.

[DEMO] **SynKB: Semantic Search for Synthetic Procedures**
*Fan Bai, Alan Ritter, Peter Madrid, Dayne Freitag and John Niekrasz* 15:30-17:00 (Hall B)
In this paper we present SynKB, an open-source, automatically extracted knowledge base of chemical synthesis protocols. Similar to proprietary chemistry databases such as Reaxsys, SynKB allows chemists to retrieve structured knowledge about synthetic procedures. By taking advantage of recent advances in natural language processing for procedural texts, SynKB supports more flexible queries about reaction conditions, and thus has the potential to help chemists search the literature for conditions used in relevant reactions as they design new synthetic routes. Using customized Transformer models to automatically extract information from 6 million synthesis procedures described in U.S. and EU patents, we show that for many queries, SynKB has higher recall than Reaxsys, while maintaining high precision. We plan to make SynKB available as an open-source tool; in contrast, proprietary chemistry databases require costly subscriptions.

## Virtual Portal 18

15:30-17:00 (Collaboratorium)

**Unsupervised Tokenization Learning**
*Anton Kolonin and Vignav Ramesh* 15:30-17:00 (Collaboratorium)
In the presented study, we discover that the so-called "transition freedom" metric appears superior for unsupervised tokenization purposes in comparison to statistical metrics such as mutual information and conditional probability, providing F-measure scores in range from 0.71 to 1.0 across explored multilingual corpora. We find that different languages require different offshoots of that metric (such as derivative, variance, and "peak values") for successful tokenization. Larger training corpora do not necessarily result in better tokenization quality, while compressing the models by eliminating statistically weak evidence tends to improve performance. The proposed unsupervised tokenization technique provides quality better than or comparable to lexicon-based ones, depending on the language.

**Few-shot Query-Focused Summarization with Prefix-Merging**
*Ruifeng Yuan, zili Wang, Ziqiang Cao and Wenjie Li* 15:30-17:00 (Collaboratorium)
Query-focused summarization has been considered as an important extension for text summarization. It aims to generate a concise highlight for a given query. Different from text summarization, query-focused summarization has long been plagued by the problem of lacking high-quality large-scale datasets. In this paper, we investigate the idea that whether we can integrate and transfer the knowledge of text summarization and question answering to assist the few-shot learning in query-focused summarization. Here, we propose prefix-merging, a prefix-based pretraining strategy for few-shot learning in query-focused summarization. Drawn inspiration from prefix-tuning, we are allowed to integrate the task knowledge from text summarization and question answering into a properly designed prefix and apply the merged prefix to query-focused summarization. With only a small amount of trainable parameters, prefix-merging outperforms fine-tuning on query-focused summarization. We further discuss the influence of different prefix designs and propose a visualized explanation for how prefix-merging works.

**FastClass: A Time-Efficient Approach to Weakly-Supervised Text Classification**
*Tingyu Xia, Yue Wang, Yuan Tian and Yi Chang* 15:30-17:00 (Collaboratorium)
Weakly-supervised text classification aims to train a classifier using only class descriptions and unlabeled data. Recent research shows that keyword-driven methods can achieve state-of-the-art performance on various tasks. However, these methods not only rely on carefully-crafted class descriptions to obtain class-specific keywords but also require substantial amount of unlabeled data and takes a long time to train. This paper proposes FastClass, an efficient weakly-supervised classification approach. It uses dense text representation to retrieve class-relevant documents from external unlabeled corpus and selects an optimal subset to train a classifier. Compared to keyword-driven methods, our approach is less reliant on initial class descriptions as it no longer needs to expand each class description into a set of class-specific keywords.Experiments on a wide range of classification tasks show that the proposed approach frequently outperforms keyword-driven models in terms of classification accuracy and often enjoys orders-of-magnitude faster training speed.

**RelCLIP: Adapting Language-Image Pretraining for Visual Relationship Detection via Relational Contrastive Learning**
*Yi Zhu, Zhaoqing Zhu, Bingqian Lin, Xiaodan Liang, Feng Zhao and Jianzhuang Liu* 15:30-17:00 (Collaboratorium)
Conventional visual relationship detection models only use the numeric ids of relation labels for training, but ignore the semantic correlation between the labels, which leads to severe training biases and harms the generalization ability of representations. In this paper, we introduce compact language information of relation labels for regularizing the representation learning of visual relations. Specifically, we propose a simple yet effective visual Relationship prediction framework that transfers natural language knowledge learned from Contrastive Language-Image Pre-training (CLIP) models to enhance the relationship prediction, termed RelCLIP. Benefiting from the powerful visual-semantic alignment ability of CLIP at image level, we introduce a novel Relational Contrastive Learning (RCL) approach which explores relation-level visual-semantic alignment via learning to match cross-modal relational embeddings. By collaboratively learning the semantic coherence and discrepancy from relation triplets, the model can generate more discriminative and robust representations. Experimental results on the Visual Genome dataset show that RelCLIP achieves significant improvements over strong baselines under full (provide accurate labels) and distant supervision (provide noise labels), demonstrating its powerful generalization ability in learning relationship representations. Code will be available at https://gitee.com/mindspore/models/tree/master/research/cv/RelCLIP.

**Discrete Cross-Modal Alignment Enables Zero-Shot Speech Translation**
*Chen Wang, Yuchen Liu, Boxing Chen, Jiajun Zhang, Wei Luo, Zhongqiang Huang and Chengqing Zong* 15:30-17:00 (Collaboratorium)
End-to-end Speech Translation (ST) aims at translating the source language speech into target language text without generating the intermediate transcriptions. However, the training of end-to-end methods relies on parallel ST data, which are difficult and expensive to obtain. Fortunately, the supervised data for automatic speech recognition (ASR) and machine translation (MT) are usually more accessible, making zero-shot speech translation a potential direction. Existing zero-shot methods fail to align the two modalities of speech and text into a shared semantic space, resulting in much worse performance compared to the supervised ST methods. In order to enable zero-shot ST, we propose a novel Discrete Cross-Modal Alignment (DCMA) method that employs a shared discrete vocabulary space to accommodate and match both modalities of speech and text. Specifically, we introduce a vector quantization module to discretize the continuous representations of speech and text into a finite set of virtual tokens, and use ASR data to map corresponding speech and text to the same virtual token in a shared codebook. This way, source language speech can be embedded in the same semantic space as the source language text, which can be then transformed into target language text with an MT module. Experiments on multiple language pairs demonstrate that our zero-shot ST method significantly improves the SOTA, and even performers on par with the strong supervised ST baselines.

**GHAN: Graph-Based Hierarchical Aggregation Network for Text-Video Retrieval**
*Yahan Yu, Bojie Hu and Yu Li* 15:30-17:00 (Collaboratorium)
Text-video retrieval focuses on two aspects: cross-modality interaction and video-language encoding. Currently, the mainstream approach is to train a joint embedding space for multimodal interactions. However, there are structural and semantic differences between text and video, making this approach challenging for fine-grained understanding. In order to solve this, we propose an end-to-end graph-based hierarchical aggregation network for text-video retrieval according to the hierarchy possessed by text and video. We design a token-level weighted network to refine intra-modality representations and construct a graph-based message passing attention network for global-local alignment across modality. We conduct experiments on the public datasets MSR-VTT-9K, MSR-VTT-7K and MSVD, and achieve Recall@1 of 73.0

**R-TeaFor: Regularized Teacher-Forcing for Abstractive Summarization**
*Guan-Yu Lin and Pu-Jen Cheng* 15:30-17:00 (Collaboratorium)
Teacher-forcing is widely used in training sequence generation models to improve sampling efficiency and to stabilize training. However, teacher-forcing is vulnerable to the exposure bias problem. Previous works have attempted to address exposure bias by modifying the training data to simulate model-generated results. Nevertheless, they do not consider the pairwise relationship between the original training data and the modified ones, which provides more information during training. Hence, we propose Regularized Teacher-Forcing (R-TeaFor) to utilize this relationship for better regularization. Empirically, our experiments show that R-TeaFor outperforms previous summarization state-of-the-

art models, and the results can be generalized to different pre-trained models.

**Open-Domain Sign Language Translation Learned from Online Video**
*Bowen Shi, Diane Brentari, Gregory Shakhnarovich and Karen Livescu* 15:30-17:00 (Collaboratorium)
Existing work on sign language translation – that is, translation from sign language videos into sentences in a written language – has focused mainly on (1) data collected in a controlled environment or (2) data in a specific domain, which limits the applicability to real-world settings. In this paper, we introduce OpenASL, a large-scale American Sign Language (ASL) - English dataset collected from online video sites (e.g., YouTube). OpenASL contains 288 hours of ASL videos in multiple domains from over 200 signers and is the largest publicly available ASL translation dataset to date. To tackle the challenges of sign language translation in realistic settings and without glosses, we propose a set of techniques including sign search as a pretext task for pre-training and fusion of mouthing and handshape features. The proposed techniques produce consistent and large improvements in translation quality, over baseline models based on prior work.

**SEMGraph: Incorporating Sentiment Knowledge and Eye Movement into Graph Model for Sentiment Analysis**
*Bingbing Wang, Bin Liang, Jiachen Du, Min Yang and Ruifeng Xu* 15:30-17:00 (Collaboratorium)
This paper investigates the sentiment analysis task from a novel perspective by incorporating sentiment knowledge and eye movement into a graph architecture, aiming to draw the eye movement-based sentiment relationships for learning the sentiment expression of the context. To be specific, we first explore a linguistic probing eye movement paradigm to extract eye movement features based on the close relationship between linguistic features and the early and late processes of human reading behavior. Furthermore, to derive eye movement features with sentiment concepts, we devise a novel weighting strategy to integrate sentiment scores extracted from affective commonsense knowledge into eye movement features, called sentiment-eye movement weights. Then, the sentiment-eye movement weights are exploited to build the sentiment-eye movement guided graph (SEMGraph) model, so as to model the intricate sentiment relationships in the context. Experimental results on two sentiment analysis datasets with eye movement signals and three sentiment analysis datasets without eye movement signals show that the proposed SEMGraph achieves state-of-the-art performance, and can also be directly generalized to those sentiment analysis datasets without eye movement signals.

**Towards Summary Candidates Fusion**
*Mathieu Ravaut, Shafiq Joty and Nancy Chen* 15:30-17:00 (Collaboratorium)
Sequence-to-sequence deep neural models fine-tuned for abstractive summarization can achieve great performance on datasets with enough human annotations. Yet, it has been shown that they have not reached their full potential, with a wide gap between the top beam search output and the oracle beam. Recently, re-ranking methods have been proposed, to learn to select a better summary candidate. However, such methods are limited by the summary quality aspects captured by the first-stage candidates. To bypass this limitation, we propose a new paradigm in second-stage abstractive summarization called SummaFusion that fuses several summary candidates to produce a novel abstractive second-stage summary. Our method works well on several summarization datasets, improving both the ROUGE scores and qualitative properties of fused summaries. It is especially good when the candidates to fuse are worse, such as in the few-shot setup where we set a new state-of-the art. We will make our code and checkpoints available at https://github.com/ntunlp/SummaFusion/.

**Contrastive Learning with Expectation-Maximization for Weakly Supervised Phrase Grounding**
*Keqin Chen, Richong Zhang, Samuel Mensah and Yongyi Mao* 15:30-17:00 (Collaboratorium)
Weakly supervised phrase grounding aims to learn an alignment between phrases in a caption and objects in a corresponding image using only caption-image annotations, i.e., without phrase-object annotations. Previous methods typically use a caption-image contrastive loss to indirectly supervise the alignment between phrases and objects, which hinders the maximum use of the intrinsic structure of the multimodal data and leads to unsatisfactory performance. In this work, we directly use the phrase-object contrastive loss in the condition that no positive annotation is available in the first place. Specifically, we propose a novel contrastive learning framework based on the expectation-maximization algorithm that adaptively refines the target prediction. Experiments on two widely used benchmarks, Flickr30K Entities and RefCOCO+, demonstrate the effectiveness of our framework. We obtain 63.05% top-1 accuracy on Flickr30K Entities and 59.51%/43.46% on RefCOCO+ TestA/TestB, outperforming the previous methods by a large margin, even surpassing a previous SoTA that uses a pre-trained vision-language model. Furthermore, we deliver a theoretical analysis of the effectiveness of our method from the perspective of the maximum likelihood estimate with latent variables.

**Weakly-Supervised Temporal Article Grounding**
*Long Chen, Yulei Niu, Brian Chen, Xudong Lin, Guangxing Han, Christopher Thomas, Hammad Ayyubi, Heng Ji and Shih-Fu Chang* 15:30-17:00 (Collaboratorium)
Given a long untrimmed video and natural language queries, video grounding (VG) aims to temporally localize the semantically-aligned video segments. Almost all existing VG work holds two simple but unrealistic assumptions: 1) All query sentences can be grounded in the corresponding video. 2) All query sentences for the same video are always at the same semantic scale. Unfortunately, both assumptions make today's VG models fail to work in practice. For example, in real-world multimodal assets (eg, news articles), most of the sentences in the article can not be grounded in their affiliated videos, and they typically have rich hierarchical relations (ie, at different semantic scales). To this end, we propose a new challenging grounding task: Weakly-Supervised temporal Article Grounding (WSAG). Specifically, given an article and a relevant video, WSAG aims to localize all "groundable" sentences to the video, and these sentences are possibly at different semantic scales. Accordingly, we collect the first WSAG dataset to facilitate this task: YouwikiHow, which borrows the inherent multi-scale descriptions in wikiHow articles and plentiful YouTube videos. In addition, we propose a simple but effective method DualMIL for WSAG, which consists of a two-level MIL loss and a single-/cross- sentence constraint loss. These training objectives are carefully designed for these relaxed assumptions. Extensive ablations have verified the effectiveness of DualMIL.

**HEGEL: Hypergraph Transformer for Long Document Summarization**
*Haopeng Zhang, Xiao Liu and Jiawei Zhang* 15:30-17:00 (Collaboratorium)
Extractive summarization for long documents is challenging due to the extended structured input context. The long-distance sentence dependency hinders cross-sentence relations modeling, the critical step of extractive summarization. This paper proposes HEGEL, a hypergraph neural network for long document summarization by capturing high-order cross-sentence relations. HEGEL updates and learns effective sentence representations with hypergraph transformer layers and fuses different types of sentence dependencies, including latent topics, keywords coreference, and section structure. We validate HEGEL by conducting extensive experiments on two benchmark datasets, and experimental results demonstrate the effectiveness and efficiency of HEGEL.

**FaD-VLP: Fashion Vision-and-Language Pre-training towards Unified Retrieval and Captioning**
*Suvir Mirchandani, Licheng Yu, Mengjiao Wang, Animesh Sinha, Wenwen Jiang, Tao Xiang and Ning Zhang* 15:30-17:00 (Collaboratorium)
Multimodal tasks in the fashion domain have significant potential for e-commerce, but involve challenging vision-and-language learning problems—e.g., retrieving a fashion item given a reference image plus text feedback from a user. Prior works on multimodal fashion tasks have either been limited by the data in individual benchmarks, or have leveraged generic vision-and-language pre-training but have not taken advantage of the characteristics of fashion data. Additionally, these works have mainly been restricted to multimodal understanding tasks.

To address these gaps, we make two key contributions. First, we propose a novel fashion-specific pre-training framework based on weakly-supervised triplets constructed from fashion image-text pairs. We show the triplet-based tasks are an effective addition to standard multimodal pre-training tasks. Second, we propose a flexible decoder-based model architecture capable of both fashion retrieval and captioning tasks. Together, our model design and pre-training approach are competitive on a diverse set of fashion tasks, including cross-modal retrieval, image retrieval with text feedback, image captioning, relative image captioning, and multimodal categorization.

**End-to-End Unsupervised Vision-and-Language Pre-training with Referring Expression Matching**
*Chi Chen, Peng Li, Maosong Sun and Yang Liu* 15:30-17:00 (Collaboratorium)
Recently there has been an emerging interest in unsupervised vision-and-language pre-training (VLP) that learns multimodal representations without parallel image-caption data. These pioneering works significantly reduce the cost of VLP on data collection and achieve promising results compared to supervised VLP. However, existing unsupervised VLP methods take as input pre-extracted region-based visual features from external object detectors, which both limits flexibility and reduces computational efficiency. In this paper, we explore end-to-end unsupervised VLP with a vision encoder to directly encode images. The vision encoder is pre-trained on image-only data and jointly optimized during multimodal pre-training. To further enhance the learned cross-modal features, we propose a novel pre-training task that predicts which patches contain an object referred to in natural language from the encoded visual features. Extensive experiments on four vision-and-language tasks show that our approach outperforms previous unsupervised VLP methods and obtains new state-of-the-art results.

**CiteSum: Citation Text-guided Scientific Extreme Summarization and Domain Adaptation with Limited Supervision**
*Yuning Mao, Ming Zhong and Jiawei Han* 15:30-17:00 (Collaboratorium)
Scientific extreme summarization (TLDR) aims to form ultra-short summaries of scientific papers. Previous efforts on curating scientific TLDR datasets failed to scale up due to the heavy human annotation and domain expertise required. In this paper, we propose a simple yet effective approach to automatically extracting TLDR summaries for scientific papers from their citation texts. Based on the proposed approach, we create a new benchmark CiteSum without human annotation, which is around 30 times larger than the previous human-curated dataset SciTLDR. We conduct a comprehensive analysis of CiteSum, examining its data characteristics and establishing strong baselines. We further demonstrate the usefulness of CiteSum by adapting models pre-trained on CiteSum (named CITES) to new tasks and domains with limited supervision. For scientific extreme summarization, CITES outperforms most fully-supervised methods on SciTLDR without any fine-tuning and obtains state-of-the-art results with only 128 examples. For news extreme summarization, CITES achieves significant gains on XSum over its base model (not pre-trained on CiteSum), e.g., +7.2 ROUGE-1 zero-shot performance and state-of-the-art few-shot performance. For news headline generation, CITES performs the best among unsupervised and zero-shot methods on Gigaword.

**Retrieval Augmented Visual Question Answering with Outside Knowledge**
*Weizhe Lin and Bill Byrne* 15:30-17:00 (Collaboratorium)
Outside-Knowledge Visual Question Answering (OK-VQA) is a challenging VQA task that requires retrieval of external knowledge to answer questions about images. Recent OK-VQA systems use Dense Passage Retrieval (DPR) to retrieve documents from external knowledge bases, such as Wikipedia, but with DPR trained separately from answer generation, introducing a potential limit on the overall system performance. Instead, we propose a joint training scheme which includes differentiable DPR integrated with answer generation so that the system can be trained in an end-to-end fashion. Our experiments show that our scheme outperforms recent OK-VQA systems with strong DPR for retrieval. We also introduce new diagnostic metrics to analyze how retrieval and generation interact. The strong retrieval ability of our model significantly reduces the number of retrieved documents needed in training, yielding significant benefits in answer quality and computation required for training.

**CycleKQR: Unsupervised Bidirectional Keyword-Question Rewriting**
*Andrea Iovine, Anjie Fang, Besnik Fetahu, Jie Zhao, Oleg Rokhlenko and Shervin Malmasi* 15:30-17:00 (Collaboratorium)
Users expect their queries to be answered by search systems, regardless of the query's surface form, which include keyword queries and natural questions. Natural Language Understanding (NLU) components of Search and QA systems may fail to correctly interpret semantically equivalent inputs if this deviates from how the system was trained, leading to suboptimal understanding capabilities. We propose the keyword-question rewriting task to improve query understanding capabilities of NLU systems for all surface forms. To achieve this, we present CycleKQR, an unsupervised approach, enabling effective rewriting between keyword and question queries using non-parallel data. Empirically we show the impact on QA performance of unfamiliar query forms for open domain and Knowledge Base QA systems (trained on either keywords or natural language questions). We demonstrate how CycleKQR significantly improves QA performance by rewriting queries into the appropriate form, while at the same time retaining the original semantic meaning of input queries, allowing CycleKQR to improve performance by up to 3% over supervised baselines. Finally, we release a dataset of 66k keyword-question pairs.

## Poster Sessions 17 & 18

15:30-17:00 (Atrium)

**ExPUNations: Augmenting Puns with Keywords and Explanations**
*Jiao Sun, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Tagyoung Chung, Jing Huang, Yang Liu and Nanyun Peng*15:30-17:00 (Atrium)
The tasks of humor understanding and generation are challenging and subjective even for humans, requiring commonsense and real-world knowledge to master. Puns, in particular, add the challenge of fusing that knowledge with the ability to interpret lexical-semantic ambiguity. In this paper, we present the ExPUNations (ExPUN) dataset, in which we augment an existing dataset of puns with detailed crowdsourced annotations of keywords denoting the most distinctive words that make the text funny, pun explanations describing why the text is funny, and fine-grained funniness ratings. This is the first humor dataset with such extensive and fine-grained annotations specifically for puns. Based on these annotations, we propose two tasks: explanation generation to aid with pun classification and keyword-conditioned pun generation, to challenge the current state-of-the-art natural language understanding and generation models' ability to understand and generate humor. We showcase that the annotated keywords we collect are helpful for generating better novel humorous texts in human evaluation, and that our natural language explanations can be leveraged to improve both the accuracy and robustness of humor classifiers.

**Self-supervised Graph Masking Pre-training for Graph-to-Text Generation**
*Jiuzhou Han and Ehsan Shareghi* 15:30-17:00 (Atrium)
Large-scale pre-trained language models (PLMs) have advanced Graph-to-Text (G2T) generation by processing the linearised version of a graph. However, the linearisation is known to ignore the structural information. Additionally, PLMs are typically pre-trained on free text which introduces domain mismatch between pre-training and downstream G2T generation tasks. To address these shortcomings, we propose graph masking pre-training strategies that neither require supervision signals nor adjust the architecture of the underlying pre-trained encoder-

decoder model. When used with a pre-trained T5, our approach achieves new state-of-the-art results on WebNLG+2020 and EventNarrative G2T generation datasets. Our method also shows to be very effective in the low-resource setting.

**STRUDEL: Structured Dialogue Summarization for Dialogue Comprehension**
*Borui Wang, Chengcheng Feng, Arjun Nair, Madelyn Mao, Jai Desai, Asli Celikyilmaz, Haoran Li, Yashar Mehdad and Dragomir Radev*
15:30-17:00 (Atrium)
Abstractive dialogue summarization has long been viewed as an important standalone task in natural language processing, but no previous work has explored the possibility of whether abstractive dialogue summarization can also be used as a means to boost an NLP system's performance on other important dialogue comprehension tasks. In this paper, we propose a novel type of dialogue summarization task - STRUctured DiaLoguE Summarization (STRUDEL) - that can help pre-trained language models to better understand dialogues and improve their performance on important dialogue comprehension tasks. In contrast to the holistic approach taken by the traditional free-form abstractive summarization task for dialogues, STRUDEL aims to decompose and imitate the hierarchical, systematic and structured mental process that we human beings usually go through when understanding and analyzing dialogues, and thus has the advantage of being more focused, specific and instructive for dialogue comprehension models to learn from. We further introduce a new STRUDEL dialogue comprehension modeling framework that integrates STRUDEL into a dialogue reasoning module over transformer encoder language models to improve their dialogue comprehension ability. In our empirical experiments on two important downstream dialogue comprehension tasks - dialogue question answering and dialogue response prediction - we demonstrate that our STRUDEL dialogue comprehension models can significantly improve the dialogue comprehension performance of transformer encoder language models.

**Fine-tuned Language Models are Continual Learners**
*Thomas Scialom, Tuhin Chakrabarty and Smaranda Muresan* 15:30-17:00 (Atrium)
Recent work on large language models relies on the intuition that most natural language processing tasks can be described via natural language instructions and that models trained on these instructions show strong zero-shot performance on several standard datasets. However, these models even though impressive still perform poorly on a wide range of tasks outside of their respective training and evaluation sets.To address this limitation, we argue that a model should be able to keep extending its knowledge and abilities, without forgetting previous skills. In spite of the limited success of Continual Learning, we show that *Fine-tuned Language Models can be continual learners.*We empirically investigate the reason for this success and conclude that Continual Learning emerges from self-supervision pre-training. Our resulting model Continual-T0 (CT0) is able to learn 8 new diverse language generation tasks, while still maintaining good performance on previous tasks, spanning in total of 70 datasets. Finally, we show that CT0 is able to combine instructions in ways it was never trained for, demonstrating some level of instruction compositionality.

**Boosting Natural Language Generation from Instructions with Meta-Learning**
*Budhaditya Deb, Ahmed Hassan Awadallah and Guoqing Zheng* 15:30-17:00 (Atrium)
Recent work has shown that language models (LMs) trained with multi-task *instructional learning* (MTIL) can solve diverse NLP tasks in zero- and few-shot settings with improved performance compared to prompt tuning. MTIL illustrates that LMs can extract and use information about the task from instructions beyond the surface patterns of the inputs and outputs. This suggests that meta-learning may further enhance the utilization of instructions for effective task transfer. In this paper we investigate whether meta-learning applied to MTIL can further improve generalization to unseen tasks in a zero-shot setting. Specifically, we propose to adapt meta-learning to MTIL in three directions: 1) Model Agnostic Meta Learning (MAML), 2) Hyper-Network (HNet) based adaptation to generate task specific parameters conditioned on instructions, and 3) an approach combining HNet and MAML. Through extensive experiments on the large scale Natural Instructions V2 dataset, we show that our proposed approaches significantly improve over strong baselines in zero-shot settings. In particular, meta-learning improves the effectiveness of instructions and is most impactful when the test tasks are strictly zero-shot (i.e. no similar tasks in the training set) and are "hard" for LMs, illustrating the potential of meta-learning for MTIL for out-of-distribution tasks.

**GREENER: Graph Neural Networks for News Media Profiling**
*Panayot Panayotov, Utsav Shukla, Husrev Taha Sencar, Mohamed Nabeel and Preslav Nakov* 15:30-17:00 (Atrium)
We study the problem of profiling news media on the Web with respect to their factuality of reporting and bias. This is an important but under-studied problem related to disinformation and "fake news" detection, but it addresses the issue at a coarser granularity compared to looking at an individual article or an individual claim. This is useful as it allows to profile entire media outlets in advance. Unlike previous work, which has focused primarily on text (e.g., on the text of the articles published by the target website, or on the textual description in their social media profiles or in Wikipedia), here our main focus is on modeling the similarity between media outlets based on the overlap of their audience. This is motivated by homophily considerations, i.e., the tendency of people to have connections to people with similar interests, which we extend to media, hypothesizing that similar types of media would be read by similar kinds of users. In particular, we propose GREENER (GRaph nEural nEtwork for News mEdia pRofiling), a model that builds a graph of inter-media connections based on their audience overlap, and then uses graph neural networks to represent each medium. We find that such representations are quite useful for predicting the factuality and the bias of news media outlets, yielding improvements over state-of-the-art results reported on two datasets. When augmented with conventionally used representations obtained from news articles, Twitter, YouTube, Facebook, and Wikipedia, prediction accuracy is found to improve by 2.5-27 macro-F1 points for the two tasks.

**Towards Opening the Black Box of Neural Machine Translation: Source and Target Interpretations of the Transformer**
*Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano and Marta R. Costa-jussà* 15:30-17:00 (Atrium)
In Neural Machine Translation (NMT), each token prediction is conditioned on the source sentence and the target prefix (what has been previously translated at a decoding step). However, previous work on interpretability in NMT has mainly focused solely on source sentence tokens' attributions. Therefore, we lack a full understanding of the influences of every input token (source sentence and target prefix) in the model predictions. In this work, we propose an interpretability method that tracks input tokens' attributions for both contexts. Our method, which can be extended to any encoder-decoder Transformer-based model, allows us to better comprehend the inner workings of current NMT models. We apply the proposed method to both bilingual and multilingual Transformers and present insights into their behaviour.

**A Major Obstacle for NLP Research: Let's Talk about Time Allocation!**
*Katharina Kann, Shiran Dudy and Arya D. McCarthy* 15:30-17:00 (Atrium)
The field of natural language processing (NLP) has grown over the last few years: conferences have become larger, we have published an incredible amount of papers, and state-of-the-art research has been implemented in a large variety of customer-facing products. However, this paper argues that we have been less successful than we *should* have been and reflects on where and how the field fails to tap its full potential. Specifically, we demonstrate that, in recent years, **subpar time allocation has been a major obstacle for NLP research**. We outline multiple concrete problems together with their negative consequences and, importantly, suggest remedies to improve the status quo. We hope that this paper will be a starting point for discussions around which common practices are – or are *not* – beneficial for NLP research.

**Overcoming Catastrophic Forgetting in Zero-Shot Cross-Lingual Generation**
*Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer and Noah Constant* 15:30-17:00 (Atrium)

In this paper, we explore the challenging problem of performing a generative task in a target language when labeled data is only available in English, using summarization as a case study. We assume a strict setting with no access to parallel data or machine translation and find that common transfer learning approaches struggle in this setting, as a generative multilingual model fine-tuned purely on English catastrophically forgets how to generate non-English. Given the recent rise of parameter-efficient adaptation techniques, we conduct the first investigation into how one such method, prompt tuning (Lester et al., 2021), can overcome catastrophic forgetting to enable zero-shot cross-lingual generation. Our experiments show that parameter-efficient prompt tuning provides gains over standard fine-tuning when transferring between less-related languages, e.g., from English to Thai. However, a significant gap still remains between these methods and fully-supervised baselines. To improve cross-lingual transfer further, we explore several approaches, including: (1) mixing in unlabeled multilingual data, and (2) explicitly factoring prompts into recombinable language and task components. Our approaches can provide further quality gains, suggesting that robust zero-shot cross-lingual generation is within reach.

**Unsupervised Entity Linking with Guided Summarization and Multiple-Choice Selection**
*Young Min Cho, Li Zhang and Chris Callison-Burch* 15:30-17:00 (Atrium)
Entity linking, the task of linking potentially ambiguous mentions in texts to corresponding knowledge-base entities, is an important component for language understanding. We address two challenge in entity linking: how to leverage wider contexts surrounding a mention, and how to deal with limited training data. We propose a fully unsupervised model called SumMC that first generates a guided summary of the contexts conditioning on the mention, and then casts the task to a multiple-choice problem where the model chooses an entity from a list of candidates. In addition to evaluating our model on existing datasets that focus on named entities, we create a new dataset that links noun phrases from WikiHow to Wikidata. We show that our SumMC model achieves state-of-the-art unsupervised performance on our new dataset and on exiting datasets.

**RobustLR: A Diagnostic Benchmark for Evaluating Logical Robustness of Deductive Reasoners**
*Soumya Sanyal, Zeyi Liao and Xiang Ren* 15:30-17:00 (Atrium)
Transformers have been shown to be able to perform deductive reasoning on inputs containing rules and statements written in the English natural language. However, it is unclear if these models indeed follow rigorous logical reasoning to arrive at the prediction or rely on spurious correlation patterns in making decisions. A strong deductive reasoning model should consistently understand the semantics of different logical operators. To this end, we present RobustLR, a diagnostic benchmark that evaluates the robustness of language models to minimal logical edits in the inputs and different logical equivalence conditions. In our experiments with RoBERTa, T5, and GPT3 we show that the models trained on deductive reasoning datasets do not perform consistently on the RobustLR test set, thus showing that the models are not robust to our proposed logical perturbations. Further, we observe that the models find it especially hard to learn logical negation operators. Our results demonstrate the shortcomings of current language models in logical reasoning and call for the development of better inductive biases to teach the logical semantics to language models. All the datasets and code base have been made publicly available.

**Algorithms for Weighted Pushdown Automata**
*Alexandra Butoi, Brian DuSell, Tim Vieira, Ryan Cotterell and David Chiang* 15:30-17:00 (Atrium)
Weighted pushdown automata (WPDAs) are at the core of many natural language processing tasks, like syntax-based statistical machine translation and transition-based dependency parsing. As most existing dynamic programming algorithms are designed for context-free grammars (CFGs), algorithms for PDAs often resort to a PDA-to-CFG conversion. In this paper, we develop novel algorithms that operate directly on WPDAs. Our algorithms are inspired by Lang's algorithm, but use a more general definition of pushdown automaton and either reduce the space requirements by a factor of |Gamma| (the size of the stack alphabet) or reduce the runtime by a factor of more than |Q| (the number of states). When run on the same class of PDAs as Lang's algorithm, our algorithm is both more space-efficient by a factor of |Gamma| and more time-efficient by a factor of |Q| x |Gamma|.

**On the Evaluation Metrics for Paraphrase Generation**
*Lingfeng Shen, Lemao Liu, Haiyun Jiang and Shuming Shi* 15:30-17:00 (Atrium)
In this paper we revisit automatic metrics for paraphrase evaluation and obtain two findings that disobey conventional wisdom: (1) Reference-free metrics achieve better performance than their reference-based counterparts. (2) Most commonly used metrics do not align well with human annotation. Underlying reasons behind the above findings are explored through additional experiments and in-depth analyses. Based on the experiments and analyses, we propose ParaScore, a new evaluation metric for paraphrase generation. It possesses the merits of reference-based and reference-free metrics and explicitly models lexical divergence. Based on our analysis and improvements, our proposed reference-based outperforms than reference-free metrics. Experimental results demonstrate that ParaScore significantly outperforms existing metrics.

**CONQRR: Conversational Query Rewriting for Retrieval with Reinforcement Learning**
*Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, Hannaneh Hajishirzi, Mari Ostendorf and Gaurav Singh Tomar* 15:30-17:00 (Atrium)
Compared to standard retrieval tasks, passage retrieval for conversational question answering (CQA) poses new challenges in understanding the current user question, as each question needs to be interpreted within the dialogue context. Moreover, it can be expensive to re-train well-established retrievers such as search engines that are originally developed for non-conversational queries. To facilitate their use, we develop a query rewriting model CONQRR that rewrites a conversational question in the context into a standalone question. It is trained with a novel reward function to directly optimize towards retrieval using reinforcement learning and can be adapted to any off-the-shelf retriever. CONQRR achieves state-of-the-art results on a recent open-domain CQA dataset containing conversations from three different sources, and is effective for two different off-the-shelf retrievers. Our extensive analysis also shows the robustness of CONQRR to out-of-domain dialogues as well as to zero query rewriting supervision.

**Nearest Neighbor Zero-Shot Inference**
*Weijia Shi, Julian Michael, Suchin Gururangan and Luke Zettlemoyer* 15:30-17:00 (Atrium)
Retrieval-augmented language models (LMs) use non-parametric memory to substantially outperform their non-retrieval counterparts on perplexity-based evaluations, but it is an open question whether they achieve similar gains in few- and zero-shot end-task accuracy. We extensively study one such model, the k-nearest neighbor LM (kNN-LM), showing that the gains marginally transfer. The main challenge is to achieve coverage of the verbalizer tokens that define the different end-task class labels. To address this challenge, we also introduce kNN-Prompt, a simple and effective kNN-LM with automatically expanded fuzzy verbalizers (e.g. to expand "terrible" to also include "silly" and other task-specific synonyms for sentiment classification). Across nine diverse end-tasks, using kNN-Prompt with GPT-2 large yields significant performance boosts over strong zeroshot baselines (13.4% absolute improvement over the base LM on average). We also show that other advantages of non-parametric augmentation hold for end tasks; kNN-Prompt is effective for domain adaptation with no further training, and gains increase with the size of the retrieval model.

**Topic Modeling With Topological Data Analysis**
*Ciarán Byrne, Danijela Horak, Karo Moilanen and Amandla Mabona* 15:30-17:00 (Atrium)
Recent unsupervised topic modelling ap- proaches that use clustering techniques on word, token or document embeddings can ex- tract coher-

ent topics. A common limitation of such approaches is that they reveal noth- ing about inter-topic relationships which are essential in many real-world application do- mains. We present an unsupervised topic mod- elling method which harnesses Topological Data Analysis (TDA) to extract a topological skeleton of the manifold upon which contextu- alised word embeddings lie. We demonstrate that our approach, which performs on par with a recent baseline, is able to construct a network of coherent topics together with meaningful relationships between them.

**Fixing Model Bugs with Natural Language Patches**
*Shikhar Murty, Christopher Manning, Scott Lundberg and Marco Tulio Ribeiro* 15:30-17:00 (Atrium)
Current approaches for fixing systematic problems in NLP models (e.g., regex patches, finetuning on more data) are either brittle, or labor-intensive and liable to shortcuts. In contrast, humans often provide corrections to each other through natural language. Taking inspiration from this, we explore natural language patches—declarative statements that allow developers to provide corrective feedback at the right level of abstraction, either overriding the model ("if a review gives 2 stars, the sentiment is negative") or providing additional information the model may lack ("if something is described as the bomb, then it is good"). We model the task of determining if a patch applies separately from the task of integrating patch information, and show that with a small amount of synthetic data, we can teach models to effectively use real patches on real data—1 to 7 patches improve accuracy by 1–4 accuracy points on different slices of a sentiment analysis dataset, and F1 by 7 points on a relation extraction dataset. Finally, we show that finetuning on as many as 100 labeled examples may be needed to match the performance of a small set of language patches.

**SCROLLS: Standardized CompaRison Over Long Language Sequences**
*Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant and Omer Levy* 15:30-17:00 (Atrium)
NLP benchmarks have largely focused on short texts, such as sentences and paragraphs, even though long texts comprise a considerable amount of natural language in the wild. We introduce SCROLLS, a suite of tasks that require reasoning over long texts. We examine existing long-text datasets, and handpick ones where the text is naturally long, while prioritizing tasks that involve synthesizing information across the input. SCROLLS contains summarization, question answering, and natural language inference tasks, covering multiple domains, including literature, science, business, and entertainment. Initial baselines, including Longformer Encoder-Decoder, indicate that there is ample room for improvement on SCROLLS. We make all datasets available in a unified text-to-text format and host a live leaderboard to facilitate research on model architecture and pretraining methods.

**Logical Neural Networks for Knowledge Base Completion with Embeddings & Rules**
*Prithviraj Sen, Breno William Carvalho, Ibrahim Abdelaziz, Pavan Kapanipathi, Salim Roukos and Alexander Gray* 15:30-17:00 (Atrium)
Knowledge base completion (KBC) has benefitted greatly by learning explainable rules in an human-interpretable dialect such as first-order logic. Rule-based KBC has so far, mainly focussed on learning one of two types of rules: conjunction-of-disjunctions and disjunction-of-conjunctions. We qualitatively show, via examples, that one of these has an advantage over the other when it comes to achieving high quality KBC. To the best of our knowledge, we are the first to propose learning both kinds of rules within a common framework. To this end, we propose to utilize logical neural networks (LNN), a powerful neuro-symbolic AI framework that can express both kinds of rules and learn these end-to-end using gradient-based optimization. Our in-depth experiments show that our LNN-based approach to learning rules for KBC leads to roughly 10% relative improvements, if not more, over SotA rule-based KBC methods. Moreover, by showing how to combine our proposed methods with knowledge graph embeddings we further achieve an additional 7.5% relative improvement.

**Passage-Mask: A Learnable Regularization Strategy for Retriever-Reader Models**
*Shujian Zhang, Chengyue Gong and Xingchao Liu* 15:30-17:00 (Atrium)
Retriever-reader models achieve competitive performance across many different NLP tasks such as open question answering and dialogue conversations. In this work, we notice these models easily overfit the top-rank retrieval passages and standard training fails to reason over the entire retrieval passages. We introduce a learnable passage mask mechanism which desensitizes the impact from the top-rank retrieval passages and prevents the model from overfitting. Controlling the gradient variance with fewer mask candidates and selecting the mask candidates with one-shot bi-level optimization, our learnable regularization strategy enforces the answer generation to focus on the entire retrieval passages. Experiments on different tasks across open question answering, dialogue conversation, and fact verification show that our method consistently outperforms its baselines. Extensive experiments and ablation studies demonstrate that our method can be general, effective, and beneficial for many NLP tasks.

**Semantic-aware Contrastive Learning for More Accurate Semantic Parsing**
*Shan Wu, Chunlei Xin, Bo Chen, Xianpei Han and Le Sun* 15:30-17:00 (Atrium)
Since the meaning representations are detailed and accurate annotations which express fine-grained sequence-level semtantics, it is usually hard to train discriminative semantic parsers via Maximum Likelihood Estimation (MLE) in an autoregressive fashion. In this paper, we propose a semantic-aware contrastive learning algorithm, which can learn to distinguish fine-grained meaning representations and take the overall sequence-level semantic into consideration. Specifically, a multi-level online sampling algorithm is proposed to sample confusing and diverse instances. Three semantic-aware similarity functions are designed to accurately measure the distance between meaning representations as a whole. And a ranked contrastive loss is proposed to pull the representations of the semantic-identical instances together and push negative instances away. Experiments on two standard datasets show that our approach achieves significant improvements over MLE baselines and gets state-of-the-art performances by simply applying semantic-aware contrastive learning on a vanilla Seq2Seq model.

**Hardness-guided domain adaptation to recognise biomedical named entities under low-resource scenarios**
*Ngoc Dang Nguyen, Lan Du, Wray Buntine, Changyou Chen and Richard Beare* 15:30-17:00 (Atrium)
Domain adaptation is an effective solution to data scarcity in low-resource scenarios. However, when applied to token-level tasks such as bioNER, domain adaptation methods often suffer from the challenging linguistic characteristics that clinical narratives possess, which leads to unsatsifactory performance. In this paper, we present a simple yet effective hardness-guided domain adaptation framework for bioNER tasks that can effectively leverage the domain hardness information to improve the adaptability of the learnt model in the low-resource scenarios. Experimental results on biomedical datasets show that our model can achieve significant performance improvement over the recently published state-of-the-art (SOTA) MetaNER model.

[CL] **Nucleus Composition in Transition-Based Dependency Parsing**
*Joakim Nivre, Ali Basirat, Luise Dürlich and Adam Moss* 15:30-17:00 (Atrium)
Dependency-based approaches to syntactic analysis assume that syntactic structure can be analyzed in terms of binary asymmetric dependency relations holding between elementary syntactic units. Computational models for dependency parsing almost universally assume that an elementary syntactic unit is a word, while the influential theory of Lucien Tesnière instead posits a more abstract notion of nucleus, which may be realized as one or more words. In this article, we investigate the effect of enriching computational parsing models with a concept of nucleus inspired by Tesnière. We begin by reviewing how the concept of nucleus can be defined in the framework of Universal Dependencies, which has become the de facto standard for training and evaluating supervised dependency parsers, and explaining how composition functions can be used to make neural transition-based dependency parsers aware of the nuclei thus defined. We then perform an extensive experimental

study, using data from 20 languages to assess the impact of nucleus composition across languages with different typological characteristics, and employing a variety of analytical tools including ablation, linear mixed-effects models, diagnostic classifiers and dimensionality reduction. The analysis reveals that nucleus composition gives small but consistent improvements in parsing accuracy for most languages, and that the improvement mainly concerns the analysis of main predicates, nominal dependents, clausal dependents and coordination structures. Significant factors explaining the rate of improvement across languages include entropy in coordination structures and frequency of certain function words, in particular determiners. Analysis using dimensionality reduction and diagnostic classifiers suggests that nucleus composition increases the similarity of vectors representing nuclei of the same syntactic type.

# Plenary: Best Papers

17:00-17:45 - **Hall B**

# 9 Workshops

## Overview

During the days of the workshops, **Registration** will be held from 07:30.

### Wednesday, December 7, 2022



### Thursday, December 8, 2022

# W1 - The 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text

**Organizers:**
Ali Hürriyetoğlu, Hristo Tanev, Vanni Zavarella, Erdem Yörük

https://emw.ku.edu.tr/case-2022/
Venue: Capital Suite 4
**Wednesday, December 7, 2022 - Thursday, December 8, 2022**

Nowadays, the unprecedented quantity of easily accessible data on social, political, and economic processes offers ground-breaking potential in guiding data-driven analysis in social and human sciences and in driving informed policy-making processes. Governments, multilateral organizations, local and global NGOs, and present an increasing demand for high-quality information about a wide variety of events ranging from political violence, environmental catastrophes, and conflict, to international economic and health crises (Coleman et al. 2014; Porta and Diani, 2015) to prevent or resolve conflicts, provide relief for those that are afflicted, or improve the lives of and protect citizens in a variety of ways. Black Lives Matter protests and conflicts in Syria are only two examples where we must understand, analyze, and improve the real-life situations using such data. Finally, these efforts respond to "growing public interest in up-to-date information on crowds" as well.

Event extraction has long been a challenge for the natural language processing (NLP) community as it requires sophisticated methods in defining event ontologies, creating language resources, and developing algorithmic approaches (Pustojevsky et al. 2003; Boroş, 2018; Chen et al. 2021). Social and political scientists have been working to create socio-political event (SPE) databases such as ACLED, EMBERS, GDELT, ICEWS, MMAD, PHOENIX, POLDEM, SPEED, TERRIER, and UCDP following similar steps for decades. These projects and the new ones increasingly rely on machine learning (ML), deep learning (DL), and NLP methods to deal better with the vast amount and variety of data in this domain (Hürriyetoğlu et al. 2020). Automation offers scholars not only the opportunity to improve existing practices, but also to vastly expand the scope of data that can be collected and studied, thus potentially opening up new research frontiers within the field of SPEs, such as political violence and social movements. But automated approaches as well suffer from major issues like bias, generalizability, class imbalance, training data limitations, and ethical issues that have the potential to affect the results and their use drastically (Lau and Baldwin 2020; Bhatia et al. 2020; Chang et al. 2019). Moreover, the results of the automated systems for SPE information collection have neither been comparable to each other nor been of sufficient quality (Wang et al. 2016; Schrodt 2020).

SPEs are varied and nuanced. Both the political context and the local language used may affect whether and how they are reported. Therefore, all steps of information collection (event definition, language resources, and manual or algorithmic steps) may need to be constantly updated, leading to a series of challenging questions: Do events related to minority groups are represented well? Are new types of events covered? Are the event definitions and their operationalization comparable across systems? This workshop aims to seek answers to these questions as well. Inspiring innovative technological and scientific solutions for tackling these issues and quantifying the quality of the results are the main goals of CASE workshop series.

| | |
|---|---|
| 09:00 - 18:30 | ***Day 1*** |
| 09:00 - 17:30 | ***Tutorials*** |
| 11:00 - 12:30 | ***Poster Session (S2)*** |

| | |
|---|---|
| 12:30 - 14:00 | ***Lunch Break (LB)*** |
| 14:00 - 15:30 | ***Afternoon Session (S3)*** |
| 15:30 - 16:00 | ***Afternoon Coffee Break (B2)*** |
| 16:00 - 17:30 | ***Afternoon Session (S4)*** |
| 17:30 - 18:30 | ***Keynote 1 Session (S5)*** |
| 09:00 - 18:30 | ***Day 2*** |
| 09:00 - 17:30 | ***Tutorials*** |
| 11:00 - 12:30 | ***Poster Session (S2)*** |
| 12:30 - 14:00 | ***Lunch Break (LB)*** |
| 14:00 - 15:30 | ***Afternoon Session (S3)*** |
| 15:30 - 16:00 | ***Afternoon Coffee Break (B2)*** |
| 16:00 - 17:30 | ***Afternoon Session (S4)*** |
| 17:30 - 18:30 | ***Keynote 2 Session (S5)*** |

# W2 - The 26th Conference on Computational Natural Language Learning

**Organizers:**
Antske Fokkens, Vivek Srikumar

https://www.conll.org/
Venue: Capital Suite 1
**Wednesday, December 7, 2022 - Thursday, December 8, 2022**

CoNLL is a yearly conference organized by SIGNLL (ACL's Special Interest Group on Natural Language Learning), focusing on theoretically, cognitively and scientifically motivated approaches to computational linguistics. Such approaches include computational learning theory and other techniques for theoretical analysis of machine learning models for NLP; models of first, second and bilingual language acquisition by humans; models of language evolution and change; computational simulation and analysis of findings from psycholinguistic and neurolinguistic experiments; analysis and interpretation of NLP models, using methods inspired by cognitive science or linguistics or other methods; data resources, techniques and tools for scientifically-oriented research in computational linguistics; connections between computational models and formal languages or linguistic theories; linguistic typology, translation, and other multilingual work; and theoretically, cognitively and scientifically motivated approaches to text generation.

| | |
|---|---|
| 09:00 - 09:10 | ***Opening Remarks*** |
| 09:10 - 10:30 | ***Keynote 1: Noah Goodman*** |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 12:30 | ***Oral Session 1: Machine Learning for NLP, Model Interpretation*** |
| 11:00-11:20 | *Continual Learning for Natural Language Generations with Transformer Calibration*<br>Peng Yang, Dingcheng Li and Ping Li |
| 11:20-11:40 | *Towards More Natural Artificial Languages*<br>Mark Hopkins |
| 11:40-12:00 | *Probing for targeted syntactic knowledge through grammatical error detection*<br>Christopher Davis, Christopher Bryant, Andrew Caines, Marek Rei and Paula Buttery |
| 12:00-12:30 | *Enhancing the Transformer Decoder with Transition-based Syntax*<br>Leshem Choshen and Omri Abend |
| 12:30 - 14:00 | ***Lunch Break*** |
| 14:00 - 15:30 | ***Oral Session 2: Multilingual Work and Translation*** |
| 14:00-14:20 | *A Multilingual Bag-of-Entities Model for Zero-Shot Cross-Lingual Text Classification*<br>Sosuke Nishikawa, Ikuya Yamada, Yoshimasa Tsuruoka and Isao Echizen |
| 14:20-14:40 | *Combining Noisy Semantic Signals with Orthographic Cues: Cognate Induction for the Indic Dialect Continuum*<br>Niyati Bafna, Josef van Genabith, Cristina España-Bonet and Zdeněk Žabokrtský |
| 14:40-15:00 | *On Neurons Invariant to Sentence Structural Changes in Neural Machine Translation*<br>Gal Patel, Leshem Choshen and Omri Abend |

| | |
|---|---|
| 15:00-15:30 | *On Language Spaces, Scales and Cross-Lingual Transfer of UD Parsers*<br>Tanja Samardžić, Ximena Gutierrez-Vasques, Rob van der Goot, Max Müller-Eberstein, Olga Pelloni and Barbara Plank |
| 15:30 - 16:00 | ***Coffee Break*** |
| 16:00 - 17:30 | ***Poster Session 1: Virtual*** |
| | *How Hate Speech Varies by Target Identity: A Computational Analysis*<br>Michael Yoder, Lynnette Ng, David West Brown and Kathleen Carley |
| | *OpenStance: Real-world Zero-shot Stance Detection*<br>Hanzi Xu, Slobodan Vucetic and Wenpeng Yin |
| | *Characterizing Verbatim Short-Term Memory in Neural Language Models*<br>Kristijan Armeni, Christopher Honey and Tal Linzen |
| | *Parsing as Deduction Revisited: Using an Automatic Theorem Prover to Solve an SMT Model of a Minimalist Parser*<br>Sagar Indurkhya |
| | *An Alignment-based Approach to Text Segmentation Similarity Scoring*<br>Gerardo Ocampo Diaz and Jessica Ouyang |
| | *A Fine-grained Interpretability Evaluation Benchmark for Neural NLP*<br>Lijie Wang, Yaozong Shen, Shuyuan Peng, Shuai Zhang, Xinyan Xiao, Hao Liu, Hongxuan Tang, Ying Chen, Hua Wu and Haifeng Wang |
| 09:10 - 10:30 | ***Keynote 2: Allyson Ettinger*** |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 12:30 | ***Poster session 2: In-person*** |
| | *That's so cute!: The CARE Dataset for Affective Response Detection*<br>Jane Yu and Alon Halevy |
| | *Causal Analysis of Syntactic Agreement Neurons in Multilingual Language Models*<br>Aaron Mueller, Yu Xia and Tal Linzen |
| | *Detecting Unintended Social Bias in Toxic Language Datasets*<br>Nihar Sahoo, Himanshu Gupta and Pushpak Bhattacharyya |
| | *Leveraging a New Spanish Corpus for Multilingual and Cross-lingual Metaphor Detection*<br>Elisa Sanchez-Bayona and Rodrigo Agerri |
| | *Cognitive Simplification Operations Improve Text Simplification*<br>Eytan Chamovitz and Omri Abend |
| | *Optimizing text representations to capture (dis)similarity between political parties*<br>Tanise Ceron, Nico Blokker and Sebastian Padó |
| | *PIE-QG: Paraphrased Information Extraction for Unsupervised Question Generation from Small Corpora*<br>Dinesh Nagumothu, Bahadorreza Ofoghi, Guangyan Huang and Peter Eklund |
| 12:30 - 14:00 | ***Lunch Break*** |
| 14:00 - 15:30 | ***Oral Session 3: Psycholinguistics and Language Models*** |
| 14:00-14:20 | *Collateral facilitation in humans and language models*<br>James Michaelov and Benjamin Bergen |
| 14:20-14:40 | *Incremental Processing of Principle B: Mismatches Between Neural Models and Humans*<br>Forrest Davis |
| 14:40-15:00 | *Syntactic Surprisal From Neural Models Predicts, But Underestimates, Human Processing Difficulty From Syntactic Ambiguities*<br>Suhas Arehalli, Brian Dillon and Tal Linzen |
| 15:00-15:30 | *Computational cognitive modeling of predictive sentence processing in a second language* |

| | |
|---|---|
| | Umesh Patil and Sol Lago |
| 15:30 - 16:00 | ***Coffee Break*** |
| 16:00 - 17:00 | ***Oral Session 4: Semantics and Grounding*** |
| 16:00-16:20 | *Entailment Semantics Can Be Extracted from an Ideal Language Model*<br>William Merrill, Alex Warstadt and Tal Linzen |
| 16:20-16:40 | *Shared knowledge in natural conversations: can entropy metrics shed light on information transfers?*<br>Eliot Maës, Philippe Blache and Leonor Becerra |
| 16:40-17:00 | *Visual Semantic Parsing: From Images to Abstract Meaning Representation*<br>Mohamed Ashraf Abdelsalam, Zhan Shi, Federico Fancellu, Kalliopi Basioti, Dhaivat Bhatt, Vladimir Pavlovic and Afsaneh Fazly |
| 17:00 - 17:15 | ***Best Paper Awards and Closing*** |

# W3 - Seventh Conference on Machine Translation

**Organizers:**
Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Marco Turchi, Marcos Zampieri

https://www.statmt.org/wmt22/
Venue: Capital Suite 13
**Wednesday, December 7, 2022 - Thursday, December 8, 2022**

The Seventh Conference on Machine Translaton (WMT) features research papers and shared tasks on any aspect of machine translation.
This year's conference will feature the following shared tasks: a general MT translation task (former News task), a biomedical translation task, a code-mixing translation task, an unsupervised and very low resource translation task, an automatic post-editing task, a sign language translation task, a word-level autocompletion task, a metrics task (assess MT quality with or without reference translation), a quality estimation task (assess MT quality without access to any reference), an MT efficiency task, a translation suggestion task (generate the suggestions for incorrect spans of the MT sentence), a chat translation task>, a sign language translation task, a large-scale multilingual MT task, and an unsupervised and very low resource supervised translation task

In addition to the shared tasks, the conference will also feature scientific papers on topics related to MT.

An invited talk by Ondrej Bojar (Charles University, Prague) on "Speech Translation: When Two Superhuman Technologies Combined Fail" will be presented on December 8 at 2pm local time.

| | |
|---|---|
| 09:00 - 09:10 | ***Opening Remarks*** |
| 09:10 - 10:30 | ***Session 1 — Shared Task Overview Papers I*** |
| 09:10-09:30 | *Findings of the 2022 Conference on Machine Translation (WMT22)*<br>Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel and Maja Popović |
| 09:30-09:50 | *Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust*<br>Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie and André F. T. Martins |
| 09:50-10:00 | *Findings of the WMT 2022 Shared Task on Quality Estimation*<br>Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. De Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins and Lucia Specia |
| 10:00-10:15 | *Findings of the WMT 2022 Shared Task on Efficient Translation*<br>Kenneth Heafield, Biao Zhang, Graeme Nail, Jelmer Van Der Linde and Nikolay Bogoychev |
| 10:15-10:30 | *Findings of the WMT 2022 Shared Task on Automatic Post-Editing* |

| | |
|---|---|
| | Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri and Marco Turchi |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 12:40 | ***Session 2 — Research Papers on Evaluation and Bias*** |
| 11:00-11:20 | *Embarrassingly Easy Document-Level MT Metrics: How to Convert Any Pretrained Metric into a Document-Level Metric*<br>Giorgos Vernikos, Brian Thompson, Prashant Mathur and Marcello Federico |
| 11:20-11:40 | *Searching for a Higher Power in the Human Evaluation of MT*<br>Johnny Wei, Tom Kocmi and Christian Federmann |
| 11:40-12:00 | *Analyzing the Use of Influence Functions for Instance-Specific Data Filtering in Neural Machine Translation*<br>Tsz Kin Lam, Eva Hasler and Felix Hieber |
| 12:00-12:20 | *Continuous Rating as Reliable Human Evaluation of Simultaneous Speech Translation*<br>Dávid Javorský, Dominik Macháček and Ondřej Bojar |
| 12:20-12:40 | *Gender Bias Mitigation for NMT Involving Genderless Languages*<br>Ander Corral and Xabier Saralegi |
| 12:40 - 14:00 | ***Lunch Break*** |
| 14:00 - 15:20 | ***Session 3 — Research Papers on Multilingual, Multimodal, Multidomain Translation*** |
| 14:00-14:20 | *Exploring the Benefits and Limitations of Multilinguality for Non-autoregressive Machine Translation*<br>Sweta Agrawal, Julia Kreutzer and Colin Cherry |
| 14:20-14:40 | *Learning an Artificial Language for Knowledge-Sharing in Multilingual Translation*<br>Danni Liu and Jan Niehues |
| 14:40-15:00 | *Don't Discard Fixed-Window Audio Segmentation in Speech-to-Text Translation*<br>Chantal Amrhein and Barry Haddow |
| 15:00-15:20 | *Additive Interventions Yield Robust Multi-Domain Machine Translation Models*<br>Elijah Rippeth and Matt Post |
| 15:20-15:40 | *Test Set Sampling Affects System Rankings: Expanded Human Evaluation of WMT20 English-Inuktitut Systems*<br>Rebecca Knowles and Chi-kiu Lo |
| 15:40-16:00 | *Can Domains Be Transferred across Languages in Multi-Domain Multilingual Neural Machine Translation?*<br>Thuy-trang Vu, Shahram Khadivi, Xuanli He, Dinh Phung and Gholamreza Haffari |
| 15:20 - 16:00 | ***Coffee Break*** |
| 16:00 - 17:30 | ***Session 4 — Shared Task Sytem Description Posters I*** |
| 16:00 - 17:30 | ***News Translation Task*** |
| 16:00-17:30 | *Inria-ALMAnaCH at WMT 2022: Does Transcription Help Cross-Script Machine Translation?*<br>Jesujoba Alabi, Lydia Nishimwe, Benjamin Muller, Camille Rey, Benoît Sagot and Rachel Bawden |
| 16:00-17:30 | *NAIST-NICT-TIT WMT22 General MT Task Submission*<br>Hiroyuki Deguchi, Kenji Imamura, Masahiro Kaneko, Yuto Nishida, Yusuke Sakai, Justin Vasselli, Huy Hien Vu and Taro Watanabe |
| 16:00-17:30 | *Samsung R&D Institute Poland Participation in WMT 2022*<br>Adam Dobrowolski, Mateusz Klimaszewski, Adam Myśliwy, Marcin Szymański, Jakub Kowalski, Kornelia Szypuła, Paweł Przewłocki and Paweł Przybysz |

16:00-17:30 *Tencent AI Lab - Shanghai Jiao Tong University Low-Resource Translation System for the WMT22 Translation Task*
Zhiwei He, Xing Wang, Zhaopeng Tu, Shuming Shi and Rui Wang

16:00-17:30 *Lan-Bridge MT's Participation in the WMT 2022 General Translation Shared Task*
Bing Han, Yangjian Wu, Gang Hu and Qiulin Chen

16:00-17:30 *Manifold's English-Chinese System at WMT22 General MT Task*
Chang Jin, Tingxun Shi, Zhengshan Xue and Xiaodong Lin

16:00-17:30 *CUNI-Bergamot Submission at WMT22 General Translation Task*
Josef Jon, Martin Popel and Ondřej Bojar

16:00-17:30 *KYB General Machine Translation Systems for WMT22*
Shivam Kalkar, Yoko Matsuzaki and Ben Li

16:00-17:30 *The AISP-SJTU Translation System for WMT 2022*
Guangfeng Liu, Qinpei Zhu, Xingyu Chen, Renjie Feng, Jianxin Ren, Renshou Wu, Qingliang Miao, Rui Wang and Kai Yu

16:00-17:30 *NT5 at WMT 2022 General Translation Task*
Makoto Morishita, Keito Kudo, Yui Oka, Katsuki Chousa, Shun Kiyono, Sho Takase and Jun Suzuki

16:00-17:30 *Adam Mickiewicz University at WMT 2022: NER-Assisted and Quality-Aware Neural Machine Translation*
Artur Nowakowski, Gabriela Pałka, Kamil Guttmann and Mikołaj Pokrywka

16:00-17:30 *Evaluating Corpus Cleanup Methods in the WMT'22 News Translation Task*
Marilena Malli and George Tambouratzis

16:00-17:30 *PROMT Systems for WMT22 General Translation Task*
Alexander Molchanov, Vladislav Kovalenko and Natalia Makhamalkina

16:00-17:30 *eTranslation's Submissions to the WMT22 General Machine Translation Task*
Csaba Oravecz, Katina Bontcheva, David Kolovratnìk, Bogomil Kovachev and Christopher Scott

16:00-17:30 *CUNI Systems for the WMT 22 Czech-Ukrainian Translation Task*
Martin Popel, Jindřich Libovický and Jindřich Helcl

16:00-17:30 *The ARC-NKUA Submission for the English-Ukrainian General Machine Translation Shared Task at WMT22*
Dimitrios Roussis and Vassilis Papavassiliou

16:00-17:30 *The NiuTrans Machine Translation Systems for WMT22*
Weiqiao Shan, Zhiquan Cao, Yuchen Han, Siming Wu, Yimin Hu, Jie Wang, Yi Zhang, Hou Baoyu, Hang Cao, Chenghao Gao, Xiaowen Liu, Tong Xiao, Anxiang Ma and Jingbo Zhu

16:00-17:30 *Teaching Unseen Low-resource Languages to Large Translation Models*
Maali Tars, Taido Purason and Andre Tättar

16:00-17:30 *DUTNLP Machine Translation System for WMT22 General MT Task*
Ting Wang, Huan Liu, Junpeng Liu and Degen Huang

16:00-17:30 *HW-TSC's Submissions to the WMT 2022 General Machine Translation Shared Task*
Daimeng Wei, Zhiqiang Rao, Zhanglin Wu, Shaojun Li, Yuanchang Luo, Yuhao Xie, Xiaoyu Chen, Hengchao Shang, Zongyao Li, Zhengzhe Yu, Jinlong Yang, Miaomiao Ma, Lizhi Lei, Hao Yang and Ying Qin

16:00-17:30 *Vega-MT: The JD Explore Academy Machine Translation System for WMT22*
Changtong Zan, Keqin Peng, Liang Ding, Baopu Qiu, Boan Liu, Shwai He, Qingyu Lu, Zheng Zhang, Chuang Liu, Weifeng Liu, Yibing Zhan and Dacheng Tao

16:00-17:30 *No Domain Left behind*
Hui Zeng

16:00-17:30 *GTCOM Neural Machine Translation Systems for WMT22*
Hao Zong and Chao Bei

16:00 - 17:30 ***Test Suites***

16:00-17:30 *Linguistically Motivated Evaluation of the 2022 State-of-the-art Machine Translation Systems for Three Language Directions*
Vivien Macketanz, Shushen Manakhimova, Eleftherios Avramidis, Ekaterina Lapshinova-koltunski, Sergei Bagdasarov and Sebastian Möller

16:00-17:30 *Automated Evaluation Metric for Terminology Consistency in MT*
Kirill Semenov and Ondřej Bojar

16:00-17:30 *Test Suite Evaluation: Morphological Challenges and Pronoun Translation*
Marion Weller-di Marco and Alexander Fraser

16:00 - 17:30 ***Metrics Task***

16:00-17:30 *Robust MT Evaluation with Sentence-level Multilingual Augmentation*
Duarte Alves, Ricardo Rei, Ana C Farinha, José G. C. De Souza and André F. T. Martins

16:00-17:30 *ACES: Translation Accuracy Challenge Sets for Evaluating Machine Translation Metrics*
Chantal Amrhein, Nikita Moghe and Liane Guillou

16:00-17:30 *Linguistically Motivated Evaluation of Machine Translation Metrics Based on a Challenge Set*
Eleftherios Avramidis and Vivien Macketanz

16:00-17:30 *Exploring Robustness of Machine Translation Metrics: A Study of Twenty-Two Automatic Metrics in the WMT22 Metric Task*
Xiaoyu Chen, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Ting Zhu, Mengli Zhu, Ning Xie, Lizhi Lei, Shimin Tao, Hao Yang and Ying Qin

16:00-17:30 *MS-COMET: More and Better Human Judgements Improve Metric Performance*
Tom Kocmi, Hitokazu Matsushita and Christian Federmann

16:00-17:30 *Partial Could Be Better than Whole. HW-TSC 2022 Submission for the Metrics Shared Task*
Yilun Liu, Xiaosong Qiao, Zhanglin Wu, Su Chang, Min Zhang, Yanqing Zhao, Song Peng, Shimin Tao, Hao Yang, Ying Qin, Jiaxin Guo, Minghan Wang, Yinglu Li, Peng Li and Xiaofeng Zhao

16:00-17:30 *Unsupervised Embedding-based Metric for MT Evaluation with Improved Human Correlation*
Ananya Mukherjee and Manish Shrivastava

16:00-17:30 *REUSE: REference-free UnSupervised Quality Estimation Metric*
Ananya Mukherjee and Manish Shrivastava

16:00-17:30 *MaTESe: Machine Translation Evaluation as a Sequence Tagging Problem*
Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo and Roberto Navigli

16:00-17:30 *COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task*
Ricardo Rei, José G. C. De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur and André F. T. Martins

16:00-17:30 *Alibaba-Translate China's Submission for WMT2022 Metrics Shared Task*
Yu Wan, Keqin Bao, Dayiheng Liu, Baosong Yang, Derek F. Wong, Lidia S. Chao, Wenqiang Lei and Jun Xie

16:00 - 17:30 ***Quality Estimation Task***

16:00-17:30 *Quality Estimation via Backtranslation at the WMT 2022 Quality Estimation Task*
Sweta Agrawal, Nikita Mehandru, Niloufar Salehi and Marine Carpuat

16:00-17:30 *Alibaba-Translate China's Submission for WMT 2022 Quality Estimation Shared Task*
Keqin Bao, Yu Wan, Dayiheng Liu, Baosong Yang, Wenqiang Lei, Xiangnan He, Derek F. Wong and Jun Xie

16:00-17:30 *KU X Upstage's Submission for the WMT22 Quality Estimation: Critical Error Detection Shared Task*
Sugyeong Eo, Chanjun Park, Hyeonseok Moon, Jaehyung Seo and Heuiseok Lim

16:00-17:30 *NJUNLP's Participation for the WMT2022 Quality Estimation Shared Task*
Xiang Geng, Yu Zhang, Shujian Huang, Shimin Tao, Hao Yang and Jiajun Chen

16:00-17:30 *BJTU-Toshiba's Submission to WMT22 Quality Estimation Shared Task*
Hui Huang, Hui Di, Chunyou Li, Hanming Wu, Kazushige Ouchi, Yufeng Chen, Jian Liu and Jinan Xu

16:00-17:30 *Papago's Submission to the WMT22 Quality Estimation Shared Task*
Seunghyun Lim and Jeonghyeok Park

16:00-17:30 *CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task*
Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. De Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie and André F. T. Martins

16:00-17:30 *CrossQE: HW-TSC 2022 Submission for the Quality Estimation Shared Task*
Shimin Tao, Su Chang, Ma Miaomiao, Hao Yang, Xiang Geng, Shujian Huang, Min Zhang, Jiaxin Guo, Minghan Wang and Yinglu Li

16:00-17:30 *Welocalize-ARC/NKUA's Submission to the WMT 2022 Quality Estimation Shared Task*
Eirini Zafeiridou and Sokratis Sofianopoulos

16:00 - 17:30 ***Efficient Translaton Task***

16:00-17:30 *Edinburgh's Submission to the WMT 2022 Efficiency Task*
Nikolay Bogoychev, Maximiliana Behnke, Jelmer Van Der Linde, Graeme Nail, Kenneth Heafield, Biao Zhang and Sidharth Kashyap

16:00-17:30 *CUNI Non-Autoregressive System for the WMT 22 Efficient Translation Shared Task*
Jindřich Helcl

16:00-17:30 *The RoyalFlush System for the WMT 2022 Efficiency Task*
Bo Qin, Aixin Jia, Qiang Wang, Jianning Lu, Shuqin Pan, Haibo Wang and Ming Chen

16:00-17:30 *HW-TSC's Submission for the WMT22 Efficiency Task*
Hengchao Shang, Ting Hu, Daimeng Wei, Zongyao Li, Xianzhi Yu, Jianfei Feng, Ting Zhu, Lizhi Lei, Shimin Tao, Hao Yang, Ying Qin, Jinlong Yang, Zhiqiang Rao and Zhengzhe Yu

16:00 - 17:30 ***Automatic Post-Editing Task***

16:00-17:30 *IIT Bombay's WMT22 Automatic Post-Editing Shared Task Submission*
Sourabh Deoghare and Pushpak Bhattacharyya

16:00-17:30 *LUL's WMT22 Automatic Post-Editing Shared Task Submission*
Xiaoying Huang, Xingrui Lou, Fan Zhang and Tu Mei

16:00 - 17:30 ***Papers from the Findings of the EMNLP***

16:00-17:30 *Translating Hanja Historical Documents to Contemporary Korean and English*
Juhee Son, Jiho Jin, Haneul Yoo, JinYeong Bak, Kyunghyun Cho and Alice Oh

16:00-17:30 *Data Selection Curriculum for Neural Machine Translation*
Tasnim Mohiuddin, Philipp Koehn, Vishrav Chaudhary, James Cross, Shruti Bhosale and Shafiq Joty

16:00-17:30 *Bitext Mining Using Distilled Sentence Representations for Low-Resource Languages*
Kevin Heffernan, Onur Çelebi and Holger Schwenk

16:00-17:30 *$m^4$ Adapter: Multilingual Multi-Domain Adaptation for Machine Translation with a Meta-Adapter*
Wen Lai, Alexandra Chronopoulou and Alexander Fraser

16:00-17:30 *What Do Compressed Multilingual Machine Translation Models Forget?*
Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson and Laurent Besacier

16:00-17:30 *Improving HowNet-Based Chinese Word Sense Disambiguation with Translations*
Xiang Zhang, Bradley Hauer and Grzegorz Kondrak

16:00-17:30 *Finding Memo: Extractive Memorization in Constrained Sequence Generation Tasks*
Vikas Raunak and Arul Menezes

16:00-17:30 *SALTED: A Framework for SAlient Long-tail Translation Error Detection*
Vikas Raunak, Matt Post and Arul Menezes

16:00-17:30 *SALTED: A Framework for SAlient Long-tail Translation Error Detection*
Vikas Raunak, Matt Post and Arul Menezes

16:00-17:30 *SALTED: A Framework for SAlient Long-tail Translation Error Detection*
Vikas Raunak, Matt Post and Arul Menezes

16:00-17:30 *SALTED: A Framework for SAlient Long-tail Translation Error Detection*
Vikas Raunak, Matt Post and Arul Menezes

16:00-17:30 *Data Cartography for Low-Resource Neural Machine Translation*
Aquia Richburg and Marine Carpuat

16:00-17:30 *Improving Zero-Shot Multilingual Translation with Universal Representations and Cross-Mapping*
Shuhao Gu and Yang Feng

16:00-17:30 *Guiding Neural Machine Translation with Semantic Kernels*
Ping Guo, Yue Hu, Xiangpeng Wei, Yubing Ren, Yunpeng Li, Luxi Xing and Yuqiang Xie

16:00-17:30 *Does Simultaneous Speech Translation need Simultaneous Models?*
Sara Papi, Marco Gaido, Matteo Negri and Marco Turchi

16:00-17:30 *Utilizing Language-Image Pretraining for Efficient and Robust Bilingual Word Alignment*
Tuan Dinh, Jy-yong Sohn, Shashank Rajput, Timothy Ossowski, Yifei Ming, Junjie Hu, Dimitris Papailiopoulos and Kangwook Lee

11:00 - 12:40 ***Session 6 — Research Papers on Practical Aspects of Machine Translation***

11:00-11:20 *Focused Concatenation for Context-Aware Neural Machine Translation*
Lorenzo Lupo, Marco Dinarelli and Laurent Besacier

11:20-11:40 *Does Sentence Segmentation Matter for Machine Translation?*
Rachel Wicks and Matt Post

11:40-12:00 *Revisiting Locality Sensitive Hashing for Vocabulary Selection in Fast Neural Machine Translation*
Hieu Hoang, Marcin Junczys-dowmunt, Roman Grundkiewicz and Huda Khayrallah

12:00-12:20 *Too Brittle to Touch: Comparing the Stability of Quantization and Distillation towards Developing Low-Resource MT Models*
Harshita Diddee, Sandipan Dandapat, Monojit Choudhury, Tanuja Ganu and Kalika Bali

12:20-12:40 *Data Augmentation for Inline Tag-Aware Neural Machine Translation*
Yonghyun Ryu, Yoonjung Choi and Sangha Kim

12:40 - 14:00 ***Lunch Break***

14:00 - 15:30 ***Session 7 — Invited Talk by Ondrej Bojar***

09:00 - 09:10 ***Session 5 — Shared Task Overview Papers II***

09:00-09:10 *Findings of the WMT 2022 Biomedical Translation Shared Task: Monolingual Clinical Case Reports*

Mariana Neves, Antonio Jimeno Yepes, Amy Siu, Roland Roller, Philippe Thomas, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, Giorgio Maria Di Nunzio, Federica Vezzani, Christel Gerardin, Rachel Bawden, Darryl Johan Estrada, Salvador Lima-lopez, Eulalia Farremaduel, Martin Krallinger, Cristian Grozea and Aurelie Neveol

09:10-09:20 *Findings of the WMT 2022 Shared Task on Chat Translation*
Ana C Farinha, M. Amin Farajian, Marianna Buchicchio, Patrick Fernandes, José G. C. De Souza, Helena Moniz and André F. T. Martins

09:20-09:35 *Findings of the First WMT Shared Task on Sign Language Translation (WMT-SLT22)*
Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-miserez and Katja Tissi

09:35-09:50 *Findings of the WMT'22 Shared Task on Large-Scale Machine Translation Evaluation for African Languages*
David Adelani, Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Akshita Bhagia, Marta R. Costa-jussà, Jesse Dodge, Fahim Faisal, Christian Federmann, Natalia Fedorova, Francisco Guzmán, Sergey Koshelev, Jean Maillard, Vukosi Marivate, Jonathan Mbuya, Alexandre Mourachko, Safiyyah Saleem, Holger Schwenk and Guillaume Wenzek

09:50-10:00 *Findings of the WMT 2022 Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT*
Marion Weller-di Marco and Alexander Fraser

10:00-10:10 *Overview and Results of MixMT Shared-Task at WMT 2022*
Vivek Srivastava and Mayank Singh

10:10-10:20 *Findings of the Word-Level AutoCompletion Shared Task in WMT 2022*
Francisco Casacuberta, George Foster, Guoping Huang, Philipp Koehn, Geza Kovacs, Lemao Liu, Shuming Shi, Taro Watanabe and Chengqing Zong

10:20-10:30 *Findings of the WMT 2022 Shared Task on Translation Suggestion*
Zhen Yang, Fandong Meng, Yingxue Zhang, Ernan Li and Jie Zhou

10:30 - 11:00 ***Coffee Break***

15:30 - 16:00 ***Coffee Break***

16:00 - 17:30 ***Session 8 — Shared Task Sytem Description Papers II***

16:00 - 17:30 ***Biomedical Translation Task***

16:00-17:30 *The SPECTRANS System Description for the WMT22 Biomedical Task*
Nicolas Ballier, Jean-baptiste Yunès, Guillaume Wisniewski, Lichao Zhu and Maria Zimina

16:00-17:30 *SRT's Neural Machine Translation System for WMT22 Biomedical Translation Task*
Yoonjung Choi, Jiho Shin, Yonghyun Ryu and Sangha Kim

16:00-17:30 *Examining Large Pre-Trained Language Models for Machine Translation: What You Don't Know about It*
Lifeng Han, Gleb Erofeev, Irina Sorokina, Serge Gladkoff and Goran Nenadic

16:00-17:30 *Summer: WeChat Neural Machine Translation Systems for the WMT22 Biomedical Translation Task*
Ernan Li, Fandong Meng and Jie Zhou

16:00-17:30 *Optum's Submission to WMT22 Biomedical Translation Tasks*
Sahil Manchanda and Saurabh Bhagwat

16:00-17:30 *Huawei BabelTar NMT at WMT22 Biomedical Translation Task: How We Further Improve Domain-specific NMT*
Weixuan Wang, Xupeng Meng, Suqing Yan, Ye Tian and Wei Peng

16:00-17:30 *HW-TSC Translation Systems for the WMT22 Biomedical Translation Task*

Zhanglin Wu, Jinlong Yang, Zhiqiang Rao, Zhengzhe Yu, Daimeng Wei, Xiaoyu Chen, Zongyao Li, Hengchao Shang, Shaojun Li, Ming Zhu, Yuanchang Luo, Yuhao Xie, Miaomiao Ma, Ting Zhu, Lizhi Lei, Song Peng, Hao Yang and Ying Qin

16:00 - 17:30 ***Chat Translation Task***

16:00-17:30 *Unbabel-IST at the WMT Chat Translation Shared Task*
João Alves, Pedro Henrique Martins, José G. C. De Souza, M. Amin Farajian and André F. T. Martins

16:00-17:30 *Investigating Effectiveness of Multi-Encoder for Conversational Neural Machine Translation*
Baban Gain, Ramakrishna Appicharla, Soumya Chennabasavaraj, Nikesh Garera, Asif Ekbal and Muthusamy Chelliah

16:00-17:30 *BJTU-WeChat's Systems for the WMT22 Chat Translation Task*
Yunlong Liang, Fandong Meng, Jinan Xu, Yufeng Chen and Jie Zhou

16:00-17:30 *HW-TSC Translation Systems for the WMT22 Chat Translation Task*
Jinlong Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, Xiaoyu Chen, Zhengzhe Yu, Zhiqiang Rao, Shaojun Li, Zhanglin Wu, Yuhao Xie, Yuanchang Luo, Ting Zhu, Yanqing Zhao, Lizhi Lei, Hao Yang and Ying Qin

16:00 - 17:30 ***Sign Language Translation Task***

16:00-17:30 *Clean Text and Full-Body Transformer: Microsoft's Submission to the WMT22 Shared Task on Sign Language Translation*
Subhadeep Dey, Abhilash Pal, Cyrine Chaabani and Oscar Koller

16:00-17:30 *Spatio-temporal Sign Language Representation and Translation*
Yasser Hamidullah, Josef Van Genabith and Cristina España-bonet

16:00-17:30 *Experimental Machine Translation of the Swiss German Sign Language via 3D Augmentation of Body Keypoints*
Lorenz Hufe and Eleftherios Avramidis

16:00-17:30 *TTIC's WMT-SLT 22 Sign Language Translation System*
Bowen Shi, Diane Brentari, Gregory Shakhnarovich and Karen Livescu

16:00-17:30 *Tackling Low-Resourced Sign Language Translation: UPC at WMT-SLT 22*
Laia Tarres, Gerard I. Gállego, Xavier Giro-i-nieto and Jordi Torres

16:00 - 17:30 ***African Languages Translation Task***

16:00-17:30 *Separating Grains from the Chaff: Using Data Filtering to Improve Multilingual Translation for Low-Resourced African Languages*
Idris Abdulmumin, Michael Beukman, Jesujoba Alabi, Chris Chinenye Emezue, Everlyn Chimoto, Tosin Adewumi, Shamsuddeen Muhammad, Mofetoluwa Adeyemi, Oreen Yousuf, Sahib Singh and Tajuddeen Gwadabe

16:00-17:30 *Language Adapters for Large-Scale MT: The GMU System for the WMT 2022 Large-Scale Machine Translation Evaluation for African Languages Shared Task*
Md Mahfuz Ibn Alam and Antonios Anastasopoulos

16:00-17:30 *Samsung Research Philippines - Datasaur AI's Submission for the WMT22 Large Scale Multilingual Translation Task*
Jan Christian Blaise Cruz and Lintang Sutawika

16:00-17:30 *University of Cape Town's WMT22 System: Multilingual Machine Translation for Southern African Languages*
Khalid Elmadani, Francois Meyer and Jan Buys

16:00-17:30 *Tencent's Multilingual Machine Translation System for WMT22 Large-Scale African Languages*
Wenxiang Jiao, Zhaopeng Tu, Jiarui Li, Wenxuan Wang, Jen-tse Huang and Shuming Shi

16:00-17:30 *DENTRA: Denoising and Translation Pre-training for Multilingual Machine Translation*

Samta Kamboj, Sunil Kumar Sahu and Neha Sengupta

16:00-17:30 *The VolcTrans System for WMT22 Multilingual Machine Translation Task*
Xian Qian, Kai Hu, Jiaqiang Wang, Yifeng Liu, Xingyuan Pan, Jun Cao and Mingxuan Wang

16:00-17:30 *WebCrawl African : A Multilingual Parallel Corpora for African Languages*
Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Abhinav Mishra, Prashant Banjare, Prasanna K R and Chitra Viswanathan

16:00-17:30 *ANVITA-African: A Multilingual Neural Machine Translation System for African Languages*
Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Prasanna K R and Chitra Viswanathan

16:00 - 17:30 ***Unsupervised and Very Low Resource Translation Task***

16:00-17:30 *HW-TSC Systems for WMT22 Very Low Resource Supervised MT Task*
Shaojun Li, Yuanchang Luo, Daimeng Wei, Zongyao Li, Hengchao Shang, Xiaoyu Chen, Zhanglin Wu, Jinlong Yang, Zhiqiang Rao, Zhengzhe Yu, Yuhao Xie, Lizhi Lei, Hao Yang and Ying Qin

16:00-17:30 *Unsupervised and Very-Low Resource Supervised Translation on German and Sorbian Variant Languages*
Rahul Tangsali, Aditya Vyawahare, Aditya Mandke, Onkar Litake and Dipali Kadam

16:00-17:30 *MUNI-NLP Systems for Lower Sorbian-German and Lower Sorbian-Upper Sorbian Machine Translation @ WMT22*
Edoardo Signoroni and Pavel Rychlý

16:00-17:30 *The AIC System for the WMT 2022 Unsupervised MT and Very Low Resource Supervised MT Task*
Ahmad Shapiro, Mahmoud Salama, Omar Abdelhakim, Mohamed Fayed, Ayman Khalafallah and Noha Adly

16:00 - 17:30 ***Code-Mixed Translation Task***

16:00-17:30 *NICT at MixMT 2022: Synthetic Code-Mixed Pre-training and Multi-way Fine-tuning for Hinglish–English Translation*
Raj Dabre

16:00-17:30 *The NiuTrans Machine Translation Systems for WMT22*
Weiqiao Shan, Zhiquan Cao, Yuchen Han, Siming Wu, Yimin Hu, Jie Wang, Yi Zhang, Hou Baoyu, Hang Cao, Chenghao Gao, Xiaowen Liu, Tong Xiao, Anxiang Ma and Jingbo Zhu

16:00-17:30 *MUCS@MixMT: IndicTrans-based Machine Translation for Hinglish Text*
Asha Hegde and Shashirekha Lakshmaiah

16:00-17:30 *SIT at MixMT 2022: Fluent Translation Built on Giant Pre-trained Models*
Abdul Khan, Hrishikesh Kanade, Girish Budhrani, Preet Jhanglani and Jia Xu

16:00-17:30 *The University of Edinburgh's Submission to the WMT22 Code-Mixing Shared Task (MixMT)*
Faheem Kirefu, Vivek Iyer, Pinzhen Chen and Laurie Burchell

16:00-17:30 *Findings of the WMT 2022 Biomedical Translation Shared Task: Monolingual Clinical Case Reports*
Mariana Neves, Antonio Jimeno Yepes, Amy Siu, Roland Roller, Philippe Thomas, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, Giorgio Maria Di Nunzio, Federica Vezzani, Christel Gerardin, Rachel Bawden, Darryl Johan Estrada, Salvador Lima-lopez, Eulalia Farremaduel, Martin Krallinger, Cristian Grozea and Aurelie Neveol

16:00-17:30 *Domain Curricula for Code-Switched MT at MixMT 2022*
Lekan Raheem, Maab Elrashid, Melvin Johnson and Julia Kreutzer

16:00 - 17:30 ***Word-Level Autocompletion Task***

16:00-17:30 *Lingua Custodia's Participation at the WMT 2022 Word-Level Auto-completion Shared Task*
Melissa Ailem, Jingshu Liu, Jean-gabriel Barthelemy and Raheel Qader

16:00-17:30 *Translation Word-Level Auto-Completion: What Can We Achieve Out of the Box?*
Yasmin Moslem, Rejwanul Haque and Andy Way

16:00-17:30 *PRHLT's Submission to WLAC 2022*
Angel Navarro, Miguel Domingo and Francisco Casacuberta

16:00-17:30 *IIGROUP Submissions for WMT22 Word-Level AutoCompletion Task*
Cheng Yang, Siheng Li, Chufan Shi and Yujiu Yang

16:00-17:30 *HW-TSC's Submissions to the WMT22 Word-Level Auto Completion Task*
Hao Yang, Hengchao Shang, Zongyao Li, Daimeng Wei, Xianghui He, Xiaoyu Chen, Zhengzhe Yu, Jiaxin Guo, Jinlong Yang, Shaojun Li, Yuanchang Luo, Yuhao Xie, Lizhi Lei and Ying Qin

16:00 - 17:30 ***Translation Suggestion Task***

16:00-17:30 *TSMind: Alibaba and Soochow University's Submission to the WMT22 Translation Suggestion Task*
Xin Ge, Ke Wang, Jiayi Wang, Nini Xiao, Xiangyu Duan, Yu Zhao and Yuqi Zhang

16:00-17:30 *Transn's Submissions to the WMT22 Translation Suggestion Task*
Mao Hongbao, Zhang Wenbo, Cai Jie and Cheng Jianwei

16:00-17:30 *Improved Data Augmentation for Translation Suggestion*
Hongxiao Zhang, Siyu Lai, Songming Zhang, Hui Huang, Yufeng Chen, Jinan Xu and Jian Liu

# W4 - The First Workshop on Ever Evolving NLP

**Organizers:**
Francesco Barbieri, Jose Camacho-Collados, Bhuwan Dhingra, Luis Espinosa-Anke, Elena Gribovskaya, Angeliki Lazaridou, Daniel Loureiro, Leonardo Neves

https://sites.google.com/view/evonlp/
Venue: Capital Suite 3
**Wednesday, December 7, 2022**

EvoNLP is a forum to discuss the challenges posed by the dynamic nature of language in the specific context of the current NLP paradigm, dominated by language models. This year, the program includes a regular session, a session dedicated to the time-aware Word-in-Context classification shared task, as well as non-archival presentations and Findings of EMNLP papers. Finally, are delighted to have the following renowned invited speakers: Eunsol Choi, Jacob Eisenstein, Adam Jatowt, Ozan Sener and Nazneen Rajani.

| | |
|---|---|
| 09:00 - 09:30 | ***Opening remarks*** |
| 09:30 - 10:00 | ***Jacob Eisenstein - What can we learn from language change?*** |
| 10:00 - 10:30 | ***Eunsol Choi - Knowledge-rich NLP models in a dynamic real world*** |
| 10:30 - 11:00 | ***Adam Jatowt - Automatic Question Answering over Temporal News Collections*** |
| 11:00 - 12:30 | ***Workshop poster session (virtual and on-site)*** |
| 12:30 - 14:00 | ***Lunch break*** |
| 14:00 - 15:00 | ***Findings and non-archival session*** |
| 15:00 - 15:30 | ***Coffee break*** |
| 15:30 - 16:00 | ***Nazneen Rajani - Takeaways from a systematic study of 75K models on Hugging Face*** |
| 16:00 - 16:30 | ***Ozan Sener - Going from Continual Learning Algorithms to Continual Learning Systems*** |
| 16:30 - 17:00 | ***Workshop oral session*** |
| 17:00 - 17:30 | ***Shared task session*** |
| 17:30 - 18:00 | ***Best paper awards and closing*** |

# W5 - 2nd Workshop on Natural Language Generation, Evaluation, and Metrics

**Organizers:**
Antoine Bosselut, Khyathi Chandu, Kaustubh Dhole, Varun Gangal, Sebastian Gehrmann, Yacine Jernite, Jekaterina Novikova, Laura Perez-Beltrachini

https://gem-benchmark.com/workshop
Venue: Capital Suite 5
**Wednesday, December 7, 2022**

Natural language generation is one of the most active research fields within NLP and its barrier of entry has reduced dramatically. While applying supervised state of the art models to new data sets is becoming easier, the evaluation of models is becoming more challenging as models can produce completely fluent but meaningless or subtly flawed output. This leads to a disconnect between real-world needs of generation models and published research. Most of the disconnect can be bridged via in-depth evaluation and documentation of both data and models.
To that end, the GEM workshop has three core goals: (1) Encourage the development of (semi-) automatic model audits and improved human evaluation strategies, (2) Popularize model evaluations in languages beyond English, (3) Provide a platform for discussions around evaluations to bridge the gap between industry and academia. All 80+ committee members are collaborating on developing and improving the GEM benchmark infrastructure which enables model comparisons on different datasets, assessment of evaluation metrics, and leads to a fast application of findings from the workshop.

Based on this infrastructure, the workshop features a modeling shared task with a focus on low-resource generation, a theme which was suggested by participants of the first shared task. All shared task submissions are evaluated through a combination of human and automatic metrics and model outputs and ratings will be made publicly available as part of the GEM Resources (https://gem-benchmark.com/resources) and may be used for future evaluation shared tasks.

| | |
|---|---|
| 09:00 - 10:30 | ***Opening Remarks and Keynote (Sean Welleck)*** |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 12:30 | ***Talk Session*** |
| 12:30 - 14:00 | ***Lunch Break*** |
| 14:00 - 15:30 | ***Poster Session*** |
| 15:30 - 16:00 | ***Coffee Break*** |
| 16:00 - 17:00 | ***Keynote (Timo Schick)*** |
| 17:00 - 18:30 | ***Talk Session*** |
| 20:00 - 21:00 | ***Virtual Keynote (Emily Dinan)*** |
| 21:00 - 22:30 | ***Virtual Poster Session*** |

# W6 - 13th International Workshop on Health Text Mining and Information Analysis

**Organizers:**
Alberto Lavelli, Eben Holderness, Antonio Jimeno Yepes, Anne-Lyse Minard, James Pustejovsky, Fabio Rinaldi

https://louhi2022.fbk.eu/
Venue: Capital Suite 21C
**Wednesday, December 7, 2022**

The 13th International Workshop on Health Text Mining and Information Analysis (LOUHI) provides an interdisciplinary forum for researchers interested in automated processing of health documents. Health documents encompass electronic health records, clinical guidelines, spontaneous reports for pharmacovigilance, biomedical literature, health forums/blogs or any other type of health-related documents.
The LOUHI workshop series fosters interactions between the Computational Linguistics, Medical Informatics and Artificial Intelligence communities. It started in 2008 in Turku, Finland and has been organized 12 times: LOUHI 2010 was co-located with NAACL in Los Angeles, CA; LOUHI 2011 was co-located with Artificial Intelligence in Medicine (AIME) in Bled, Slovenia; LOUHI 2013 was held in Sydney, Australia during NICTA Techfest; LOUHI 2014 was co-located with EACL in Gothenburg, Sweden; LOUHI 2015 was co-located with EMNLP in Lisbon, Portugal; LOUHI 2016 was co-located with EMNLP in Austin, Texas; LOUHI 2017 was held in Sydney, Australia; LOUHI 2018 was co-located with EMNLP in Brussels, Belgium; LOUHI 2019 was co-located with EMNLP in Hong Kong; LOUHI 2020 was co-located with EMNLP; and LOUHI 2021 was co-located with EACL.

The aim of the LOUHI 2022 workshop is to bring together research work on topics related to health documents, particularly emphasizing multidisciplinary aspects of health documentation and the interplay between nursing and medical sciences, information systems, computational linguistics and computer science.

| | |
|---|---|
| 09:00 - 09:10 | ***Opening Remarks*** |
| 09:10 - 10:00 | ***Invited Talk*** |
| 10:00 - 10:30 | ***TBD*** |
| 10:00-10:30 | *Can Current Explainability Help Provide References in Clinical Notes to Support Humans Annotate Medical Codes?*<br>Byung-Hak Kim, Zhongfen Deng, Philip Yu and Varun Ganapathi |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 12:30 | ***Session 2*** |
| 11:00-11:30 | *Assessing the Limits of Straightforward Models for Nested Named Entity Recognition in Spanish Clinical Narratives*<br>Matias Rojas, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Jocelyn Dunstan and Marta Villegas |
| 11:30-12:00 | *A Quantitative and Qualitative Analysis of Schizophrenia Language*<br>Amal Alqahtani, Efsun Sarioglu Kayi, Sardar Hamidian, Michael Compton and Mona Diab |
| 12:00-12:30 | *Enriching Deep Learning with Frame Semantics for Empathy Classification in Medical Narrative Essays*<br>Priyanka Dey and Roxana Girju |
| 12:30 - 14:00 | ***Lunch Break*** |

| | |
|---|---|
| 14:00 - 15:30 | ***Session 3*** |
| 14:00-14:30 | *Exploring the Influence of Dialog Input Format for Unsupervised Clinical Questionnaire Filling*<br>Farnaz Ghassemi Toudeshki, Anna Liednikova, Philippe Jolivet and Claire Gardent |
| 14:30-15:00 | *DDI-MuG: Multi-aspect Graphs for Drug-Drug Interaction Extraction*<br>Jie Yang, Yihao Ding, Siqu Long, Josiah Poon and Soyeon Caren Han |
| 15:00-15:30 | *Divide and Conquer: An Extreme Multi-Label Classification Approach for Coding Diseases and Procedures in Spanish*<br>Jose Barros, Matias Rojas, Jocelyn Dunstan and Andres Abeliuk |
| 15:30 - 16:00 | ***Coffee Break*** |
| 16:00 - 17:15 | ***Session 4 (Poster Session)*** |
| 17:15 - 17:30 | ***Mini Break*** |
| 17:30 - 19:00 | ***Session 5*** |
| 17:30-18:00 | *Integration of Heterogeneous Knowledge Sources for Biomedical Text Processing*<br>Parsa Bagherzadeh and Sabine Bergler |
| 18:00-18:30 | *How Long Is Enough? Exploring the Optimal Intervals of Long-Range Clinical Note Language Modeling*<br>Samuel Cahyawijaya, Bryan Wilie, Holy Lovenia, Huan Zhong, MingQian Zhong, Yuk-Yu Nancy Ip and Pascale Fung |
| 18:30-19:00 | *Proxy-based Zero-Shot Entity Linking by Effective Candidate Retrieval*<br>Maciej Wiatrak, Eirini Arvaniti, Angus Brayne, Jonas Vetterle and Aaron Sim |

# W7 - Massively Multilingual Natural Language Understanding 2022

**Organizers:**
Jack FitzGerald, Kay Rottmann, Julia Hirschberg, Mohit Bansal, Anna Rumshisky, Charith Peris, Christopher Hench

https://mmnlu-22.github.io/
Venue: Capital Suite 10
**Wednesday, December 7, 2022**

*Let's scale natural language understanding technology to every language on Earth!*
By 2023 there will be over 8 billion virtual assistants worldwide, the majority of which will be on smartphones. Additionally, over 100 million smart speakers have been sold, most of which exclusively use a voice interface and require Natural Language Understanding (NLU) during every user interaction in order to function. However, even as we approach the point in which there will be more virtual assistants than people in the world, major virtual assistants still only support a small fraction of the world's languages. This limitation is driven by the lack of labeled data, the expense associated with human-based quality assurance, model maintenance and update costs, and more. Innovation is how we will jump these hurdles. The vision of this workshop is to help propel natural language understanding technology into the 50-language, 100-language, and even the 1,000-language regime, both for production systems and for research endeavors.

| | |
|---|---|
| 09:00 - 09:30 | ***Introduction and Shared Task Overview*** |
| 09:30 - 10:00 | ***Fine-grained Multi-lingual Disentangled Autoencoder for Language-Agnostic Representation Learning*** |
| 10:00 - 10:30 | ***Invited Talk by Mahdi Namazifar: Towards Efficient Transfer Learning Across Languages*** |
| 10:30 - 11:00 | ***Break*** |
| 11:00 - 11:30 | ***Zero-Shot Shared Task Winners: Massimo Nicosia and Francesco Piccinno, Google*** |
| 11:30 - 12:00 | ***Organizers' Choice Award: Maxime De Bruyn and the bolleke team*** |
| 12:00 - 12:30 | ***Invited Talk by Géraldine Damnati, Orange Labs: Multilingual NLP for Customer Relationship Management*** |
| 12:30 - 13:30 | ***Lunch*** |
| 13:30 - 14:00 | ***Invited Talk by Sebastian Ruder, Google: Towards Massively Multilingual Modular Models*** |
| 14:00 - 14:30 | ***Best Paper Award and Full-Data Shared Task Winner: Bo Zheng and the HIT-SCIR team*** |
| 14:30 - 15:30 | ***Poster Session*** |
| 15:30 - 16:00 | ***Break*** |
| 16:00 - 16:30 | ***Invited Talk by David Yarowsky, JHU: Massively Multilingual NLP in 1600+ Languages*** |
| 16:30 - 17:00 | ***Invited Talk by Anna Rumshisky, UMass Lowell: Learning in the Wild: Modeling Language in Real-World Scenarios*** |
| 17:00 - 17:30 | ***Invited Talk by Heng Ji, U of Illinois Urbana-Champaign: Multilingual Information Extraction for Thousands of Types*** |
| 17:30 - 18:30 | ***Networking*** |

# W8 - The Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)

**Organizers:**
David Bamman, Dirk Hovy, David Jurgens, Katherine Keith, Brendan O'Connor, Svitlana Volkova

https://sites.google.com/site/nlpandcss/
Venue: Capital Suite 2
**Wednesday, December 7, 2022**

The Natural Language Processing (NLP) and Computational Social Science (CSS) workshop brings together researchers from a variety of disciplines to discuss how to integrate NLP techniques and insights in CSS research as well as improve NLP through insights from the social sciences. This Fifth iteration of the workshop builds on a successful string of iterations with interdisciplinary contributions from many social sciences. This workshop's distinctly interdisciplinary nature aims to foster the exchange of knowledge, create new collaborations, and create opportunities for cross-discipline dialog.

| | |
|---|---|
| 09:00 - 09:10 | ***Opening Remarks*** |
| 09:10 - 10:00 | ***Invited Speaker: Anjalie Field*** |
| 10:00 - 10:30 | ***Synchronous Talk Session 1: Influence due to Linguistic Variation*** |
| 10:00-10:10 | *Lexical Choice and Projected Power: A Case Study of Implicit Gender Information in English CVs*<br>Jinrui Yang, Sheilla Njoto, Marc Cheong, Leah Ruppanner and Lea Frermann |
| 10:10-10:10 | *Conditional Language Models for Community-Level Linguistic Variation*<br>Bill Noble and Jean-philippe Bernardy |
| 10:20-10:10 | *Predicting Long-Term Citations from Short-Term Linguistic Influence*<br>Sandeep Soni, David Bamman and Jacob Eisenstein |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 11:30 | ***Synchronous Talk Session 2: Political Frames and Stances*** |
| 11:00-11:10 | *Quotatives Indicate Decline in Objectivity in U.S. Political News*<br>Tiancheng Hu, Manoel Horta Ribeiro, Robert West and Andreas Spitz |
| 11:10-11:20 | *Capturing Topic Framing via Masked Language Modeling*<br>Xiaobo Guo, Weicheng Ma and Soroush Vosoughi |
| 11:20-11:30 | *Examining Political Rhetoric with Epistemic Stance Detection*<br>Ankita Gupta, Su Lin Blodgett, Justin Gross and Brendan O'connor |
| 11:30 - 12:30 | ***In-person Poster Session*** |
| | *Improving the Generalizability of Text-Based Emotion Detection by Leveraging Transformers with Psycholinguistic Features*<br>Sourabh Zanwar, Daniel Wiechmann, Yu Qiao and Elma Kerz |
| | *Lexical Choice and Projected Power: A Case Study of Implicit Gender Information in English CVs*<br>Jinrui Yang, Sheilla Njoto, Marc Cheong, Leah Ruppanner and Lea Frermann |
| | *Quotatives Indicate Decline in Objectivity in U.S. Political News* |

Tiancheng Hu, Manoel Horta Ribeiro, Robert West and Andreas Spitz

*Measuring Harmful Representations in Scandinavian Language Models*
Samia Touileb and Debora Nozza

*Fine-Grained Extraction and Classification of Skill Requirements in German-Speaking Job Ads*
Ann-sophie Gnehm, Eva Bühlmann, Helen Buchs and Simon Clematide

*Extracting Associations of Intersectional Identities with Discourse about Institution from Nigeria*
Pavan Kantharaju and Sonja Schmer-galunder

*No Word Embedding Model Is Perfect: Evaluating the Representation Accuracy for Social Bias in the Media*
Maximilian Spliethöver, Maximilian Keiff and Henning Wachsmuth

*You Are What You Talk About: Inducing Evaluative Topics for Personality Analysis*
Josip Jukić, Iva Vukojević and Jan Snajder

*Capturing Topic Framing via Masked Language Modeling*
Xiaobo Guo, Weicheng Ma and Soroush Vosoughi

*Opening up Minds with Argumentative Dialogues*
Youmna Farag, Charlotte Brand, Jacopo Amidei, Paul Piwek, Tom Stafford, Svetlana Stoyanchev and Andreas Vlachos

*Are Neural Topic Models Broken?*
Alexander Miserlis Hoyle, Rupak Sarkar, Pranav Goel and Philip Resnik

*Logical Fallacy Detection*
Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea and Bernhard Schoelkopf

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

14:00 - 15:00 ***Invited Speaker: Jisun An***

15:00 - 15:30 ***Synchronous Talk Session 2: Heterogenous Real-World Social Data***

15:00-15:10 *Fine-Grained Extraction and Classification of Skill Requirements in German-Speaking Job Ads*
Ann-sophie Gnehm, Eva Bühlmann, Helen Buchs and Simon Clematide

15:10-15:10 *Analyzing Norm Violations in Real-Time Live-Streaming Chat*
Jihyung Moon, Dong-ho Lee, Hyundong Cho, Woojeong Jin, Chan Young Park, Minwoo Kim, Jay Pujara and Sungjoon Park

15:20-15:30 *Status Biases in Deliberation Online: Evidence from a Randomized Experiment on ChangeMyView*
Emaad Manzoor, Yohan Jo and Alan Montgomery

15:30 - 16:00 ***Coffee Break***

16:00 - 17:00 ***Invited Speaker: Elliot Ash***

17:00 - 20:00 ***Break***

20:00 - 21:00 ***Virtual Poster Session***

*Fine-Grained Extraction and Classification of Skill Requirements in German-Speaking Job Ads*
Ann-sophie Gnehm, Eva Bühlmann, Helen Buchs and Simon Clematide

*Experiencer-Specific Emotion and Appraisal Prediction*
Maximilian Wegge, Enrica Troiano, Laura Ana Maria Oberlaender and Roman Klinger

*Understanding Narratives from Demographic Survey Data: a Comparative Study with Multiple Neural Topic Models*
Xiao Xu, Gert Stulp, Antal Van Den Bosch and Anne Gauthier

*Human Counts Extraction from Text*
Mian Zhong, Shehzaad Dhuliawala and Niklas Stoehr

*To Prefer or to Choose? Generating Agency and Power Counterfactuals Jointly for Gender Bias Mitigation*
Maja Stahl, Maximilian Spliethöver and Henning Wachsmuth

*Conspiracy Narratives in the Protest Movement Against COVID-19 Restrictions in Germany. A Long-term Content Analysis of Telegram Chat Groups.*
Manuel Weigand, Maximilian Weber and Johannes Gruber

*Conditional Language Models for Community-Level Linguistic Variation*
Bill Noble and Jean-philippe Bernardy

*Understanding Interpersonal Conflict Types and their Impact on Perception Classification*
Charles Welch, Joan Plepi, Béla Neuendorf and Lucie Flek

*Examining Political Rhetoric with Epistemic Stance Detection*
Ankita Gupta, Su Lin Blodgett, Justin Gross and Brendan O'connor

*Linguistic Elements of Engaging Customer Service Discourse on Social Media*
Sonam Singh and Anthony Rios

*Measuring Harmful Representations in Scandinavian Language Models*
Samia Touileb and Debora Nozza

*Can Contextualizing User Embeddings Improve Sarcasm and Hate Speech Detection?*
Kim Breitwieser

*Lexical Choice and Projected Power: A Case Study of Implicit Gender Information in English CVs*
Jinrui Yang, Sheilla Njoto, Marc Cheong, Leah Ruppanner and Lea Frermann

*Detecting Dissonant Stance in Social Media: The Role of Topic Exposure*
Vasudha Varadarajan, Nikita Soni, Weixi Wang, Christian Luhmann, H. Andrew Schwartz and Naoya Inoue

*An Analysis of Acknowledgments in NLP Conference Proceedings*
Winston Wu

*Extracting Associations of Intersectional Identities with Discourse about Institution from Nigeria*
Pavan Kantharaju and Sonja Schmer-galunder

*OLALA: Object-Level Active Learning for Efficient Document Layout Annotation*
Zejiang Shen, Weining Li, Jian Zhao, Yaoliang Yu and Melissa Dell

*Towards Few-Shot Identification of Morality Frames using In-Context Learning*
Shamik Roy, Nishanth Sridhar Nakshatri and Dan Goldwasser

*Analyzing Norm Violations in Real-Time Live-Streaming Chat*
Jihyung Moon, Dong-ho Lee, Hyundong Cho, Woojeong Jin, Chan Young Park, Minwoo Kim, Jay Pujara and Sungjoon Park

*Utilizing Weak Supervision to Create S3D: A Sarcasm Annotated Dataset*
Jordan Painter, Helen Treharne and Diptesh Kanojia

*A Robust Bias Mitigation Procedure Based on the Stereotype Content Model*
Eddie Ungless, Amy Rafferty, Hrichika Nag and Björn Ross

*Who is GPT-3? An exploration of personality, values and demographics*
Marilù Miotto, Nicola Rossberg and Bennett Kleinberg

*No Word Embedding Model Is Perfect: Evaluating the Representation Accuracy for Social Bias in the Media*
Maximilian Spliethöver, Maximilian Keiff and Henning Wachsmuth

*You Are What You Talk About: Inducing Evaluative Topics for Personality Analysis*
Josip Jukić, Iva Vukojević and Jan Snajder

*Status Biases in Deliberation Online: Evidence from a Randomized Experiment on ChangeMyView*
Emaad Manzoor, Yohan Jo and Alan Montgomery

*Predicting Long-Term Citations from Short-Term Linguistic Influence*
Sandeep Soni, David Bamman and Jacob Eisenstein

*Capturing Topic Framing via Masked Language Modeling*
Xiaobo Guo, Weicheng Ma and Soroush Vosoughi

*MBTI Personality Prediction for Fictional Characters Using Movie Scripts*
Yisi Sang, Xiangyang Mou, Mo Yu, Dakuo Wang, Jing Li and Jeffrey Stanton

*Are Neural Topic Models Broken?*
Alexander Miserlis Hoyle, Rupak Sarkar, Pranav Goel and Philip Resnik

*Logical Fallacy Detection*
Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea and Bernhard Schoelkopf

*A Critical Reflection and Forward Perspective on Empathy and Natural Language Processing*
Allison Lahnala, Charles Welch, David Jurgens and Lucie Flek

# W9 - Second Workshop on NLP for Positive Impact

**Organizers:**
Laura Biester, Dorottya Demszky, Zhijing Jin, Mrinmaya Sachan, Joel Tetreault, Steven Wilson, Lu Xiao, Jieyu Zhao

https://sites.google.com/view/nlp4positiveimpact
Venue: Capital Suite 21A
**Wednesday, December 7, 2022**

The widespread and indispensable use of language-oriented AI systems presents new opportunities to have a positive social impact. Much existing work on NLP for social good focuses on detecting or preventing harm, such as classifying hate speech, mitigating bias, or identifying signs of depression. However, NLP research also offers the potential for positive proactive applications that can improve user and public well-being or foster constructive conversations. Nevertheless, "positive impact" remains difficult to define, and well-intentioned NLP technology can raise concerns about ethics and privacy. In addition to technical papers, this workshop also features invited keynote speakers and panelists to facilitate discussion and enhance knowledge of NLP for positive impact.

| | |
|---|---|
| 13:15 - 14:00 | ***Opening Remarks and Introduction to NLP for Social Good*** |
| 14:00 - 14:30 | ***Invited Talk by Preslav Nakov*** |
| 14:30 - 14:45 | ***Preslav Nakov Live Q&A*** |
| 14:45 - 15:00 | ***Break 1*** |
| 15:00 - 16:00 | ***Physical Poster Session*** |
| 15:00-16:00 | *Securely Capturing People's Interactions with Voice Assistants at Home: A Bespoke Tool for Ethical Data Collection*<br>Angus Addlesee |
| 15:00-16:00 | *Towards Countering Essentialism through Social Bias Reasoning*<br>Emily Allaway, Nina Taneja, Sarah-jane Leslie and Maarten Sap |
| 15:00-16:00 | *Enhancing Crisis-Related Tweet Classification with Entity-Masked Language Modeling and Multi-Task Learning*<br>Philipp Seeberger and Korbinian Riedhammer |
| 15:00-16:00 | *Misinformation Detection in the Wild: News Source Classification as a Proxy for Non-article Texts*<br>Matyas Bohacek |
| 15:00-16:00 | *Breaking through Inequality of Information Acquisition among Social Classes: A Modest Effort on Measuring "Fun"*<br>Chenghao Xiao, Baicheng Sun, Jindi Wang, Mingyue Liu and Jiayi Feng |
| 15:00-16:00 | *Hate-CLIPper: Multimodal Hateful Meme Classification based on Cross-modal Interaction of CLIP Features*<br>Gokul Karthik Kumar and Karthik Nandakumar |
| 15:00-16:00 | *Participatory Translations of Oshiwambo: Towards Culture Preservation with Language Technology*<br>Wilhelmina Nekoto, Julia Kreutzer, Jenalea Rajab, Millicent Ochieng and Jade Abbott |
| 15:00-16:00 | *Challenges and Opportunities in Information Manipulation Detection: An Examination of Wartime Russian Media*<br>Chan Young Park, Julia Mendelsohn, Anjalie Field and Yulia Tsvetkov |

| | |
|---|---|
| 15:00-16:00 | *Gender Bias in Meta-Embeddings*<br>Masahiro Kaneko, Danushka Bollegala and Naoaki Okazaki |
| 15:00-16:00 | *Fair NLP Models with Differentially Private Text Encoders*<br>Gaurav Maheshwari, Pascal Denis, Mikaela Keller and Aurélien Bellet |
| 15:00-16:00 | *Mitigating Covertly Unsafe Text within Natural Language Systems*<br>Alex Mei, Anisha Kabir, Sharon Levy, Melanie Subbiah, Emily Allaway, John Judge, Desmond Patton, Bruce Bimber, Kathleen McKeown and William Yang Wang |
| 15:00-16:00 | *Logical Fallacy Detection*<br>Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea and Bernhard Schoelkopf |
| 15:00 - 16:00 | ***Virtual Poster Session 1*** |
| 15:00-16:00 | *A Dataset of Sustainable Diet Arguments on Twitter*<br>Marcus Hansen and Daniel Hershcovich |
| 15:00-16:00 | *ClimaBench: A Benchmark Dataset For Climate Change Text Understanding in English*<br>Tanmay Laud, Daniel Spokoyny, Tom Corringham and Taylor Berg-kirkpatrick |
| 15:00-16:00 | *Modelling Persuasion through Misuse of Rhetorical Appeals*<br>Amalie Pauli, Leon Derczynski and Ira Assent |
| 15:00-16:00 | *Identifying Condescending Language: A Tale of Two Distinct Phenomena?*<br>Carla Perez Almendros and Steven Schockaert |
| 15:00-16:00 | *Beyond Model Interpretability: On the Faithfulness and Adversarial Robustness of Contrastive Textual Explanations*<br>Julia El Zini and Mariette Awad |
| 15:00-16:00 | *HARALD: Augmenting Hate Speech Data Sets with Real Data*<br>Tal Ilan and Dan Vilenchik |

| | |
|---|---|
| 16:00 - 17:00 | ***Lightning Talk Session*** |
| 17:00 - 17:30 | ***Break 2*** |
| 17:30 - 18:00 | ***Invited Talk by Mike Bailey*** |
| 18:00 - 18:30 | ***Invited Talk by Milind Tambe*** |
| 18:30 - 18:45 | ***Mike Bailey & Milind Tambe Live Q&A*** |
| 18:45 - 19:15 | ***Invited Talk by Rada Mihalcea*** |
| 19:15 - 19:45 | ***Invited Talk by Sam Bowman*** |
| 19:45 - 20:00 | ***Rada Mihalcea and Sam Bowman Live Q&A*** |
| 20:00 - 20:15 | ***Break 3*** |
| 20:15 - 21:00 | ***Panel*** |
| 21:00 - 21:20 | ***Interactive Session*** |
| 21:20 - 21:30 | ***Closing Remarks and Best Paper Awards*** |
| 21:30 - 22:30 | ***Virtual Poster Session 2*** |
| 21:00-22:00 | *A unified framework for cross-domain and cross-task learning of mental health conditions*<br>Huikai Chua, Andrew Caines and Helen Yannakoudakis |
| 21:00-22:00 | *Critical Perspectives: A Benchmark Revealing Pitfalls in PerspectiveAPI*<br>Lucas Rosenblatt, Lorena Piedras and Julia Wilkins |

21:00-22:00 *Impacts of Low Socio-economic Status on Educational Outcomes: A Narrative Based Analysis*
Motti Kelbessa, Ilyas Jamil and Labiba Jahan

21:00-22:00 *Using NLP to Support English Teaching in Rural Schools*
Luis Chiruzzo, Laura Musto, Santiago Gongora, Brian Carpenter, Juan Filevich and Aiala Rosa

21:00-22:00 *"Am I Answering My Job Interview Questions Right?": A NLP Approach to Predict Degree of Explanation in Job Interview Responses*
Raghu Verrap, Ehsanul Nirjhar, Ani Nenkova and Theodora Chaspari

21:00-22:00 *Generate Me a Bedtime Story: Leveraging Natural Language Processing for Early Vocabulary Enhancement*
Trevor Hall, Maria Valentini, Eliana Colunga and Katharina Kann

21:00-22:00 *BELA: Bot for English Language Acquisition*
Muskan Mahajan

21:00-22:00 *Transformers-Based Approach for a Sustainability Term-Based Sentiment Analysis (STBSA)*
Blaise Sandwidi and Suneer Pallitharammal Mukkolakal

21:00-22:00 *Understanding COVID-19 Vaccine Campaign on Facebook using Minimal Supervision*
Tunazzina Islam and Dan Goldwasser

21:00-22:00 *Conditional Supervised Contrastive Learning for Fair Text Classification*
Jianfeng Chi, William Shand, Yaodong Yu, Kai-Wei Chang, Han Zhao and Yuan Tian

21:00-22:00 *Handling and Presenting Harmful Text in NLP Research*
Hannah Kirk, Abeba Birhane, Bertie Vidgen and Leon Derczynski

21:00-22:00 *Don't Just Clean It, Proxy Clean It: Mitigating Bias by Proxy in Pre-Trained Models*
Swetasudha Panda, Ari Kobren, Michael Wick and Qinlan Shen

# W10 - Towards Semi-Supervised and Reinforced Task-Oriented Dialog Systems

**Organizers:**
Zhijian Ou, Junlan Feng, Juanzi Li

http://seretod.org/
Venue: Capital Suite 12A
**Wednesday, December 7, 2022**

Task-oriented dialog (TOD) systems are designed to assist users to accomplish their goals. Recently, neural generative approaches have received increasing attention. Unfortunately, building TOD systems remains as a label-intensive and time-consuming task. The process still heavily relies on manually labeled dialog data and annotated task-related knowledge base. However, unlabeled data are often easily available in many forms such as human-to-human dialogs, open-domain text corpus, and unstructured knowledge documents. Moreover, the nature of sequential decision making is not fully considered in current TOD systems. The purpose of this workshop is to invite researchers from both academia and industry to share their perspectives on building semi-supervised and reinforced TOD systems, discuss challenges and advance the field in joint effort. In parallel, a challenge is organized to foster this line of research, with a newly released, large-scale, human-human dialog dataset, called the MobileCS (Mobile Customer Service) dataset, which consists of 100K real-world dialogs. The challenge includes two tracks: Information extraction from dialog transcripts (Track 1), and Task-oriented dialog systems (Track 2).

| | |
|---|---|
| 04:50 - 05:00 | ***Opening Remarks*** |
| 05:00 - 05:40 | ***Invited Talk 1 - Dilek Hakkani-Tur*** |
| 05:40 - 06:00 | ***Break*** |
| 06:00 - 07:00 | ***Oral Session 1 (Semi-Supervised Dialogue Systems)*** |
| 06:00-06:15 | *Semi-Supervised Knowledge-Grounded Pre-training for Task-Oriented Dialog Systems*<br>Weihao Zeng, Keqing He, Zechen Wang, Dayuan Fu, Guanting Dong, Ruotong Geng, Pei Wang, Jingang Wang, Chaobo Sun, Wei Wu and Weiran Xu |
| 06:15-06:30 | *Prompt Learning for Domain Adaptation in Task-Oriented Dialogue*<br>Makesh Narsimhan Sreedhar and Christopher Parisien |
| 07:00 - 08:00 | ***Oral Session 2 (Information Extraction and Knowledge-Grounded Dialogue Systems)*** |
| 07:00-07:15 | *Information Extraction and Human-Robot Dialogue towards Real-life Tasks A Baseline Study with the MobileCS Dataset*<br>Hong Liu, Hao Peng, Zhijian Ou, Juanzi Li, Yi Huang and Junlan Feng |
| 08:00 - 09:00 | ***Lunch*** |
| 09:00 - 09:40 | ***Invited Talk 2 - Jason Williams*** |
| 09:40 - 10:40 | ***Oral Session 3 (Reinforced Dialogue Systems)*** |
| 09:40-09:55 | *A Generative User Simulator with GPT-based Architecture and Goal State Tracking for Reinforced Multi-Domain Dialog Systems*<br>Hong Liu, Yucheng Cai, Zhijian Ou, Yi Huang and Junlan Feng |
| 09:55-10:10 | *Offline-to-Online Co-Evolutional User Simulator and Dialogue System* |

| | |
|---|---|
| | Dafeng Chi, Yuzheng Zhuang, Yao Mu, Bin Wang, Jianzhu Bao, Yasheng Wang, Yuhan Dong, Xin Jiang, Qun Liu and Jianye Hao |
| 10:10-10:25 | *State-Aware Adversarial Training for Utterance-Level Dialogue Generation*<br>Yi Huang, Xiaoting Wu, Wei Hu, Junlan Feng and Chao Deng |
| 10:40 - 11:00 | ***Break*** |
| 11:00 - 11:40 | ***Invited Talk 3 - Pascale Fung*** |
| 11:40 - 12:40 | ***Oral Session 4 (Dialogue Datasets)*** |
| 11:40-11:55 | *CMCC: A Comprehensive and Large-Scale Human-Human Dataset for Dialogue Systems*<br>Yi Huang, Xiaoting Wu, Si Chen, Wei Hu, Qing Zhu, Junlan Feng, Chao Deng, Zhijian Ou and Jiangjiang Zhao |
| 12:40 - 13:40 | ***Poster Session*** |
| | *A GlobalPointer based Robust Approach for Information Extraction from Dialog Transcripts*<br>Yanbo J. Wang, Sheng Chen, Hengxing Cai, Wei Wei, Kuo Yan, Zhe Sun, Hui Qin, Yuming Li and Xiaochen Cai |
| | *A Token-pair Framework for Information Extraction from Dialog Transcripts in SereTOD Challenge*<br>Chenyue Wang, Xiangxing Kong, Mengzuo Huang, Feng Li, Jian Xing, Weidong Zhang and Wuhe Zou |
| | *Disentangling Confidence Score Distribution for Out-of-Domain Intent Detection with Energy-Based Learning*<br>Yanan Wu, Zhiyuan Zeng, Keqing He, Yutao Mou, Pei Wang, Yuanmeng Yan and Weiran Xu |
| | *Oh My Mistake!: Toward Realistic Dialogue State Tracking including Turnback Utterances*<br>Takyoung Kim, Yukyung Lee, Hoonsang Yoon, Pilsung Kang, Junseong Bang and Misuk Kim |
| | *History-Aware Hierarchical Transformer for Multi-session Open-domain Dialogue System*<br>Lizhen Cui, Chunyan Miao, Yuan You, Pengwei Wang, Zhiwei Zeng, Boyang Li, Yong Liu and Tong Zhang |
| | *Modeling Complex Dialogue Mappings via Sentence Semantic Segmentation Guided Conditional Variational Auto-Encoder*<br>Kan Li, Yitong Li, Fei Mi, Weichao Wang, Yiwei Li, Shaoxiong Feng and Bin Sun |
| | *Diving Deep into Modes of Fact Hallucinations in Dialogue Systems*<br>Rohini Srihari, Yiwei Li, Sougata Saha and Souvik Das |
| | *Keep Me Updated! Memory Management in Long-term Conversations*<br>Nako Sung, Woomyoung Park, Sang-Woo Lee, Hyeri Kim, Yuin Jeong, Sungdong Kim, Min Young Lee, Soyoung Kang, Donghyun Kwak and Sanghwan Bae |
| 13:40 - 13:50 | ***SereTOD Challenge Awards*** |
| 13:50 - 14:30 | ***Panel, Closing*** |

# W11 - The Third Workshop on Simple and Efficient Natural Language Processing

**Organizers:**
Angela Fan, Iryna Gurevych, Yufang Hou, Zornitsa Kozareva, Sasha Luccioni, Nafise Sadat Moosavi, Sujith Ravi, Gyuwan Kim, Roy Schwartz, Andreas Rücklé

https://sites.google.com/view/sustainlp2022/home
Venue: Capital Suite 6
**Wednesday, December 7, 2022**

It is our great pleasure to welcome you to the third edition of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing.
The Natural Language Processing community has, in recent years, demonstrated a notable focus on improving higher scores on standard benchmarks and taking the lead on community-wide leaderboards (e.g., GLUE, SentEval). While this aspiration has led to improvements in benchmark performance of (predominantly neural) models, it has also came at a cost, i.e., increased model complexity and the evergrowing amount of computational resources required for training and using the current state-of-the- art models. Moreover, the recent research efforts have, for the most part, failed to identify sources of empirical gains in models, often failing to empirically justify the model complexity beyond benchmark performance.
Because of these easily observable trends, we organized the SustaiNLP workshop with the goal of promoting more sustainable NLP research and practices, with two main objectives: (1) encouraging development of more efficient NLP models; and (2) providing simpler architectures and empirical justification of model complexity. For both aspects, we encouraged submissions from all topical areas of NLP.
Besides the original research papers (short and long), we encouraged cross-submissions of work that has been published at other events as well as extended abstracts of work in progress that fit the scope and aims of the workshop (only the original research papers, however, are included in these workshop proceedings).
This year, we received 20 submissions from ARR, proposing a multitude of viable resource-efficient NLP methods and spanning a wide range of NLP applications. We have selected 13 submissions for presentation at the workshop, yielding an acceptance rate of $\sim$65
Many thanks to the ARR program committee and our senior area chairs for their thorough and thoughtful reviews. We would also like to thank to our panelists and invited speakers whose discussions and talks we strongly believe will make the workshop exciting and memorable.
We are looking forward to the third edition of the SustaiNLP workshop!
SustaiNLP Organizers November 2022

09:00 - 10:30 ***Opening Remarks and Gather Town Session 1***

*Who Says Elephants Can't Run: Bringing Large Scale MoE Models into Cloud Scale Production*
Young Jin Kim, Rawn Henry, Raffy Fahim and Hany Hassan

*AfroLM: A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages*
Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi and Chris Chinenye Emezue

*Data-Efficient Auto-Regressive Document Retrieval for Fact Verification*
James Thorne

*AlphaTuning: Quantization-Aware Parameter-Efficient Adaptation of Large-Scale Pre-Trained Language Models*
Se Jung Kwon, Jeonghoon Kim, Jeongin Bae, Kang Min Yoo, Jin-Hwa Kim, Baeseong Park, Byeongwook Kim, Jung-Woo Ha, Nako Sung and Dongsoo Lee

*Towards Fair Dataset Distillation for Text Classification*
Xudong Han, Aili Shen, Yitong Li, Lea Frermann, Timothy Baldwin and Trevor Cohn

*Mask More and Mask Later: Efficient Pretraining of Masked Language Models by Disentangling the [MASK] Token*
Baohao Liao, David Thulke, Sanjika Hewavitharana, Hermann Ney and Christof Monz

*Look Ma, Only 400 Samples! Revisiting the Effectiveness of Automatic N-Gram Rule Generation for Spelling Normalization in Filipino*
Lorenzo Jaime Yu Flores and Dragomir Radev

*AutoCAD: Automatically Generate Counterfactuals for Mitigating Shortcut Learning*
Minlie Huang, Jie Zhou, Jinchao Zhang, Yeshuang Zhu and Jiaxin Wen

*Contrastive Demonstration Tuning for Pre-trained Language Models*
Huajun Chen, Chuanqi Tan, Zhenru Zhang, Siyuan Cheng, Ningyu Zhang and Xiaozhuan Liang

*Reconciliation of Pre-trained Models and Prototypical Neural Networks in Few-shot Named Entity Recognition*
Jiancheng Lv, Jie Fu, Wenqiang Lei and Youcheng Huang

*Improving the Sample Efficiency of Prompt Tuning with Domain Adaptation*
Han Yu, Boyang Li and Xu Guo

*Partitioned Gradient Matching-based Data Subset Selection for Compute-Efficient Robust ASR Training*
Ganesh Ramakrishnan, Preethi Jyothi, Rishabh Iyer, Durga Sivasubramanian and Ashish Mittal

*Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again*
Yu Su, Huan Sun, Lang Li, You Chen, Clayton Washington, Nikolas McNeal and Bernal Jimenez Gutierrez

*Summarization as Indirect Supervision for Relation Extraction*
Muhao Chen, Mingyu Derek Ma, Wenxuan Zhou, I-Hung Hsu and Keming Lu

*Bridging the Training-Inference Gap for Dense Phrase Retrieval*
William Yang Wang, Yashar Mehdad, Yizhe Zhang, Wenhan Xiong, Barlas Oguz, Jinhyuk Lee and Gyuwan Kim

*Ensemble Transformer for Efficient and Accurate Ranking Tasks: an Application to Question Answering Systems*
Alessandro Moschitti, Eric Lind, Luca Soldaini and Yoshitomo Matsubara

*Plug-and-Play VQA: Zero-shot VQA by Conjoining Large Pretrained Models with Zero Training*
Steven C.H. Hoi, Silvio Savarese, Boyang Li, Junnan Li and Anthony Meng Huat Tiong

*Train Flat, Then Compress: Sharpness-Aware Minimization Learns More Compressible Models*
Emma Strubell, Sanket Vaibhav Mehta and Clara Na

*Sparse Mixers: Combining MoE and Mixing to build a more efficient BERT*
Joshua Ainslie and James Lee-Thorp

*XDoc: Unified Pre-training for Cross-Format Document Understanding*
Furu Wei, Cha Zhang, Lei Cui, Tengchao Lv and Jingye Chen

*Scaling Laws Under the Microscope: Predicting Transformer Performance from Small Scale Experiments*
Jonathan Berant, Yair Carmon and Maor Ivgi

11:00 - 12:00 ***Oral Presentation 1***

*Quadapter: Adapter for GPT-2 Quantization*
Simyung Chang, Markus Nagel, Jaeseong You and Minseop Park

*Quadapter: Adapter for GPT-2 Quantization*
Simyung Chang, Markus Nagel, Jaeseong You and Minseop Park

*AfroLM: A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages*
Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi and Chris Chinenye Emezue

*Who Says Elephants Can't Run: Bringing Large Scale MoE Models into Cloud Scale Production*
Young Jin Kim, Rawn Henry, Raffy Fahim and Hany Hassan

*Effective Pretraining Objectives for Transformer-based Autoencoders*
Alessandro Moschitti, Matteo Gabburo and Luca Di Liello

14:30 - 15:00 ***Invited Talk (Hinrich Schutze)***

15:00 - 15:30 ***Oral Presentation 2***

*Algorithmic Diversity and Tiny Models: Comparing Binary Networks and the Fruit Fly Algorithm on Document Representation Tasks*
Tanise Ceron, Nhut Truong and Aurelie Herbelot

*Scaling Laws Under the Microscope: Predicting Transformer Performance from Small Scale Experiments*
Jonathan Berant, Yair Carmon and Maor Ivgi

16:00 - 17:30 ***Gather Town Session 2***

*Efficient Two-Stage Progressive Quantization of BERT*
Charles Le, Arash Ardakani, Amir Ardakani, Hang Zhang, Yuyan Chen, James J. Clark, Brett H. Meyer and Warren J. Gross

*KGRefiner: Knowledge Graph Refinement for Improving Accuracy of Translational Link Prediction Methods*
Mohammad Javad Saeedizade, Najmeh Torabian and Behrouz Minaei-Bidgoli

*Algorithmic Diversity and Tiny Models: Comparing Binary Networks and the Fruit Fly Algorithm on Document Representation Tasks*
Tanise Ceron, Nhut Truong and Aurelie Herbelot

*HyperMixer: An MLP-based Green AI Alternative to Transformers*
Florian Mai, Arnaud Pannatier, Fabio James Fehr, Haolin Chen, Francois Marelli, François Fleuret and James Henderson

*A Few More Examples May Be Worth Billions of Parameters*
Omer Levy, Sebastian Riedel, Patrick Lewis and Yuval Kirstain

*Few-shot initializing of Active Learner via Meta-Learning*
George Tsatsaronis, Zubair Afzal, Vikrant Yadav and Zi Long Zhu

*From Mimicking to Integrating: Knowledge Integration for Pre-Trained Language Models*
Xu Sun, Jie Zhou, Peng Li, Guangxiang Zhao, Xuancheng Ren, Yankai Lin and Lei Li

*FPT: Improving Prompt Tuning Efficiency via Progressive Training*
Qun Liu, Zhiyuan Liu, Maosong Sun, Yichun Yin, Huadong Wang, Yujia Qin and Yufei Huang

*Modeling Context With Linear Attention for Scalable Document-Level Translation*
Noah A. Smith, Nikolaos Pappas, Hao Peng and Zhaofeng Wu

*Quadapter: Adapter for GPT-2 Quantization*
Simyung Chang, Markus Nagel, Jaeseong You and Minseop Park

*Quadapter: Adapter for GPT-2 Quantization*
Simyung Chang, Markus Nagel, Jaeseong You and Minseop Park

*Towards Realistic Low-resource Relation Extraction: A Benchmark with Empirical Baseline Study*
Huajun Chen, Xi Chen, Xin Xie, Ningyu Zhang, Xiang Chen and Xin Xu

*DORE: Document Ordered Relation Extraction based on Generative Framework*
Zheng Zhang, Xipeng Qiu, Hang Yan, Yuqing Yang and Qipeng Guo

*On the Curious Case of l2 norm of Sense Embeddings*
Danushka Bollegala and Yi Zhou

*Generating Multiple-Length Summaries via Reinforcement Learning for Unsupervised Sentence Summarization*
Hwanjo Yu, Xing Xie, Chayoung Park, Xiting Wang and Dongmin Hyun

*Explore Unsupervised Structures in Pretrained Models for Relation Extraction*
Yuanbin Wu, Tao Ji and Xi Yang

*Improving Generalization of Pre-trained Language Models via Stochastic Weight Averaging*
Peng Lu, Ivan Kobyzev, Mehdi Rezagholizadeh, Ahmad Rashid, Ali Ghodsi and Phillippe Langlais

*Continuation KD: Improved Knowledge Distillation through the Lens of Continuation Optimization*
Aref Jafari, Ivan Kobyzev, Mehdi Rezagholizadeh, Pascal Poupart and Ali Ghodsi

*Effective Pretraining Objectives for Transformer-based Autoencoders*
Alessandro Moschitti, Matteo Gabburo and Luca Di Liello

19:00 - 20:00 ***Invited Talk (Song Han)***

20:00 - 20:30 ***Panel Discussion***

21:00 - 21:30 ***Invited Talk (Percy Liang)***

21:30 - 22:00 ***Invited Talk (Kurt Keutzer)***

22:00 - 22:30 ***Best Paper Awards and Closing Remarks***

# W12 - Unimodal and Multimodal Induction of Linguistic Structures

**Organizers:**
Wenjuan Han, Zilong Zheng, Zhouhan Lin, Lifeng Jin, Yikang Shen, Yoon Kim, Kewei Tu

https://induction-of-structure.github.io/emnlp2022/
Venue: Capital Suite 12B
**Wednesday, December 7, 2022**

Induction of structures (IoS) is the process of inducing structured objects (a general term of structured data rather than discrete or real values) from a set of observations. It is a branch of machine learning where the output space consists of discrete combinatorial objects (such as strings, trees, and graphs) and is unobserved or partially observed during learning. IoS in natural language processing has often been very focused on the problem of uncovering the syntactic structure (e.g., a constituent or dependency tree), semantic structure, label sequence, discourse structure etc from input text. Such structures have been found useful in downstream tasks such as relation extraction and machine translation. Apart from the wide usage in language, inducing the underlying structures from raw sensory inputs (e.g., vision) has been a long-standing challenge in the field of artificial intelligence.

| | |
|---|---|
| 09:00 - 09:10 | ***Opening Remark*** |
| 09:10 - 09:50 | ***Keynote 1*** |
| 09:50 - 10:30 | ***Keynote 2*** |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 11:40 | ***Keynote 3*** |
| 11:40 - 12:20 | ***Keynote 4*** |
| 12:20 - 14:00 | ***Lunch Break*** |
| 14:00 - 14:40 | ***Keynote 5*** |
| 14:40 - 15:20 | ***Keynote 6*** |
| 15:20 - 16:00 | ***Coffee Break*** |
| 16:00 - 17:00 | ***Poster session*** |
| 17:00 - 17:30 | ***Oral Presentation I*** |
| 17:30 - 17:45 | ***Mini Break*** |
| 17:45 - 18:50 | ***Oral Presentation II*** |
| 17:45-17:58 | *Named Entity Recognition as Structured Span Prediction*<br>Urchade Zaratiana, Nadi Tomeh, Pierre Holat and Thierry Charnois |
| 17:58-18:11 | *Global Span Selection for Named Entity Recognition*<br>Urchade Zaratiana, Niama Elkhbir, Pierre Holat, Nadi Tomeh and Thierry Charnois |
| 18:11-18:24 | *A Subspace-Based Analysis of Structured and Unstructured Representations in Image-Text Retrieval* |

| | |
|---|---|
| | Erica K. Shimomoto, Edison Marrese-Taylor, Hiroya Takamura, Ichiro Kobayashi and Yusuke Miyao |
| 18:24-18:37 | *StrAE: Autoencoding for Pre-Trained Embeddings using Explicit Structure*<br>Mattia Opper, Victor Prokhorov and Narayanaswamy Siddharth |
| 18:37-18:50 | *Probing Script Knowledge from Pre-Trained Models*<br>Zijia Jin, Xingyu Zhang, Mo Yu and Lifu Huang |
| 18:50 - 19:00 | ***Ending Remark*** |

# W13 - The Sixth Widening NLP Workshop

**Organizers:**
Shaily Bhatt, Sunipa Dev, Bonaventure Dossou, Tirthankar Ghosal, Hatem Haddad, Haley M. Lepp, Fatemehsadat Mireshghallah, Surangika Ranathunga, Xanda Schofield, Isidora Tourni, Weijia Xu

https://www.winlp.org/
Venue: Capital Suite 21B
**Wednesday, December 7, 2022**

The WiNLP workshop is open to all to foster an inclusive and welcoming ACL environment. It aims to promote diversity and highlight the work of underrepresented groups in NLP: anyone who self-identifies within an underrepresented group [based on gender, ethnicity, nationality, sexual orientation, disability status, or otherwise] is invited to submit a two-page abstract for a poster presentation. In our 2022 iteration, we hope to be more intentional about centering discussions of access and disability, as well as contributing to diversity in scientific background, discipline, training, obtained degrees, seniority, and communities from underrepresented languages.

The full-day event includes invited talks, oral presentations, and poster sessions. The workshop provides an excellent opportunity for junior members in the community to showcase their work and connect with senior mentors for feedback and career advice. It also offers recruitment opportunities with leading industrial labs. Most importantly, the workshop will provide an inclusive and accepting space, and work to lower structural barriers to joining and collaborating with the NLP community at large.

08:45 - 09:45 ***Poster Session***

*Lexical methods for bias exploration from a Latin American perspective*
Luciana Benotti, Laura Alonso Alemany and Lucia Gonzalez

*Low Resourced Multilingual Neural Machine Translation for Ometo-English*

*Improving neural machine translation for low-resource languages using related language resources*
Atnafu Lambebo Tonja

*Transformer Based Amharic Headline Generation using Sub-word2Vec Representation*
Mahlet Taye, Yaregal Assabie and Abebaw Eshetu

*DistillEmb: Distilling word embeddings via contrastive learning*
Amanuel Mersha and Stephen Wu

*Perturbation-based Active Learning for Question Answering*
Fan Luo and Mihai Surdeanu

*Short Comparative Analysis on Pretrained BART and RoBERTa in Detecting Hate Speech on YouTube and Reddit Platforms*
Dinuja Perera and Nisansa De Silva

*Synthesis and Evaluation of a Domain-specific Large Data Set for Dungeons & Dragons*
Akila Peiris and Nisansa De Silva

*Amharic Fake News Detection on Social Media Using Feature Fusion*

*MBTI Personality Prediction Approach on Persian Twitter*
Samin Fatehi, Zahra Anvarian, Yasmin Madani, Mohammadjavad Mehditabar and Sauleh Eetemadi

*Afaan Oromo Hate Speech Detection and Classification on Social Media*

*Amharic-Kistanigna Bi-directional Machine Translation using Deep Learning*
Mengistu Negia and Rahel Tamiru

*Exploiting Available Resources for the Training of Manglish Language Models*
Meisin Lee and Lay-ki Soon

*Towards a general purpose machine translation system for Sranantongo*
Just Zwennicker and David Stap

*Transfer Learning and Word Sense Disambiguation for Low-resource Language, the Case of Amharic*
Neima Ahmed and Million Meshesha

*Boosting the Performance of Gender Subspace in Domain-Specific Gender Bias Analysis*
Yanqin Tan, Cassandra L. Jacobs, Mimi Zhang, Marvin Thielk and Yi Chu

*Contextual Embeddings Can Distinguish Homonymy from Polysemy in a Human-Like Way*

*The BERT Walked Down the Garden Path Assigned Semantic Roles*

*The Effect of Normalization for Bi-directional Amharic-English Neural Machine Translation*

*Automatic Speech Recognition using Self-Supervised Learning Approach*

*Prosody Based Automatic Speech segmentation for Amharic*

*Before and beyond MeToo, Measuring changes in power and agency within sexual abuse news stories over time*
Gyulim Kang and Hope Schroeder

*Challenges of Amharic Hate Speech Data Annotation Using Yandex Toloka Crowdsourcing Platform*
Abinew Ali Ayele, Tadesse Destaw Belay, Seid Muhie Yimam, Skadi Dinter, Tesfa Asfaw and Chris Biemann

*Question Answering Classification for Amharic Social Media Community Based Questions*
Tadesse Destaw Belay, Seid Muhie Yimam, Abinew Ali Ayele and Chris Biemann

*On Genitalia, Reproduction and Pleasure: Biases in the Representation of Sexes*
Sadhi Vornberger and Peter Bourgonje

*An Unsupervised Learning Approach for Categorising Research Proposals and Recommending Papers*
Annie Lee, Mariia Ponomarenko and Peiyuan Zhou

*Detecting Depression on Twitter with a Time-Aware Multimodal Transformer*

*Detecting Adverse Drug Events from social media: A brief literature review*
Imane Guellil, Nidhaleddine Chenni, Yousra Berrachedi, Massinissa Abboud, Jinge Wu, Beatrice Alex and Honghan Wu

*Data Augmentation to Address the Out-of-Vocabulary Problem in Low-Resource Sinhala-English Neural Machine Translation*
Aloka Fernando and Surangika Ranathunga

*Evaluating Gender Bias in Pre-trained Indic Language Models*
Neeraja Kirtane, V Manushree and Aditya Kane

*Tackling Gender Microaggressions in Hindi Text*
Vishakha Agrawal

*ParsVQA-Caps: A Benchmark for Visual Question Answering and Image Captioning in Persian*
Shaghayegh Mobasher, Ghazal Zamaninejad, Maryam Hashemi, Melika Nobakhtian and Sauleh Eetemadi

*HERDPhobia: A Dataset for Hate Speech Detection against Fulani Herdsmen in Nigeria*
Saminu Aliyu, Gregory Wajiga, Muhammad Murtala, Shamsuddeen Muhammad, Idris Abdulmumin and Ibrahim Ahmad

*Experiments on Generalizability of BERTopic on Short Text*
Muriël De Groot, Mohammad Aliannejadi and Marcel Haas

*Domain-Specific Lexicon-Based Sentiment Analysis using Contextual Shifter Patterns*
Shamsuddeen Muhammad and Idris Abdulmumin

*Generate Answer to Visual Questions with Pre-trained Vision-and-Language Embeddings*
Hadi Sheikhi, Maryam Hashemi and Sauleh Eetemadi

09:45 - 10:00 ***Welcome***

10:00 - 10:50 ***Panel: Finding a cool job in NLP***

10:50 - 11:00 ***Coffee Break***

11:00 - 11:50 ***Keynote and Q&A: Prof. Houda Bouamor***

11:50 - 12:30 ***Break (Noon Prayer)***

12:30 - 13:30 ***Sponsors' Lunch***

13:30 - 14:20 ***Panel: Being a researcher in Arabic NLP***

14:20 - 14:30 ***Coffee Break***

14:30 - 15:10 ***Second Poster Session***

*Lexical methods for bias exploration from a Latin American perspective*
Luciana Benotti, Laura Alonso Alemany and Lucia Gonzalez

*Low Resourced Multilingual Neural Machine Translation for Ometo-English*

*Improving neural machine translation for low-resource languages using related language resources*
Atnafu Lambebo Tonja

*Transformer Based Amharic Headline Generation using Sub-word2Vec Representation*
Mahlet Taye, Yaregal Assabie and Abebaw Eshetu

*DistillEmb: Distilling word embeddings via contrastive learning*
Amanuel Mersha and Stephen Wu

*Perturbation-based Active Learning for Question Answering*
Fan Luo and Mihai Surdeanu

*Short Comparative Analysis on Pretrained BART and RoBERTa in Detecting Hate Speech on YouTube and Reddit Platforms*
Dinuja Perera and Nisansa De Silva

*Synthesis and Evaluation of a Domain-specific Large Data Set for Dungeons & Dragons*
Akila Peiris and Nisansa De Silva

*Amharic Fake News Detection on Social Media Using Feature Fusion*

*MBTI Personality Prediction Approach on Persian Twitter*
Samin Fatehi, Zahra Anvarian, Yasmin Madani, Mohammadjavad Mehditabar and Sauleh Eetemadi

*Afaan Oromo Hate Speech Detection and Classification on Social Media*

*Amharic-Kistanigna Bi-directional Machine Translation using Deep Learning*
Mengistu Negia and Rahel Tamiru

*Exploiting Available Resources for the Training of Manglish Language Models*
Meisin Lee and Lay-ki Soon

*Towards a general purpose machine translation system for Sranantongo*
Just Zwennicker and David Stap

*Transfer Learning and Word Sense Disambiguation for Low-resource Language, the Case of Amharic*
Neima Ahmed and Million Meshesha

*Boosting the Performance of Gender Subspace in Domain-Specific Gender Bias Analysis*
Yanqin Tan, Cassandra L. Jacobs, Mimi Zhang, Marvin Thielk and Yi Chu

*Contextual Embeddings Can Distinguish Homonymy from Polysemy in a Human-Like Way*

*The BERT Walked Down the Garden Path Assigned Semantic Roles*

*The Effect of Normalization for Bi-directional Amharic-English Neural Machine Translation*

*Automatic Speech Recognition using Self-Supervised Learning Approach*

*Prosody Based Automatic Speech segmentation for Amharic*

*Before and beyond MeToo, Measuring changes in power and agency within sexual abuse news stories over time*
Gyulim Kang and Hope Schroeder

*Challenges of Amharic Hate Speech Data Annotation Using Yandex Toloka Crowdsourcing Platform*
Abinew Ali Ayele, Tadesse Destaw Belay, Seid Muhie Yimam, Skadi Dinter, Tesfa Asfaw and Chris Biemann

*Question Answering Classification for Amharic Social Media Community Based Questions*
Tadesse Destaw Belay, Seid Muhie Yimam, Abinew Ali Ayele and Chris Biemann

*On Genitalia, Reproduction and Pleasure: Biases in the Representation of Sexes*
Sadhi Vornberger and Peter Bourgonje

*An Unsupervised Learning Approach for Categorising Research Proposals and Recommending Papers*
Annie Lee, Mariia Ponomarenko and Peiyuan Zhou

*Detecting Depression on Twitter with a Time-Aware Multimodal Transformer*

*Detecting Adverse Drug Events from social media: A brief literature review*
Imane Guellil, Nidhaleddine Chenni, Yousra Berrachedi, Massinissa Abboud, Jinge Wu, Beatrice Alex and Honghan Wu

*Data Augmentation to Address the Out-of-Vocabulary Problem in Low-Resource Sinhala-English Neural Machine Translation*
Aloka Fernando and Surangika Ranathunga

*Evaluating Gender Bias in Pre-trained Indic Language Models*
Neeraja Kirtane, V Manushree and Aditya Kane

*Tackling Gender Microaggressions in Hindi Text*
Vishakha Agrawal

*ParsVQA-Caps: A Benchmark for Visual Question Answering and Image Captioning in Persian*
Shaghayegh Mobasher, Ghazal Zamaninejad, Maryam Hashemi, Melika Nobakhtian and Sauleh Eetemadi

*HERDPhobia: A Dataset for Hate Speech Detection against Fulani Herdsmen in Nigeria*
Saminu Aliyu, Gregory Wajiga, Muhammad Murtala, Shamsuddeen Muhammad, Idris Abdulmumin and Ibrahim Ahmad

*Experiments on Generalizability of BERTopic on Short Text*

| | |
|---|---|
| | Muriël De Groot, Mohammad Aliannejadi and Marcel Haas |
| | *Domain-Specific Lexicon-Based Sentiment Analysis using Contextual Shifter Patterns*<br>Shamsuddeen Muhammad and Idris Abdulmumin |
| | *Generate Answer to Visual Questions with Pre-trained Vision-and-Language Embeddings*<br>Hadi Sheikhi, Maryam Hashemi and Sauleh Eetemadi |
| 15:10 - 15:30 | ***Break (Afternoon Prayer)*** |
| 15:30 - 16:20 | ***Fireside chat: Dr. Klaus Zechner*** |
| 16:30 - 16:50 | ***Closing Session*** |

# W14 - BlackboxNLP Analyzing and Interpreting Neural Networks for NLP

**Organizers:**
Jasmijn Bastings, Yonatan Belinkov, Yanai Elazar, Dieuwke Hupkes, Naomi Saphra, Sarah Wiegreffe

https://blackboxnlp.github.io/
Venue: Capital Suite 5
**Thursday, December 8, 2022**

Many recent performance improvements in NLP have come at the cost of understanding of the systems. How do we assess what representations and computations models learn? How do we formalize desirable properties of interpretable models, and measure the extent to which existing models achieve them? How can we build models that better encode these properties? What can new or existing tools tell us about systems' inductive biases?
The goal of this workshop is to bring together researchers focused on interpreting and explaining NLP models by taking inspiration from machine learning, psychology, linguistics, and neuroscience. We hope the workshop will serve as an interdisciplinary meetup that allows for cross-collaboration.

The topics of the workshop include, but are not limited to: Explanation methods such as saliency, attribution, free-text explanations, or explanations with structured properties; Probing methods for testing whether models have acquired or represent certain linguistic properties; Applying analysis techniques from other disciplines (e.g., neuroscience or computer vision); Examining model performance on simplified or formal languages; More interpretable model architectures; Open-source tools for analysis, visualization, or explanation; Evaluation of explanation methods; Opinion pieces about the state of explainable NLP.

08:00 - 09:00 ***Poster Session 0 - Virtual***

*A Minimal Model for Compositional Generalization on gSCAN*
Alice Hein and Klaus Diepold

*Sparse Interventions in Language Models with Differentiable Masking*
Nicola De Cao, Leon Schmid, Dieuwke Hupkes and Ivan Titov

*Where's the Learning in Representation Learning for Compositional Semantics and the Case of Thematic Fit*
Mughilan Muthupari, Samrat Halder, Asad B. Sayeed and Yuval Marton

*Sentence Ambiguity, Grammaticality and Complexity Probes*
Sunit Bhattacharya, Vilém Zouhar and Ondrej Bojar

*Post-Hoc Interpretation of Transformer Hyperparameters with Explainable Boosting Machines*
Kiron Deb, Xuan Zhang and Kevin Duh

*Revisit Systematic Generalization via Meaningful Learning*
Ning Shi, Boxin Wang, Wei Wang, Xiangyu Liu and Zhouhan Lin

*Is It Smaller Than a Tennis Ball? Language Models Play the Game of Twenty Questions*
Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann and Walter Daelemans

*Post-hoc analysis of Arabic transformer models*
Ahmed Abdelali, Nadir Durrani, Fahim Dalvi and Hassan Sajjad

*Universal Evasion Attacks on Summarization Scoring*
Wenchuan Mu and Kwan Hui Lim

*How (Un)Faithful is Attention?*

Hessam Amini and Leila Kosseim

*Are Multilingual Sentiment Models Equally Right for the Right Reasons?*
Rasmus Kær Jørgensen, Fiammetta Caccavale, Christian Igel and Anders Søgaard

*Probing for Understanding of English Verb Classes and Alternations in Large Pre-trained Language Models*
David K Yi, James V. Bruno, Jiayu Han, Peter Zukerman and Shane Steinert-Threlkeld

*Analyzing Gender Translation Errors to Identify Information Flows between the Encoder and Decoder of a NMT System*
Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier and François Yvon

*Human Ratings Do Not Reflect Downstream Utility: A Study of Free-Text Explanations for Model Predictions*
Jenny Kunz, Martin Jirenius, Oskar Holmström and Marco Kuhlmann

*Analyzing the Representational Geometry of Acoustic Word Embeddings*
Badr M. Abdullah and Dietrich Klakow

*Understanding Domain Learning in Language Models Through Subpopulation Analysis*
Zheng Zhao, Yftah Ziser and Shay B Cohen

*Intermediate Entity-based Sparse Interpretable Representation Learning*
Diego Garcia-Olano, Yasumasa Onoe, Joydeep Ghosh and Byron C Wallace

*Towards Procedural Fairness: Uncovering Biases in How a Toxic Language Classifier Uses Sentiment Information*
Isar Nejadgholi, Esma Balkir, Kathleen C. Fraser and Svetlana Kiritchenko

*Investigating the Characteristics of a Transformer in a Few-Shot Setup: Does Freezing Layers in RoBERTa Help?*
Digvijay Anil Ingle, Rishabh Kumar Tripathi, Ayush Kumar, Kevin Patel and Jithendra Vepa

*It Is Not Easy To Detect Paraphrases: Analysing Semantic Similarity With Antonyms and Negation Using the New SemAntoNeg Benchmark*
Teemu Vahtola, Mathias Creutz and Jörg Tiedemann

*Controlling for Stereotypes in Multimodal Language Model Evaluation*
Manuj Malik and Richard Johansson

*On the Compositional Generalization Gap of In-Context Learning*
Arian Hosseini, Ankit Vani, Dzmitry Bahdanau, Alessandro Sordoni and Aaron Courville

*Explaining Translationese: why are Neural Classifiers Better and what do they Learn?*
Kwabena Amponsah-Kaakyire, Daria Pylypenko, Josef Van Genabith and Cristina España-Bonet

*Probing GPT-3's Linguistic Knowledge on Semantic Tasks*
Lining Zhang, Mengchen Wang, Liben Chen and Wenxin Zhang

*Garden Path Traversal in GPT-2*
William Jurayj, William Rudman and Carsten Eickhof

*Testing Pre-trained Language Models' Understanding of Distributivity via Causal Mediation Analysis*
Pangbo Ban, Yifan Jiang, Tianran Liu and Shane Steinert-Threlkeld

*Using Roark-Hollingshead Distance to Probe BERT's Syntactic Competence*
Jingcheng Niu, Wenjie Lu, Eric Corlett and Gerald Penn

*DALLE-2 is Seeing Double: Flaws in Word-to-Concept Mapping in Text2Image Models*
Royi Rassin, Shauli Ravfogel and Yoav Goldberg

*Practical Benefits of Feature Feedback Under Distribution Shift*
Anurag Katakkar, Clay H. Yoo, Weiqin Wang, Zachary Chase Lipton and Divyansh Kaushik

*Identifying the Source of Vulnerability in Explanation Discrepancy: A Case Study in Neural Text Classification*
Ruixuan Tang, Hanjie Chen and Yangfeng Ji

*Probing Pretrained Models of Source Codes*
Sergey Troshin and Nadezhda Chirkova

*Probing the representations of named entities in Transformer-based Language Models*
Stefan Frederik Schouten, Peter Bloem and Piek Vossen

*Decomposing Natural Logic Inferences for Neural NLI*
Julia Rozanova, Deborah Ferreira, Mokanarangan Thayaparan, Marco Valentino and Andre Freitas

*Probing with Noise: Unpicking the Warp and Weft of Embeddings*
Filip Klubicka and John D. Kelleher

*Look to the Right: Mitigating Relative Position Bias in Extractive Question Answering*
Kazutoshi Shinoda, Saku Sugawara and Akiko Aizawa

*A Continuum of Generation Tasks for Investigating Length Bias and Degenerate Repetition*
Darcey Riley and David Chiang

*Universal and Independent: Multilingual Probing Framework for Exhaustive Model Interpretation and Evaluation*
Oleg Serikov, Vitaly Protasov, Ekaterina Voloshina, Viktoria Knyazkova and Tatiana Shavrina

*The Rediscovery Hypothesis: Language Models Need to Meet Linguistics*
Vassilina Nikoulina, Maxat Tezekbayev, Nuradil Kozhakhmet, Madina Babazhanova, Matthias Gallé and Zhenisbek Assylbekov

*The Solvability of Interpretability Evaluation Metrics*
Yilun Zhou and Julie Shah

*Discontinuous Constituency and BERT: A Case Study of Dutch*
Konstantinos Kogkalidis and Gijs Wijnholds

*The BERT Walked Down the Garden Path Assigned Semantic Roles*
Tovah Irwin, Kyra Wilson and Alec Marantz

*Analyzing Transformers in Embedding Space*
Guy Dar, Mor Geva, Ankit Gupta and Jonathan Berant

*FIDAM-Eval: A Framework for Evaluating Feature Interaction Detection and Attribution Methods*
Jaap Jumelet and Willem H. Zuidema

*Faithful, Interpretable Model Explanations via Causal Abstraction*
Atticus Geiger, Zhengxuan Wu, Karel D'Oosterlinck, Elisa Kreiss, Noah Goodman, Thomas Icard and Christopher Potts

*Behavioral Testing of Knowledge Graph Embedding Models for Link Prediction*
Wiem Ben Rim, Carolin Lawrence, Kiril Gashteovski, Mathias Niepert and Naoaki Okazaki

*Do LSTMs See Gender? Probing the Ability of LSTMs to Learn Abstract Syntactic Rules*
Priyanka Sukumaran, Conor Houghton and Nina Kazanina

*FRAME: Evaluating Rationale-Label Consistency Metrics for Free-Text Rationales*
Aaron Chan, Shaoliang Nie, Liang Tan, Xiaochang Peng, Hamed Firooz, Maziar Sanjabi and Xiang Ren

*Tracing and Manipulating Intermediate Results in Neural Math Problem Solvers*
Yuta Matsumoto, Benjamin Heinzerling, Masashi Yoshikawa and Kentaro Inui

*A Critical Look at Fine-tuning Generalization: Are Patterns All We Need?*

Marius Mosbach, Shauli Ravfogel, Dietrich Klakow and Yanai Elazar

*A Human-Centric Assessment Framework for AI*
Sascha Saralajew, Ammar Shaker, Zhao Xu, Kiril Gashteovski, Bhushan Kotnis, Wiem Ben Rim, Jürgen Quittek and Carolin Lawrence

*Do Language Models Understand Measurements?*
Sungjin Park, Seungwoo Ryu and Edward Choi

*Outlier Dimensions that Disrupt Transformers are Driven by Frequency*
Giovanni Puccetti, Anna Rogers, Aleksandr Drozd and Felice Dell'Orletta

*On the Impact of Temporal Concept Drift on Model Explanations*
Zhixue Zhao, George Chrysostomou, Kalina Bontcheva and Nikolaos Aletras

*Lexical Generalization Improves with Larger Models and Longer Training*
Elron Bandel, Yoav Goldberg and Yanai Elazar

*What Has Been Enhanced in my Knowledge-Enhanced Language Model?*
Yifan Hou, Guoji Fu and Mrinmaya Sachan

*SensePOLAR: Word sense aware interpretability for pre-trained contextual word embeddings*
Jan Engler, Sandipan Sikdar, Marlene Lutz and Markus Strohmaier

*Identifying Human Strategies for Generating Word-Level Adversarial Examples*
Maximilian Mozes, Bennett Kleinberg and Lewis Griffin

*Transformer Language Models without Positional Encodings Still Learn Positional Information*
Adi Haviv, Ori Ram, Ofir Press, Peter Izsak and Omer Levy

*Probing Relational Knowledge in Language Models via Word Analogies*
Kiamehr Rezaee and Jose Camacho-Collados

*Few-Shot Out-of-Domain Transfer Learning of Natural Language Explanations in a Label-Abundant Setup*
Yordan Yordanov, Vid Kocijan, Thomas Lukasiewicz and Oana-Maria Camburu

*The Curious Case of Absolute Position Embeddings*
Koustuv Sinha, Amirhossein Kazemnejad, Siva Reddy, Joelle Pineau, Dieuwke Hupkes and Adina Williams

*Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining*
Andreas Madsen, Nicholas Meade, Vaibhav Adlakha and Siva Reddy

*Are Large Pre-Trained Language Models Leaking Your Personal Information?*
Jie Huang, Hanyin Shao and Kevin Chen-Chuan Chang

*Exploring The Landscape of Distributional Robustness for Question Answering Models*
Anas Awadalla, Mitchell Wortsman, Gabriel Ilharco, Sewon Min, Ian Magnusson, Hannaneh Hajishirzi and Ludwig Schmidt

*Impact of Pretraining Term Frequencies on Few-Shot Numerical Reasoning*
Yasaman Razeghi, Robert L Logan IV, Matt Gardner and Sameer Singh

*What do Large Language Models Learn beyond Language?*
Avinash Madasu and Shashank Srivastava

*How Much Does Attention Actually Attend? Questioning the Importance of Attention in Pre-trained Transformers*
Michael Hassid, Hao Peng, Daniel Rotem, Jungo Kasai, Ivan Montero, Noah A. Smith and Roy Schwartz

*Probing for Constituency Structure in Neural Language Models*
David Arps, Younes Samih, Laura Kallmeyer and Hassan Sajjad

*Recursive Neural Networks with Bottlenecks Diagnose (Non-)Compositionality*
Verna Dankers and Ivan Titov

*CAT-probing: A Metric-based Approach to Interpret How Pre-trained Models for Programming Language Attend Code Structure*
Nuo Chen, Qiushi Sun, Renyu Zhu, Xiang Li, Xuesong Lu and Ming Gao

*ER-Test: Evaluating Explanation Regularization Methods for Language Models*
Brihi Joshi, Aaron Chan, Ziyi Liu, Shaoliang Nie, Maziar Sanjabi, Hamed Firooz and Xiang Ren

*Can Language Models Serve as Temporal Knowledge Bases?*
Ruilin Zhao, Feng Zhao, Guandong Xu, Sixiao Zhang and Hai Jin

*Baked-in State Probing*
Shubham Toshniwal, Sam Wiseman, Karen Livescu and Kevin Gimpel

*Towards Tracing Knowledge in Language Models Back to the Training Data*
Ekin Akyurek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas and Kelvin Guu

*Calibrating Trust of Multi-Hop Question Answering Systems with Decompositional Probes*
Kaige Xie, Sarah Wiegreffe and Mark Riedl

*CodeExp: Explanatory Code Document Generation*
Haotian Cui, Chenglong Wang, Junjie Huang, Jeevana Priya Inala, Todd Mytkowicz, Bo Wang, Jianfeng Gao and Nan Duan

*Influence Functions for Sequence Tagging Models*
Sarthak Jain, Varun Manjunatha, Byron Wallace and Ani Nenkova

*CORE: A Retrieve-then-Edit Framework for Counterfactual Data Generation*
Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi and Luke Zettlemoyer

09:00 - 09:10 ***Opening Remarks***

09:15 - 10:00 ***Invited Talk 1 - Lena Voita***

10:00 - 10:30 ***Oral Presentations 1 and 2***

10:00-10:30 *Human Ratings Do Not Reflect Downstream Utility: A Study of Free-Text Explanations for Model Predictions*
Jenny Kunz, Martin Jirenius, Oskar Holmström and Marco Kuhlmann

10:00-10:30 *Analyzing the Representational Geometry of Acoustic Word Embeddings*
Badr M. Abdullah and Dietrich Klakow

10:30 - 11:00 ***Coffee Break***

11:00 - 12:30 ***Poster Session 1 - Onsite***

*Where's the Learning in Representation Learning for Compositional Semantics and the Case of Thematic Fit*
Mughilan Muthupari, Samrat Halder, Asad B. Sayeed and Yuval Marton

*Sentence Ambiguity, Grammaticality and Complexity Probes*
Sunit Bhattacharya, Vilém Zouhar and Ondrej Bojar

*The Rediscovery Hypothesis: Language Models Need to Meet Linguistics*
Vassilina Nikoulina, Maxat Tezekbayev, Nuradil Kozhakhmet, Madina Babazhanova, Matthias Gallé and Zhenisbek Assylbekov

*The Solvability of Interpretability Evaluation Metrics*
Yilun Zhou and Julie Shah

*Analyzing Transformers in Embedding Space*
Guy Dar, Mor Geva, Ankit Gupta and Jonathan Berant

*FIDAM-Eval: A Framework for Evaluating Feature Interaction Detection and Attribution Methods*
Jaap Jumelet and Willem H. Zuidema

*Understanding Domain Learning in Language Models Through Subpopulation Analysis*
Zheng Zhao, Yftah Ziser and Shay B Cohen

*Intermediate Entity-based Sparse Interpretable Representation Learning*
Diego Garcia-Olano, Yasumasa Onoe, Joydeep Ghosh and Byron C Wallace

*Investigating the Characteristics of a Transformer in a Few-Shot Setup: Does Freezing Layers in RoBERTa Help?*
Digvijay Anil Ingle, Rishabh Kumar Tripathi, Ayush Kumar, Kevin Patel and Jithendra Vepa

*It Is Not Easy To Detect Paraphrases: Analysing Semantic Similarity With Antonyms and Negation Using the New SemAntoNeg Benchmark*
Teemu Vahtola, Mathias Creutz and Jörg Tiedemann

*Controlling for Stereotypes in Multimodal Language Model Evaluation*
Manuj Malik and Richard Johansson

*Behavioral Testing of Knowledge Graph Embedding Models for Link Prediction*
Wiem Ben Rim, Carolin Lawrence, Kiril Gashteovski, Mathias Niepert and Naoaki Okazaki

*Do LSTMs See Gender? Probing the Ability of LSTMs to Learn Abstract Syntactic Rules*
Priyanka Sukumaran, Conor Houghton and Nina Kazanina

*Using Roark-Hollingshead Distance to Probe BERT's Syntactic Competence*
Jingcheng Niu, Wenjie Lu, Eric Corlett and Gerald Penn

*DALLE-2 is Seeing Double: Flaws in Word-to-Concept Mapping in Text2Image Models*
Royi Rassin, Shauli Ravfogel and Yoav Goldberg

*Probing the representations of named entities in Transformer-based Language Models*
Stefan Frederik Schouten, Peter Bloem and Piek Vossen

*Decomposing Natural Logic Inferences for Neural NLI*
Julia Rozanova, Deborah Ferreira, Mokanarangan Thayaparan, Marco Valentino and Andre Freitas

*Tracing and Manipulating Intermediate Results in Neural Math Problem Solvers*
Yuta Matsumoto, Benjamin Heinzerling, Masashi Yoshikawa and Kentaro Inui

*Probing with Noise: Unpicking the Warp and Weft of Embeddings*
Filip Klubicka and John D. Kelleher

*A Critical Look at Fine-tuning Generalization: Are Patterns All We Need?*
Marius Mosbach, Shauli Ravfogel, Dietrich Klakow and Yanai Elazar

*A Human-Centric Assessment Framework for AI*
Sascha Saralajew, Ammar Shaker, Zhao Xu, Kiril Gashteovski, Bhushan Kotnis, Wiem Ben Rim, Jürgen Quittek and Carolin Lawrence

*Look to the Right: Mitigating Relative Position Bias in Extractive Question Answering*
Kazutoshi Shinoda, Saku Sugawara and Akiko Aizawa

*Do Language Models Understand Measurements?*
Sungjin Park, Seungwoo Ryu and Edward Choi

*Outlier Dimensions that Disrupt Transformers are Driven by Frequency*
Giovanni Puccetti, Anna Rogers, Aleksandr Drozd and Felice Dell'Orletta

*On the Impact of Temporal Concept Drift on Model Explanations*
Zhixue Zhao, George Chrysostomou, Kalina Bontcheva and Nikolaos Aletras

*Lexical Generalization Improves with Larger Models and Longer Training*
Elron Bandel, Yoav Goldberg and Yanai Elazar

*What Has Been Enhanced in my Knowledge-Enhanced Language Model?*
Yifan Hou, Guoji Fu and Mrinmaya Sachan

*Identifying Human Strategies for Generating Word-Level Adversarial Examples*
Maximilian Mozes, Bennett Kleinberg and Lewis Griffin

*Transformer Language Models without Positional Encodings Still Learn Positional Information*
Adi Haviv, Ori Ram, Ofir Press, Peter Izsak and Omer Levy

*Probing Relational Knowledge in Language Models via Word Analogies*
Kiamehr Rezaee and Jose Camacho-Collados

*The Curious Case of Absolute Position Embeddings*
Koustuv Sinha, Amirhossein Kazemnejad, Siva Reddy, Joelle Pineau, Dieuwke Hupkes and Adina Williams

*Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining*
Andreas Madsen, Nicholas Meade, Vaibhav Adlakha and Siva Reddy

*Exploring The Landscape of Distributional Robustness for Question Answering Models*
Anas Awadalla, Mitchell Wortsman, Gabriel Ilharco, Sewon Min, Ian Magnusson, Hannaneh Hajishirzi and Ludwig Schmidt

*Impact of Pretraining Term Frequencies on Few-Shot Numerical Reasoning*
Yasaman Razeghi, Robert L Logan IV, Matt Gardner and Sameer Singh

*How Much Does Attention Actually Attend? Questioning the Importance of Attention in Pre-trained Transformers*
Michael Hassid, Hao Peng, Daniel Rotem, Jungo Kasai, Ivan Montero, Noah A. Smith and Roy Schwartz

*Probing for Constituency Structure in Neural Language Models*
David Arps, Younes Samih, Laura Kallmeyer and Hassan Sajjad

*Calibrating Trust of Multi-Hop Question Answering Systems with Decompositional Probes*
Kaige Xie, Sarah Wiegreffe and Mark Riedl

*CORE: A Retrieve-then-Edit Framework for Counterfactual Data Generation*
Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi and Luke Zettlemoyer

*SensePOLAR: Word sense aware interpretability for pre-trained contextual word embeddings*
Jan Engler, Sandipan Sikdar, Marlene Lutz and Markus Strohmaier

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

14:00 - 14:45 ***Invited Talk 2 - Catherine Olsson***

14:45 - 15:30 ***Oral Presentations 3 and 4***

14:45-15:30 *Investigating the Characteristics of a Transformer in a Few-Shot Setup: Does Freezing Layers in RoBERTa Help?*
Digvijay Anil Ingle, Rishabh Kumar Tripathi, Ayush Kumar, Kevin Patel and Jithendra Vepa

14:45-15:30 *Probing with Noise: Unpicking the Warp and Weft of Embeddings*
Filip Klubicka and John D. Kelleher

15:30 - 16:00 ***Coffee Break***

16:00 - 17:30 ***Poster Session 2 - Virtual***

*A Minimal Model for Compositional Generalization on gSCAN*
Alice Hein and Klaus Diepold

*Sparse Interventions in Language Models with Differentiable Masking*
Nicola De Cao, Leon Schmid, Dieuwke Hupkes and Ivan Titov

*Where's the Learning in Representation Learning for Compositional Semantics and the Case of Thematic Fit*
Mughilan Muthupari, Samrat Halder, Asad B. Sayeed and Yuval Marton

*Sentence Ambiguity, Grammaticality and Complexity Probes*
Sunit Bhattacharya, Vilém Zouhar and Ondrej Bojar

*Post-Hoc Interpretation of Transformer Hyperparameters with Explainable Boosting Machines*
Kiron Deb, Xuan Zhang and Kevin Duh

*Revisit Systematic Generalization via Meaningful Learning*
Ning Shi, Boxin Wang, Wei Wang, Xiangyu Liu and Zhouhan Lin

*Is It Smaller Than a Tennis Ball? Language Models Play the Game of Twenty Questions*
Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann and Walter Daelemans

*Post-hoc analysis of Arabic transformer models*
Ahmed Abdelali, Nadir Durrani, Fahim Dalvi and Hassan Sajjad

*Universal Evasion Attacks on Summarization Scoring*
Wenchuan Mu and Kwan Hui Lim

*How (Un)Faithful is Attention?*
Hessam Amini and Leila Kosseim

*Are Multilingual Sentiment Models Equally Right for the Right Reasons?*
Rasmus Kær Jørgensen, Fiammetta Caccavale, Christian Igel and Anders Søgaard

*Probing for Understanding of English Verb Classes and Alternations in Large Pre-trained Language Models*
David K Yi, James V. Bruno, Jiayu Han, Peter Zukerman and Shane Steinert-Threlkeld

*Analyzing Gender Translation Errors to Identify Information Flows between the Encoder and Decoder of a NMT System*
Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier and François Yvon

*Human Ratings Do Not Reflect Downstream Utility: A Study of Free-Text Explanations for Model Predictions*
Jenny Kunz, Martin Jirenius, Oskar Holmström and Marco Kuhlmann

*Analyzing the Representational Geometry of Acoustic Word Embeddings*
Badr M. Abdullah and Dietrich Klakow

*Understanding Domain Learning in Language Models Through Subpopulation Analysis*
Zheng Zhao, Yftah Ziser and Shay B Cohen

*Intermediate Entity-based Sparse Interpretable Representation Learning*
Diego Garcia-Olano, Yasumasa Onoe, Joydeep Ghosh and Byron C Wallace

*Towards Procedural Fairness: Uncovering Biases in How a Toxic Language Classifier Uses Sentiment Information*
Isar Nejadgholi, Esma Balkir, Kathleen C. Fraser and Svetlana Kiritchenko

*Investigating the Characteristics of a Transformer in a Few-Shot Setup: Does Freezing Layers in RoBERTa Help?*
Digvijay Anil Ingle, Rishabh Kumar Tripathi, Ayush Kumar, Kevin Patel and Jithendra Vepa

*It Is Not Easy To Detect Paraphrases: Analysing Semantic Similarity With Antonyms and Negation Using the New SemAntoNeg Benchmark*
Teemu Vahtola, Mathias Creutz and Jörg Tiedemann

*Controlling for Stereotypes in Multimodal Language Model Evaluation*
Manuj Malik and Richard Johansson

*On the Compositional Generalization Gap of In-Context Learning*
Arian Hosseini, Ankit Vani, Dzmitry Bahdanau, Alessandro Sordoni and Aaron Courville

*Explaining Translationese: why are Neural Classifiers Better and what do they Learn?*

Kwabena Amponsah-Kaakyire, Daria Pylypenko, Josef Van Genabith and Cristina España-Bonet

*Probing GPT-3's Linguistic Knowledge on Semantic Tasks*
Lining Zhang, Mengchen Wang, Liben Chen and Wenxin Zhang

*Garden Path Traversal in GPT-2*
William Jurayj, William Rudman and Carsten Eickhof

*Testing Pre-trained Language Models' Understanding of Distributivity via Causal Mediation Analysis*
Pangbo Ban, Yifan Jiang, Tianran Liu and Shane Steinert-Threlkeld

*Using Roark-Hollingshead Distance to Probe BERT's Syntactic Competence*
Jingcheng Niu, Wenjie Lu, Eric Corlett and Gerald Penn

*DALLE-2 is Seeing Double: Flaws in Word-to-Concept Mapping in Text2Image Models*
Royi Rassin, Shauli Ravfogel and Yoav Goldberg

*Practical Benefits of Feature Feedback Under Distribution Shift*
Anurag Katakkar, Clay H. Yoo, Weiqin Wang, Zachary Chase Lipton and Divyansh Kaushik

*Identifying the Source of Vulnerability in Explanation Discrepancy: A Case Study in Neural Text Classification*
Ruixuan Tang, Hanjie Chen and Yangfeng Ji

*Probing Pretrained Models of Source Codes*
Sergey Troshin and Nadezhda Chirkova

*Probing the representations of named entities in Transformer-based Language Models*
Stefan Frederik Schouten, Peter Bloem and Piek Vossen

*Decomposing Natural Logic Inferences for Neural NLI*
Julia Rozanova, Deborah Ferreira, Mokanarangan Thayaparan, Marco Valentino and Andre Freitas

*Probing with Noise: Unpicking the Warp and Weft of Embeddings*
Filip Klubicka and John D. Kelleher

*Look to the Right: Mitigating Relative Position Bias in Extractive Question Answering*
Kazutoshi Shinoda, Saku Sugawara and Akiko Aizawa

*A Continuum of Generation Tasks for Investigating Length Bias and Degenerate Repetition*
Darcey Riley and David Chiang

*Universal and Independent: Multilingual Probing Framework for Exhaustive Model Interpretation and Evaluation*
Oleg Serikov, Vitaly Protasov, Ekaterina Voloshina, Viktoria Knyazkova and Tatiana Shavrina

*The Rediscovery Hypothesis: Language Models Need to Meet Linguistics*
Vassilina Nikoulina, Maxat Tezekbayev, Nuradil Kozhakhmet, Madina Babazhanova, Matthias Gallé and Zhenisbek Assylbekov

*The Solvability of Interpretability Evaluation Metrics*
Yilun Zhou and Julie Shah

*Discontinuous Constituency and BERT: A Case Study of Dutch*
Konstantinos Kogkalidis and Gijs Wijnholds

*The BERT Walked Down the Garden Path Assigned Semantic Roles*
Tovah Irwin, Kyra Wilson and Alec Marantz

*Analyzing Transformers in Embedding Space*
Guy Dar, Mor Geva, Ankit Gupta and Jonathan Berant

*FIDAM-Eval: A Framework for Evaluating Feature Interaction Detection and Attribution Methods*
Jaap Jumelet and Willem H. Zuidema

*Faithful, Interpretable Model Explanations via Causal Abstraction*
Atticus Geiger, Zhengxuan Wu, Karel D'Oosterlinck, Elisa Kreiss, Noah Goodman, Thomas Icard and Christopher Potts

*Behavioral Testing of Knowledge Graph Embedding Models for Link Prediction*
Wiem Ben Rim, Carolin Lawrence, Kiril Gashteovski, Mathias Niepert and Naoaki Okazaki

*Do LSTMs See Gender? Probing the Ability of LSTMs to Learn Abstract Syntactic Rules*
Priyanka Sukumaran, Conor Houghton and Nina Kazanina

*FRAME: Evaluating Rationale-Label Consistency Metrics for Free-Text Rationales*
Aaron Chan, Shaoliang Nie, Liang Tan, Xiaochang Peng, Hamed Firooz, Maziar Sanjabi and Xiang Ren

*Tracing and Manipulating Intermediate Results in Neural Math Problem Solvers*
Yuta Matsumoto, Benjamin Heinzerling, Masashi Yoshikawa and Kentaro Inui

*A Critical Look at Fine-tuning Generalization: Are Patterns All We Need?*
Marius Mosbach, Shauli Ravfogel, Dietrich Klakow and Yanai Elazar

*A Human-Centric Assessment Framework for AI*
Sascha Saralajew, Ammar Shaker, Zhao Xu, Kiril Gashteovski, Bhushan Kotnis, Wiem Ben Rim, Jürgen Quittek and Carolin Lawrence

*Do Language Models Understand Measurements?*
Sungjin Park, Seungwoo Ryu and Edward Choi

*Outlier Dimensions that Disrupt Transformers are Driven by Frequency*
Giovanni Puccetti, Anna Rogers, Aleksandr Drozd and Felice Dell'Orletta

*On the Impact of Temporal Concept Drift on Model Explanations*
Zhixue Zhao, George Chrysostomou, Kalina Bontcheva and Nikolaos Aletras

*Lexical Generalization Improves with Larger Models and Longer Training*
Elron Bandel, Yoav Goldberg and Yanai Elazar

*What Has Been Enhanced in my Knowledge-Enhanced Language Model?*
Yifan Hou, Guoji Fu and Mrinmaya Sachan

*SensePOLAR: Word sense aware interpretability for pre-trained contextual word embeddings*
Jan Engler, Sandipan Sikdar, Marlene Lutz and Markus Strohmaier

*Identifying Human Strategies for Generating Word-Level Adversarial Examples*
Maximilian Mozes, Bennett Kleinberg and Lewis Griffin

*Transformer Language Models without Positional Encodings Still Learn Positional Information*
Adi Haviv, Ori Ram, Ofir Press, Peter Izsak and Omer Levy

*Probing Relational Knowledge in Language Models via Word Analogies*
Kiamehr Rezaee and Jose Camacho-Collados

*Few-Shot Out-of-Domain Transfer Learning of Natural Language Explanations in a Label-Abundant Setup*
Yordan Yordanov, Vid Kocijan, Thomas Lukasiewicz and Oana-Maria Camburu

*The Curious Case of Absolute Position Embeddings*
Koustuv Sinha, Amirhossein Kazemnejad, Siva Reddy, Joelle Pineau, Dieuwke Hupkes and Adina Williams

*Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining*
Andreas Madsen, Nicholas Meade, Vaibhav Adlakha and Siva Reddy

*Are Large Pre-Trained Language Models Leaking Your Personal Information?*
Jie Huang, Hanyin Shao and Kevin Chen-Chuan Chang

*Exploring The Landscape of Distributional Robustness for Question Answering Models*
Anas Awadalla, Mitchell Wortsman, Gabriel Ilharco, Sewon Min, Ian Magnusson, Hannaneh Hajishirzi and Ludwig Schmidt

*Impact of Pretraining Term Frequencies on Few-Shot Numerical Reasoning*
Yasaman Razeghi, Robert L Logan IV, Matt Gardner and Sameer Singh

*What do Large Language Models Learn beyond Language?*
Avinash Madasu and Shashank Srivastava

*How Much Does Attention Actually Attend? Questioning the Importance of Attention in Pre-trained Transformers*
Michael Hassid, Hao Peng, Daniel Rotem, Jungo Kasai, Ivan Montero, Noah A. Smith and Roy Schwartz

*Probing for Constituency Structure in Neural Language Models*
David Arps, Younes Samih, Laura Kallmeyer and Hassan Sajjad

*Recursive Neural Networks with Bottlenecks Diagnose (Non-)Compositionality*
Verna Dankers and Ivan Titov

*CAT-probing: A Metric-based Approach to Interpret How Pre-trained Models for Programming Language Attend Code Structure*
Nuo Chen, Qiushi Sun, Renyu Zhu, Xiang Li, Xuesong Lu and Ming Gao

*ER-Test: Evaluating Explanation Regularization Methods for Language Models*
Brihi Joshi, Aaron Chan, Ziyi Liu, Shaoliang Nie, Maziar Sanjabi, Hamed Firooz and Xiang Ren

*Can Language Models Serve as Temporal Knowledge Bases?*
Ruilin Zhao, Feng Zhao, Guandong Xu, Sixiao Zhang and Hai Jin

*Baked-in State Probing*
Shubham Toshniwal, Sam Wiseman, Karen Livescu and Kevin Gimpel

*Towards Tracing Knowledge in Language Models Back to the Training Data*
Ekin Akyurek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas and Kelvin Guu

*Calibrating Trust of Multi-Hop Question Answering Systems with Decompositional Probes*
Kaige Xie, Sarah Wiegreffe and Mark Riedl

*CodeExp: Explanatory Code Document Generation*
Haotian Cui, Chenglong Wang, Junjie Huang, Jeevana Priya Inala, Todd Mytkowicz, Bo Wang, Jianfeng Gao and Nan Duan

*Influence Functions for Sequence Tagging Models*
Sarthak Jain, Varun Manjunatha, Byron Wallace and Ani Nenkova

*CORE: A Retrieve-then-Edit Framework for Counterfactual Data Generation*
Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi and Luke Zettlemoyer

17:30 - 17:45 ***Mini Break***

17:45 - 18:45 ***Invited Talk 3 - David Bau***

18:45 - 19:00 ***Closing Remarks***

# W15 - Data Science with Human-in-the-Loop (Language Advances)

**Organizers:**
Eduard Dragut, Yunyao Li, Lucian Popa, Slobodan Vucetic, Shashank Srivastava

https://www.dashworkshops.org/emnlp-2022
Venue: Capital Suite 10
**Thursday, December 8, 2022**

The aim of this workshop is to stimulate research on the cooperation between humans and computers within the broad area of natural language processing, including but not limited to information extraction, information retrieval and text mining, machine translation, dialog systems, question answering, language generation, summarization, model interpretability, evaluation, fairness, and ethics. We invite researchers and practitioners interested in understanding how to optimize human-computer cooperation and how to minimize human effort along an NLP pipeline in a wide range of tasks and applications.

We hope to bring together interdisciplinary researchers from academia, research labs and practice to share, exchange, learn, and develop preliminary results, new concepts, ideas, principles, and methodologies on understanding and improving human-computer interaction in natural language processing. We expect the workshop to help develop and grow a strong community of researchers who are interested in this topic and to yield future collaborations and scientific exchanges across the relevant areas of computational linguistics, natural language processing, data mining, machine learning, data and knowledge management, human-machine interaction, and intelligent user interfaces.

| | |
|---|---|
| 08:55 - 09:00 | ***Opening Remarks*** |
| 09:00 - 09:40 | ***Keynote Talk 1*** |
| 09:40 - 10:40 | ***Paper Session 1*** |
| 10:40 - 11:00 | ***Coffee Break*** |
| 11:00 - 11:40 | ***Keynote Talk 2*** |
| 11:40 - 12:30 | ***Paper Session 2*** |
| 12:30 - 14:00 | ***Lunch Break*** |
| 14:00 - 14:20 | ***Invited Talk 1*** |
| 14:20 - 14:40 | ***Invited Talk 2*** |
| 14:40 - 15:40 | ***Paper Session 3*** |
| 15:30 - 16:00 | ***Coffee Break*** |
| 16:00 - 16:40 | ***Keynote Talk 3 (Hybrid)*** |
| 16:40 - 17:30 | ***Panel Discussion*** |

# W16 - The Fourth Workshop on Financial Technology and Natural Language Processing

**Organizers:**
Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, Hsin-Hsi Chen

https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-2022-emnlp/
Venue: Capital Suite 3
**Thursday, December 8, 2022**

In FinNLP-2022, we have a keynote (Knowledge-Based News Event Analysis and Forecasting) from Dr. Oktie Hassanzadeh, a Senior Research Staff Member at IBM T.J. Watson Research Center, and an overview of recent FinNLP studies from the FinNLP organizer. The accepted papers cover various topics, including emerging trend identification, intent classification, market information prediction, sentiment analysis, digital strategy maturity assessment, and so on. Several kinds of financial documents are explored, such as the transcriptions of earnings calls, social media data, and news articles. The shared task participants share several approaches for evaluating the rationales of amateur investors. We hope the audiences of FinNLP-2022 can learn the latest tendency, and also have a comprehensive understanding of where we are now in financial opinion scoring.

| | |
|---|---|
| 09:00 - 09:10 | ***Opening Remarks*** |
| 09:10 - 09:45 | ***Keynote - Knowledge-Based News Event Analysis and Forecasting*** |
| 09:45 - 10:00 | ***Overview of Recent FinNLP Studies*** |
| 09:00 - 10:30 | ***Main Track I*** |
| 10:00-10:15 | *Contextualizing Emerging Trends in Financial News Articles*<br>Nhu Khoa Nguyen, Thierry Delahaut, Emanuela Boros, Antoine Doucet and Gaël Lejeune |
| 10:15-10:30 | *Contextualizing Emerging Trends in Financial News Articles*<br>Nhu Khoa Nguyen, Thierry Delahaut, Emanuela Boros, Antoine Doucet and Gaël Lejeune |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 12:30 | ***Main Track II*** |
| 11:00-11:15 | *TweetFinSent: A Dataset of Stock Sentiments on Twitter*<br>Yulong Pei, Amarachi Mbakwe, Akshat Gupta, Salwa Alamir, Hanxuan Lin, Xiaomo Liu and Sameena Shah |
| 11:15-11:30 | *Stock Price Volatility Prediction: A Case Study with AutoML*<br>Hilal Pataci, Yunyao Li, Yannis Katsis, Yada Zhu and Lucian Popa |
| 11:30-11:45 | *Learning Better Intent Representations for Financial Open Intent Classification*<br>Xianzhi Li, Will Aitken, Xiaodan Zhu and Stephen W. Thomas |
| 11:45-12:00 | *Exploring Robustness of Prefix Tuning in Noisy Data: A Case Study in Financial Sentiment Analysis*<br>Sudhandar Balakrishnan, Yihao Fang and Xiaodan Zhu |
| 12:00-12:15 | *A Taxonomical NLP Blueprint to Support Financial Decision Making through Information-Centred Interactions*<br>Siavash Kazemian, Cosmin Munteanu and Gerald Penn |

| | |
|---|---|
| 12:15-12:30 | *Toward Privacy-preserving Text Embedding Similarity with Homomorphic Encryption*<br>Donggyu Kim, Garam Lee and Sungwoo Oh |
| 12:30 - 14:00 | ***Lunch Break*** |
| 14:00 - 15:30 | ***Main Track III and Shared Task I*** |
| 14:00-14:15 | *DigiCall: A Benchmark for Measuring the Maturity of Digital Strategy through Company Earning Calls*<br>Hilal Pataci, Kexuan Sun and T. Ravichandran |
| 14:15-14:30 | *Disentangled Variational Topic Inference for Topic-Accurate Financial Report Generation*<br>Sixing Yan |
| 14:30-14:40 | *Overview of the FinNLP-2022 ERAI Task: Evaluating the Rationales of Amateur Investors*<br>Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura and Hsin-Hsi Chen |
| 14:40-14:50 | *PromptShots at the FinNLP-2022 ERAI Task: Pairwise Comparison and Unsupervised Ranking*<br>Peratham Wiriyathammabhum |
| 14:50-15:00 | *LIPI at the FinNLP-2022 ERAI Task: Ensembling Sentence Transformers for Assessing Maximum Possible Profit and Loss from Online Financial Posts*<br>Sohom Ghosh and Sudip Kumar Naskar |
| 15:00-15:10 | *DCU-ML at the FinNLP-2022 ERAI Task: Investigating the Transferability of Sentiment Analysis Data for Evaluating Rationales of Investors*<br>Chenyang Lyu, Tianbo Ji and Liting Zhou |
| 15:10-15:20 | *UOA at the FinNLP-2022 ERAI Task: Leveraging the Class Label Description for Financial Opinion Mining*<br>Jinan Zou, Haiyao Cao, Yanxi Liu, Lingqiao Liu, Ehsan Abbasnejad and Javen Qinfeng Shi |
| 15:20-15:30 | *aiML at the FinNLP-2022 ERAI Task: Combining Classification and Regression Tasks for Financial Opinion Mining*<br>Zhaoxuan Qin, Jinan Zou, Qiaoyang Luo, Haiyao Cao and Yang Jiao |
| 15:30 - 16:00 | ***Coffee Break*** |
| 16:00 - 16:30 | ***Shared Task II and EMNLP Findings*** |
| 16:00-16:10 | *Yet at the FinNLP-2022 ERAI Task: Modified models for evaluating the Rationales of Amateur Investors*<br>Yan Zhuang and Fuji Ren |
| 16:10-16:20 | *LDPP at the FinNLP-2022 ERAI Task: Determinantal Point Processes and Variational Autoencoders for Identifying High-Quality Opinions from a pool of Social Media Posts*<br>Paul Trust and Rosane Minghim |
| 16:20-16:30 | *Jetsons at the FinNLP-2022 ERAI Task: BERT-Chinese for mining high MPP posts*<br>Alolika Gon, Sihan Zha, Sai Krishna Rallabandi, Parag Pravin Dakle and Preethi Raghavan |
| 16:30 - 16:42 | ***EMNLP Findings - DialogueGAT: A Graph Attention Network for Financial Risk Prediction by Modeling the Dialogues in Earnings Conference Calls*** |
| 16:42 - 16:54 | ***EMNLP Findings - VarMAE: Pre-training of Variational Masked Autoencoder for Domain-adaptive Language Understanding*** |
| 16:54 - 17:06 | ***EMNLP Findings - ASDOT: Any-Shot Data-to-Text Generation with Pretrained Language Models*** |
| 17:06 - 17:18 | ***EMNLP Findings - DocFiNet: Augmenting Text and Speech Transformers with Semi-structured Document Representations for Financial Tasks*** |
| 17:18 - 17:30 | ***EMNLP Findings - Long Text and Multi-Table Summarization: Dataset and Method*** |
| 17:30 - 17:40 | ***Closing*** |

# W17 - 3rd Workshop on Figurative Language Processing

**Organizers:**
Debanjan Ghosh, Beata Beigman Klebanov, Smaranda Muresan, Anna Feldman, Soujanya Poria, Tuhin Chakrabarty

https://sites.google.com/view/figlang2022
Venue: Capital Suite 12A
**Thursday, December 8, 2022**

Processing of figurative language is a rapidly growing area in NLP, including computational modeling of metaphors, idioms, puns, irony, sarcasm, simile, and other figures. Characteristic to all areas of human activity (from poetic, ordinary, scientific, social media) and, thus, to all types of discourse, figurative language becomes an important problem for NLP systems. Its ubiquity in language has been established in a number of corpus studies and the role it plays in human reasoning has been confirmed in psychological experiments. This makes figurative language an important research area for computational and cognitive linguistics, and its automatic identification, interpretation and generation indispensable for any semantics-oriented NLP application. The proposed workshop will be the third edition of the biennial Workshop on Figurative Language Processing, whose first edition was held at NAACL 2018 and the second – at ACL 2020. The workshop builds upon a long series of related workshops that the current organizers have been involved with: "Metaphor in NLP" series (2013-2016) and "Computational Approaches to Linguistic Creativity" series (2009-2010). We expand the scope to incorporate various types of figurative language, with the aim of maintaining and nourishing a community of NLP researchers interested in this topic. The main focus will be on computational modeling of figurative language using state-of-the-art NLP techniques. However, papers on cognitive, linguistic, social, rhetorical, and applied aspects are also of interest, provided that they are presented within a computational, formal, or a quantitative framework. In addition, we propose a shared task on grounded understanding of figurative language, as described further below. The workshop will solicit both full papers and short papers for either oral or poster presentation.

| | |
|---|---|
| 08:50 - 09:00 | ***Opening Remarks*** |
| 09:00 - 10:30 | ***Research Track*** |
| 09:00-09:15 | *Ring That Bell: A Corpus and Method for Multimodal Metaphor Detection in Videos*<br>Khalid Alnajjar, Mika Hämäläinen and Shuo Zhang |
| 09:15-09:30 | *Food for Thought: How can we exploit contextual embeddings in the translation of idiomatic expressions?*<br>Lukas Santing, Ryan Jean-Luc Sijstermans, Giacomo Anerdi, Pedro Jeuris, Marijn ten Thij and Riza Batista-Navarro |
| 09:30-09:45 | *Distribution-Based Measures of Surprise for Creative Language: Experiments with Humor and Metaphor*<br>Razvan C. Bunescu and Oseremen O. Uduehi |
| 09:45-09:55 | *The Secret of Metaphor on Expressing Stronger Emotion*<br>Yucheng Li, Frank Guerin and Chenghua Lin |
| 09:55-10:05 | *Back to the Roots: Predicting the Source Domain of Metaphors using Contrastive Learning*<br>Meghdut Sengupta, Milad Alshomary and Henning Wachsmuth |
| 10:05-10:15 | *Can Yes-No Question-Answering Models be Useful for Few-Shot Metaphor Detection?* |

| | |
|---|---|
| | Lena Dankin, Kfir Bar and Nachum Dershowitz |
| 10:15-10:25 | *On the Cusp of Comprehensibility: Can Language Models Distinguish Between Metaphors and Nonsense?*<br>Bernadeta Griciūtė, Marc Tanti and Lucia Donatelli |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 12:30 | ***Research Track + Shared Tasks*** |
| 11:00-11:10 | *A Report on the Euphemisms Detection Shared Task*<br>Patrick Lee, Anna Feldman and Jing Peng |
| 11:10-11:20 | *EUREKA: EUphemism Recognition Enhanced through Knn-based methods and Augmentation*<br>Sedrick Scott Keh, Rohit Bharadwaj, Emmy Liu, Simone Tedeschi, Varun Gangal and Roberto Navigli |
| 11:20-11:30 | *Detecting Euphemisms with Literal Descriptions and Visual Imagery*<br>Ilker Kesen, Aykut Erdem, Erkut Erdem and Iacer Calixto |
| 11:30-11:40 | *A Report on the FigLang 2022 Shared Task on Understanding Figurative Language*<br>Arkadiy Saakyan, Tuhin Chakrabarty, Debanjan Ghosh and Smaranda Muresan |
| 11:40-11:50 | *Just-DREAM-about-it: Figurative Language Understanding with DREAM-FLUTE*<br>Yuling Gu, Yao Fu, Valentina Pyatkin, Ian Magnusson, Bhavana Dalvi Mishra and Peter Clark |
| 11:50-12:00 | *Effective Cross-Task Transfer Learning for Explainable Natural Language Inference with T5*<br>Irina Bigoulaeva, Rachneet Singh Sachdeva, Harish Tayyar Madabushi, Aline Villavicencio and Iryna Gurevych |
| 12:00-12:15 | *Drum Up SUPPORT: Systematic Analysis of Image-Schematic Conceptual Metaphors*<br>Lennart Wachowiak, Dagmar Gromann and Chao Xu |
| 12:15-12:30 | *Transfer Learning Parallel Metaphor using Bilingual Embeddings*<br>Maria Berger |
| 12:30 - 14:00 | ***Lunch Break*** |
| 14:00 - 15:00 | ***Keynote Talk 1: Aline Villavicencio: Modelling Multiword Expressions and Idiomaticity: an Acid Test for Understanding*** |
| 15:00 - 15:30 | ***Research Track*** |
| 15:00-15:10 | *An insulin pump? Identifying figurative links in the construction of the drug lexicon*<br>Antonio Reyes and Rafael Saldivar |
| 15:10-15:20 | *Picard understanding Darmok: A Dataset and Model for Metaphor-Rich Translation in a Constructed Language*<br>Peter A. Jansen and Jordan Boyd-Graber |
| 15:20-15:30 | *FigurativeQA: A Test Benchmark for Figurativeness Comprehension for Question Answering*<br>Geetanjali Rakshit and Jeffrey Flanigan |
| 15:30 - 16:00 | ***Coffee Break*** |
| 16:00 - 17:30 | ***Poster Session (Shared Tasks + Findings)*** |
| 16:00-17:30 | *TEDB System Description to a Shared Task on Euphemism Detection 2022*<br>Peratham Wiriyathammabhum |
| 16:00-17:30 | *A Prompt Based Approach for Euphemism Detection*<br>Abulimiti Maimaitituoheti, Yang Yong and Fan Xiaochao |
| 16:00-17:30 | *Euphemism Detection by Transformers and Relational Graph Attention Network*<br>Yuting Wang, Yiyi Liu, Ruqing Zhang, Yixing Fan and Jiafeng Guo |
| 16:00-17:30 | *Bayes at FigLang 2022 Euphemism Detection shared task: Cost-Sensitive Bayesian Fine-tuning and Venn-Abers Predictors for Robust Training under Class Skewed Distributions* |

| | |
|---|---|
| | Paul Trust, Kadusabe Provia and Kizito Omala |
| 16:00-17:30 | *An Exploration of Linguistically-Driven and Transfer Learning Methods for Euphemism Detection*<br>Devika Tiwari and Natalie Parde |
| 16:00-17:30 | *Adversarial Perturbations Augmented Language Models for Euphemism Identification*<br>Guneet Kohli, Prabsimran Kaur and Jatin Bedi |
| 16:00-17:30 | *Exploring Euphemism Detection in Few-Shot and Zero-Shot Settings*<br>Sedrick Scott Keh |
| 16:00-17:30 | *SBU Figures It Out: Models Explain Figurative Language*<br>Yash Kumar Lal and Mohaddeseh Bastan |
| 16:00-17:30 | *NLP@UIT at FigLang-EMNLP 2022: A Divide-and-Conquer System For Shared Task On Understanding Figurative Language*<br>Khoa Thi-Kim Phan, Duc-Vu Nguyen and Ngan Luu-Thuy Nguyen |
| 16:00-17:30 | *Visualizing the Obvious: A Concreteness-based Ensemble Model for Noun Property Prediction*<br>Chris Callison-Burch, Mark Yatskar, Marianna Apidianaki, Artemis Panagopoulou and Yue Yang |
| 16:00-17:30 | *Sarcasm Detection is Way Too Easy! An Empirical Comparison of Human and Machine Sarcasm Detection*<br>Walid Magdy, Silviu Oprea, Steven Wilson and Ibrahim Abu Farha |
| 16:00-17:30 | *A Unified Framework for Pun Generation with Humor Principles*<br>Nanyun Peng, Divyanshu Sheth and Yufei Tian |
| 16:00-17:30 | *It's Better to Teach Fishing than Giving a Fish: An Auto-Augmented Structure-aware Generative Model for Metaphor Detection*<br>Qianli Ma and Huawen Feng |
| 16:00-17:30 | *Systematicity in GPT-3's Interpretation of Novel English Noun Compounds*<br>Christopher Potts, Riley Carlson and Siyan Li |
| 16:00-17:30 | *PoeLM: A Meter- and Rhyme-Controllable Language Model for Unsupervised Poetry Generation*<br>Eneko Agirre, Aitor Soroa, Manex Agirrezabal, Mikel Artetxe and Aitor Ormazabal |
| 16:00-17:30 | *Scientific and Creative Analogies in Pretrained Language Models*<br>Ekaterina Shutova, Pushkar Mishra, Helen Yannakoudakis and Tamara Czinczoll |
| 16:00-17:30 | *Cards Against AI: Predicting Humor in a Fill-in-the-blank Party Game*<br>Dafna Shahaf and Dan Ofer |
| 17:30 - 17:55 | ***Break*** |
| 17:55 - 19:00 | ***Keynote Talk 2: Penny M. Pexman: Irony Acquisition: How Children Learn to Detect Sarcasm*** |

# W18 - 1st Workshop on Mathematical Natural Language Processing

**Organizers:**
Deborah Ferreira, Marco Valentino, Andre Freitas, Sean Welleck, Moritz Schubotz

https://sites.google.com/view/1st-mathnlp/home
Venue: Capital Suite 6
**Thursday, December 8, 2022**

Articulating mathematical arguments is a fundamental part of scientific reasoning and communication. Across many disciplines, expressing relations and interdependencies between quantities (usually in an equational form) is at the center of scientific argumentation. One can easily find examples of mathematical discourse across different scientific contributions and textbooks. Nevertheless, the application of contemporary models for performing inference over mathematical text still needs to be explored despite its importance. Creating methods and models that can understand mathematical text and discourse will pave the path toward developing systems capable of complex mathematical inference, leading to automated scientific discovery in fields that depend on mathematical knowledge. However, there are still technical gaps that need to be addressed, such as the availability of datasets and evaluation tasks, techniques for the joint interpretation of different modalities present in the mathematical text (equational and natural language), the understanding of unique aspects of mathematical discourse and multi-hop models for mathematical inference. This workshop will be an initial community-building venue for addressing these challenges by connecting experts in this field.

| | |
|---|---|
| 09:00 - 09:15 | ***Opening Remarks*** |
| 09:15 - 10:30 | ***Oral Session 1*** |
| 09:15-09:30 | *Tracing and Manipulating intermediate values in Neural Math Problem Solvers*<br>Yuta Matsumoto, Benjamin Heinzerling, Masashi Yoshikawa and Kentaro Inui |
| 09:30-09:45 | *Evaluating Token-Level and Passage-Level Dense Retrieval Models for Math Information Retrieval*<br>Wei Zhong, Jheng-Hong Yang, YUQING XIE and Jimmy Lin |
| 09:45-10:00 | *Investigating Math Word Problems using Pretrained Multilingual Language Models*<br>Minghuan Tan, Lei Wang, Lingxiao Jiang and Jing Jiang |
| 10:00-10:15 | *Induced Natural Language Rationales and Interleaved Markup Tokens Enable Extrapolation in Large Language Models*<br>Mirelle Candida Bueno, Carlos Gemmell, Jeff Dalton, Roberto Lotufo and Rodrigo Nogueira |
| 10:15-10:30 | *Towards Autoformalization of Mathematics and Code Correctness: Experiments with Elementary Proofs*<br>Garett Cunningham, Razvan Bunescu and David Juedes |
| 10:30 - 11:00 | ***Coffee Break 1*** |
| 11:00 - 11:45 | ***Invited Talk 1: Ashwin Kalyan: LLMs-as-a-Service: Harnessing the power of Foundation Models for Challenging Reasoning Problems*** |
| 11:45 - 12:30 | ***Oral Session 2*** |
| 11:45-12:00 | *Textual Enhanced Contrastive Learning for Solving Math Word Problems*<br>Yibin Shen, Qianying Liu, Zhuoyuan Mao, Fei Cheng and Sadao Kurohashi |

| | |
|---|---|
| 12:00-12:15 | *Multi-View Reasoning: Consistent Contrastive Learning for Math Word Problem*<br>Wenqi Zhang, Yongliang Shen, Yanna Ma, Xiaoxia Cheng, Zeqi Tan, Qingpeng Nong and Weiming Lu |
| 12:15-12:30 | *LogicSolver: Towards Interpretable Math Word Problem Solving with Logical Prompt-enhanced Learning*<br>Zhicheng Yang, Jinghui Qin, Jiaqi Chen, Liang Lin and Xiaodan Liang |
| 12:30 - 14:00 | ***Lunch*** |
| 14:00 - 14:45 | ***Building the Automated Mathematician: an NLP Perspective*** |
| 14:45 - 15:30 | ***Oral Session 3*** |
| 14:45-15:00 | *Numerical Correlation in Text*<br>Daniel Spokoyny, Chien-Sheng Wu and Caiming Xiong |
| 15:00-15:15 | *Extracting Operator Trees from Model Embeddings*<br>Anja Reusch and Wolfgang Lehner |
| 15:15-15:30 | *End-to-End Evaluation of a Spoken Dialogue System for Learning Basic Mathematics*<br>Eda Okur, Saurav Sahay, Roddy Fuentes Alba and Lama Nachman |
| 15:30 - 16:00 | ***Coffee Break 2*** |
| 16:00 - 16:15 | ***Closing Remarks*** |

# W19 - The 2nd Workshop on Multi-lingual Representation Learning

**Organizers:**
Duygu Ataman, Hila Gonen, Sebastian Ruder, Orhan Firat, Gözde Gül Sahin, Jamshidbek Mirzakhalov

https://sigtyp.github.io/ws2022-mrl.html
Venue: Capital Suite 2
**Thursday, December 8, 2022**

Multilingual representation learning methods have recently been found to be extremely efficient in learning features useful for transfer learning between languages and demonstrating potential in achieving successful adaptation of natural language processing (NLP) models into languages or tasks with little to no training resources. On the other hand, there are many aspects of such models which have the potential for further development and analysis in order to prove their applicability in various contexts. These contexts include different NLP tasks and also understudied language families, which face important obstacles in achieving practical advances that could improve the state-of-the-art in NLP of various low-resource or underrepresented languages.

The 2nd edition of the Multilingual Representation Learning (MRL) workshop continues to work for its goal of bringing together the research community consisting of scientists studying different aspects in multilingual representation learning, currently the most promising approach to improve the NLP in low-resource or underrepresented languages, and provide the rapidly growing number of researchers working on the topic with a means of communication and an opportunity to present their work and exchange ideas. This year the workshop includes an excellent line of invited speakers consisting of top experts in the field working on different aspects of multilingualism and deep learning to help foster and allow new communication channels and debates on ground-breaking ideas in computational linguistics and artificial intelligence research. Our research program presents an opportunity to present and share recent findings from all members of our community studying a wide array of multi-lingual representation learning methods, including their theoretical formulation and analysis, practical aspects such as the application of current state-of-the-art approaches in transfer learning to different tasks or studies on adaptation into previously under-studied context. Research findings are available through our proceedings, as well as poster session discussions and oral presentations containing very interesting and competitive research studies from around the world. Our workshop also attracts the interest from conference presenters, and we are excited to welcome many scientists to present their studies featured in the Findings of EMNLP. In order to provide a better understanding on how the language typology may impact the applicability of multilingual representation learning methods and motivate the development of novel methods that are more generic or competitive in different languages; we organized the first shared task on multi-lingual clause-level morphology, a novel benchmark containing sentence-level annotations in six languages with distinct morphosyntactic typology for evaluating multilingual understanding and generation models.

By allowing a communication means for research groups working on machine learning, linguistic typology, or real-life applications of NLP tasks in various languages to share and discuss their recent findings, MRL is excited for the 2022 gathering of all multilingual research community and aims to support rapid development of NLP methods and tools that are accessible to more communities and applicable to a wider range of languages.

| | |
|---|---|
| 09:00 - 09:15 | ***Opening Remarks*** |
| 09:15 - 10:00 | ***Oral Session 1*** |

| | |
|---|---|
| 10:00 - 10:30 | ***Shared task session*** |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 12:30 | ***Poster Session*** |
| 12:30 - 14:00 | ***Lunch Break*** |
| 14:00 - 14:45 | ***Invited Talk by Razvan Pascanu, Deepmind*** |
| 14:45 - 15:30 | ***Oral Session 2*** |
| 15:30 - 16:00 | ***Coffee Break*** |
| 16:00 - 16:45 | ***Invited Talk by Kyunghyun Cho, NYU*** |
| 16:45 - 17:00 | ***Mini Break*** |
| 17:00 - 17:45 | ***Invited Talk by Ev Fedorenko, MIT*** |
| 17:45 - 18:00 | ***Closing Remarks*** |

# W20 - Novel Ideas in Learning-to-Learn through Interaction

**Organizers:**
Prasanna Parthasarathi, Marc-Alexandre Côté

https://www.cs.mcgill.ca/~pparth2/nilli_workshop/
Venue: Capital Suite 21C
**Thursday, December 8, 2022**

Interactive environments have played a pivotal role in the development of reasoning mechanisms in intelligent species. Recent advances in language generation, multimodal learning, interactive and embodied learning with using language as a mode of instruction for learning have increased focus to addressing challenges in this growing topic of research. In the horizon, there is scope for realizing scenarios where agents with primitive task knowledge and an interact-and-learn procedure to systematically acquire knowledge through its interactions with the environment. This Novel Ideas in Learning-to-Learn through Interaction (NILLI) workshop focuses on collecting discussions to improve clarity towards the challenges in this topic of research which require modeling sophisticated continual interactive agents across diverse tasks and mediums of interactions. This interdisciplinary research topic unifies research paradigms of lifelong learning, natural language processing, embodied learning, reinforcement learning, robot learning and multi-modal learning towards building interactive and interpretable AI.

| | |
|---|---|
| 09:00 - 09:05 | ***Opening Remarks*** |
| 09:05 - 09:50 | ***Invited Talk 1*** |
| 09:50 - 10:30 | ***Ligtning Talks (4)*** |
| | *AERBIF: Actionable Entities Recognition Benchmark for Interactive Fiction*<br>Ivan P. Yamshchikov and Alexey Tikhonov |
| | *Multimodal Contextualized Plan Prediction for Embodied Task Completion*<br>Mert Inan, Aishwarya Padmakumar, Spandana Gella, Patrick Lange and Dilek Hakkani-Tur |
| | *Thompson sampling for interactive Bayesian optimization of dynamic masking-based language model pre-training*<br>Iñigo Urteaga, Moulay-Zaïdane Draïdia, Tomer Lancewicki and Shahram Khadivi |
| | *ReAct: Synergizing Reasoning and Acting in Language Models*<br>Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan and Yuan Cao |
| 10:30 - 10:45 | ***Coffee Break*** |
| 10:45 - 11:15 | ***Ligtning Talks (3)*** |
| | *Joint Audio/Text Training for Transformer Rescorer of Streaming Speech Recognition*<br>Suyoun Kim, Ke Li, Lucas Kabela, Rongqing Huang, Jiedan Zhu, Ozlem Kalinli and Duc Le |
| | *Revisiting the Roles of "Text" in Text Games*<br>Yi Gu, Shunyu Yao, Chuang Gan, Josh Tenenbaum and Mo Yu |
| | *ComFact: A Benchmark for Linking Contextual Commonsense Knowledge*<br>Silin Gao, Jena D. Hwang, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji and Antoine Bosselut |
| 11:15 - 12:00 | ***Invited Talk 2*** |
| 12:00 - 13:00 | ***Lunch Break*** |

13:00 - 13:45 ***Invited Talk 3***

13:45 - 14:30 ***Invited Talk 4***

14:30 - 15:40 ***Ligtning Talks (7)***

*LEMON: Language-Based Environment Manipulation via Execution-Guided Pre-training*
Qi Shi, Qian Liu, Bei Chen, Yu Zhang, Ting Liu and Jian-Guang LOU

*Learn What Is Possible, Then Choose What Is Best: Disentangling One-To-Many Relations in Language Through Text-based Games*
Benjamin Towle and Ke Zhou

*Reason first, then respond: Modular Generation for Knowledge-infused Dialogue*
Leonard Adolphs, Kurt Shuster, Jack Urbanek, Arthur Szlam and Jason Weston

*Lexi: Self-Supervised Learning of the UI Language*
Pratyay Banerjee, Shweti Mahajan, Kushal Arora, Chitta Baral and Oriana Riva

*Context-aware Information-theoretic Causal De-biasing for Interactive Sequence Labeling*
Junda Wu, Rui Wang, Tong Yu, Ruiyi Zhang, Handong Zhao, Shuai Li, Ricardo Henao and Ani Nenkova

*Predicting Long-Term Citations from Short-Term Linguistic Influence*
Sandeep Soni, David Bamman and Jacob Eisenstein

*StuBot: Learning by Teaching a Conversational Agent Through Machine Reading Comprehension*
Nayoung Jin and Hana Lee

15:40 - 16:00 ***Coffee Break***

16:00 - 16:45 ***Invited Talk 5***

16:45 - 17:30 ***Invited Talk 6***

17:30 - 17:40 ***Closing Remarks***

# W21 - Natural Legal Language Processing Workshop 2022

**Organizers:**
Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, Daniel Preoțiuc-Pietro

https://nllpw.org/
Venue: Capital Suite 8
**Thursday, December 8, 2022**

Following the success of the first three editions of the workshop (NAACL 2019, KDD 2020, EMNLP 2021), the Natural Legal Language Processing (NLLP) 2022 workshop aims to bring researchers and practitioners from NLP, machine learning and other artificial intelligence disciplines together with legal practitioners and researchers. We welcome submissions describing original work on legal data, as well as data with legal relevance, such as applications of NLP to legal tasks, experimental results using and adapting NLP methods for legal data, tasks, resources, demos, industrial research and interdisciplinary position papers.

Website: https://nllpw.org/workshop; Twitter: https://twitter.com/NllpWorkshop; Youtube: https://www.youtube.com/c/Na

Mailing list: https://groups.google.com/g/nllp

| | |
|---|---|
| 09:00 - 10:40 | ***Session 1*** |
| 09:10-09:15 | *Multi-LexSum: Real-world Summaries of Civil Rights Lawsuits at Multiple Granularities*<br>Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger and Doug Downey |
| 09:15-09:20 | *Extractive Summarization of Legal Decisions using Multi-task Learning and Maximal Marginal Relevance*<br>Abhishek Agarwal, Shanshan Xu and Matthias Grabmair |
| 09:20-09:25 | *Towards Cross-Domain Transferability of Text Generation Models for Legal Text*<br>Vinayshekhar Bannihatti Kumar, Kasturi Bhattacharjee and Rashmi Gangadharaiah |
| 09:35-09:40 | *Parameter-Efficient Legal Domain Adaptation*<br>Jonathan Li, Rohan Bhambhoria and Xiaodan Zhu |
| 09:40-09:45 | *ClassActionPrediction: A Challenging Benchmark for Legal Judgment Prediction of Class Action Cases in the US*<br>Gil Semo, Dor Bernsohn, Ben Hagag, Gila Hayat and Joel Niklaus |
| 09:45-09:50 | *Zero Shot Transfer of Legal Judgement Prediction as Article-aware Entailment for the European Court of Human Rights*<br>T.y.s.s Santosh, Oana Ichim and Matthias Grabmair |
| 09:50-09:55 | *Revisiting Transformer-based Models for Long Document Classification*<br>Xiang Dai, Ilias Chalkidis, Sune Darkner and Desmond Elliott |
| 09:55-10:00 | *Attack on Unfair ToS Clause Detection: A Case Study using Universal Adversarial Triggers*<br>Shanshan Xu, Irina Broda, Rashid Haddad, Marco Negrini and Matthias Grabmair |
| 10:00-10:05 | *AraLegal-BERT: A pretrained language model for Arabic Legal text*<br>Muhammad Al-qurishi, Sarah Alqaseemi and Riad Souissi |
| 10:05-10:10 | *An Efficient Active Learning Pipeline for Legal Text Classification*<br>Sepideh Mamooler, Rémi Lebret, Stephane Massonnet and Karl Aberer |
| 10:10-10:15 | *A Legal Approach to Hate Speech – Operationalizing the EU's Legal Framework against the Expression of Hatred as an NLP Task* |

| | |
|---|---|
| | Frederike Zufall, Marius Hamacher, Katharina Kloppenborg and Torsten Zesch |
| 10:15-10:20 | *Validity Assessment of Legal Will Statements as Natural Language Inference*<br>Alice Kwak, Jacob Israelsen, Clayton Morrison, Derek Bambauer and Mihai Surdeanu |
| 10:40 - 11:00 | ***Break*** |
| 11:00 - 12:00 | ***Finding the Law - Michael A. Livermore (University of Virginia School of Law)*** |
| 12:00 - 12:40 | ***Session 2*** |
| 12:00-12:05 | *Data-efficient end-to-end Information Extraction for Statistical Legal Analysis*<br>Wonseok Hwang, Saehee Eom, Hanuhl Lee, Hai Jin Park and Minjoon Seo |
| 12:05-12:10 | *Efficient Deep Learning-based Sentence Boundary Detection in Legal Text*<br>Reshma Sheik, Gokul T and S Nirmala |
| 12:10-12:15 | *Semantic Segmentation of Legal Documents via Rhetorical Roles*<br>Vijit Malik, Rishabh Sanjay, Shouvik Kumar Guha, Angshuman Hazarika, Shubham Nigam, Arnab Bhattacharya and Ashutosh Modi |
| 12:15-12:20 | *Detecting Relevant Differences Between Similar Legal Texts*<br>Xiang Li, Jiaxun Gao, Diana Inkpen and Wolfgang Alschner |
| 12:20-12:25 | *E-NER — An Annotated Named Entity Recognition Corpus of Legal Text*<br>Ting Wai Terence Au, Vasileios Lampos and Ingemar Cox |
| 12:40 - 14:00 | ***Lunch and In-Person Poster Session*** |
| 14:00 - 15:30 | ***Session 3*** |
| 14:00-14:05 | *On What it Means to Pay Your Fair Share: Towards Automatically Mapping Different Conceptions of Tax Justice in Legal Research Literature*<br>Reto Gubelmann, Peter Hongler, Elina Margadant and Siegfried Handschuh |
| 14:05-14:10 | *Combining WordNet and Word Embeddings in Data Augmentation for Legal Texts*<br>Sezen Perçin, Andrea Galassi, Francesca Lagioia, Federico Ruggeri, Piera Santin, Giovanni Sartor and Paolo Torroni |
| 14:10-14:15 | *Named Entity Recognition in Indian court judgments*<br>Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn and Vivek Raghavan |
| 14:15-14:20 | *Legal Named Entity Recognition with Multi-Task Domain Adaptation*<br>Răzvan-alexandru Smădu, Ion-robert Dinică, Andrei-marius Avram, Dumitru-clementin Cercel, Florin Pop and Mihaela-claudia Cercel |
| 14:35-14:40 | *Do Charge Prediction Models Learn Legal Theory?*<br>Zhenwei An, Quzhe Huang, Cong Jiang, Yansong Feng and Dongyan Zhao |
| 14:40-14:45 | *Legal-Tech Open Diaries: Lesson learned on how to develop and deploy light-weight models in the era of humongous Language Models*<br>Stelios Maroudas, Sotiris Legkas, Prodromos Malakasiotis and Ilias Chalkidis |
| 14:45-14:50 | *Processing Long Legal Documents with Pre-trained Transformers: Modding LegalBERT and Longformer*<br>Dimitris Mamakas, Petros Tsotsi, Ion Androutsopoulos and Ilias Chalkidis |
| 14:50-14:55 | *Privacy-Preserving Models for Legal Natural Language Processing*<br>Ying Yin and Ivan Habernal |
| 14:55-15:00 | *Automatic Classification of Legal Violations in Cookie Banner Texts*<br>Marieke Van Hofslot, Almila Akdag Salah, Albert Gatt and Cristiana Santos |
| 15:00-15:05 | *Tracking Semantic Shifts in German Court Decisions with Diachronic Word Embeddings*<br>Daniel Braun |
| 15:30 - 16:00 | ***Break*** |
| 16:00 - 17:30 | ***Session 4*** |

16:00-16:05 *Should I disclose my dataset? Caveats between reproducibility and individual data rights*
Raysa M. Benatti, Camila M. L. Villarroel, Sandra Avila, Esther L. Colombini and Fabiana Severi

16:05-16:10 *Privacy Pitfalls of Online Service Terms and Conditions: a Hybrid Approach for Classification and Summarization*
Emilia Lukose, Suparna De and Jon Johnson

16:10-16:15 *Computing and Exploiting Document Structure to Improve Unsupervised Extractive Summarization of Legal Case Decisions*
Yang Zhong and Diane Litman

16:15-16:20 *Abstractive Summarization of Dutch Court Verdicts Using Sequence-to-sequence Models*
Marijn Schraagen, Floris Bex, Nick Van De Luijtgaarden and Daniël Prijs

16:20-16:25 *Graph-based Keyword Planning for Legal Clause Generation from Topics*
Sagar Joshi, Sumanth Balaji, Aparna Garimella and Vasudeva Varma

16:25-16:30 *Text Simplification for Legal Domain: {I}nsights and Challenges*
Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya and Nandakishore Kambhatla

16:45-16:50 *On Breadth Alone: Improving the Precision of Terminology Extraction Systems on Patent Corpora*
Sean Nordquist and Adam Meyers

16:50-16:55 *The Legal Argument Reasoning Task in Civil Procedure*
Leonard Bongard, Lena Held and Ivan Habernal

16:55-17:00 *Legal and Political Stance Detection of SCOTUS Language*
Noah Bergam, Emily Allaway and Kathleen Mckeown

17:00-17:05 *Can AMR Assist Legal and Logical Reasoning?*
Nikolaus Schrack, Ruixiang Cui, Hugo López and Daniel Hershcovich

17:05-17:10 *LawngNLI: A Long-Premise Benchmark for In-Domain Generalization from Short to Long Contexts and for Implication-Based Retrieval*
William Bruno and Dan Roth

17:10-17:15 *Legal Prompting: Teaching a Language Model to Think Like a Lawyer*
Fangyi Yu, Lee Quartey and Frank Schilder

17:30 - 18:30 ***From NLLP to Legal NLP - The Future of the Field (Panel) & Best Presentation Award***

# W22 - Sharing Stories and Lessons Learned

**Organizers:**
Diyi Yang, Pradeep Dasigi, Tongshuang Wu, Tuhin Chakrabarty, Yuval Pinter, Mike Zheng Shou

https://ssll-emnlp.github.io/
Venue: Capital Suite 12B
**Thursday, December 8, 2022**

The driving forces of progress in NLP are the people behind the work. We learn from their work. But to generate such good work, what are the principles and strategies they used? What are the roadblocks, challenges, mistakes, and lessons learned? These are quite valuable to the newbies across different career stages. In fact, we always reach out to the senior people around us for advice and reflect on their stories when we start a new career chapter - (1) fresh phd students reach out to early career researchers including senior phds or recent graduate, (2) early career researchers reach out to mid/late career professors, (3) company newbies reach out to industrial leaders. But often, only a few people would be approachable around us. This workshop aims at making the sharing of successful researchers' stories and lessons learned to be accessible to everyone in our community. Such sharing would be very inspiring and helpful for those who might be struggling with making a choice or feeling lost right now. Our workshop will line up with sessions dedicated to individual career stage groups; each session will consist of 3-5 speeches and a panel QA & discussion to interact with the audience.

| | |
|---|---|
| 08:50 - 09:00 | ***Opening Remarks*** |
| 09:00 - 09:20 | ***Career and Research Opportunities in Product: My experience working on the Google Assistant. Speaker: Manaal Faruqui*** |
| 09:20 - 09:40 | ***Story behind the First NLP System Named After a Children's TV Star to Win a Best Paper. Speaker: Matthew Peters*** |
| 09:40 - 10:00 | ***Building Datasets for the Analysis of Culture. Speaker: David Bamman*** |
| 10:00 - 10:30 | ***Coffee Break*** |
| 10:30 - 11:00 | ***Panel on Stories behind Cool Work. Moderator: Tongshuang Wu*** |
| 11:00 - 11:20 | ***My Journey in NLP. Speaker: Zhijing Jin*** |
| 11:20 - 11:40 | ***Serendipity tales of a Mexican NLP gal. Speaker: Thamar Solorio*** |
| 11:40 - 12:00 | ***Improving Papers through Self-editing. Speaker: Bonnie Webber*** |
| 12:00 - 12:30 | ***Panel on Growth and Life as a Researcher. Moderator: Tuhin Chakrabarty*** |
| 12:30 - 14:30 | ***Networking / Mentoring*** |
| 15:30 - 15:50 | ***Weird Things About Professorship. Speaker: Colin Raffel*** |
| 15:50 - 16:10 | ***Lessons Learned from Interdisciplinary Collaboration. Speaker: Emma Strubell*** |
| 16:10 - 16:30 | ***Venturing Off the NLP Map into Interdisciplinary Lands. Speaker: David Jurgens*** |
| 16:30 - 17:00 | ***Panel on Going beyond the Comfort Zone. Moderator: Yuval Pinter*** |
| 16:00 - 16:30 | ***Closing Remarks*** |

# W23 - Workshop on Text Simplification, Accessibility, and Readability

**Organizers:**
Sanja Štajner, Horacio Saggion, Daniel Ferrés, Matthew Shardlow, Kim Cheng Sheang, Kai North, Marcos Zampieri, Wei Xu

https://www.taln.upf.edu/pages/tsar2022-ws/
Venue: Capital Suite 21A
**Thursday, December 8, 2022**

The Web provides an abundance of knowledge and information that reaches the global population. However, the way in which a text is written (vocabulary, syntax, or text organization/structure), or presented, can make it inaccessible, especially for non-native speakers, people with low literacy, and those with cognitive or linguistic impairments. Research on automatic text simplification (TS), textual accessibility, and readability thus have the potential to improve social inclusion of marginalized populations. These related research areas have increasingly attracted more and more attention in the past ten years, evidenced by the growing number of publications in NLP conferences. The Text Simplification, Accessibility, and Readability (TSAR) workshop brings together researchers, industry developers, public organizations representatives, and other parties interested in the problem of making information more accessible to all citizens. The workshop features two invited talks, round table discussion, presentation of novel research, and the results and participating systems of the shared task on lexical simplification for English, Spanish, and Portuguese, which was organized as a part of this workshop.

| | |
|---|---|
| 09:30 - 09:45 | ***Opening Remarks*** |
| 09:45 - 10:30 | ***Session 1*** |
| 09:45-10:00 | *Parallel Corpus Filtering for Japanese Text Simplification*<br>Koki Hatagaki, Tomoyuki Kajiwara and Takashi Ninomiya |
| 10:00-10:15 | *Patient-friendly Clinical Notes: Towards a new Text Simplification Dataset*<br>Jan Trienes, Jörg Schlötterer, Hans-Ulrich Schildhaus and Christin Seifert |
| 10:15-18:10 | *IrekiaLF_es: a New Open Benchmark and Baseline Systems for Spanish Automatic Text Simplification*<br>Itziar Gonzalez-Dios, Iker Gutiérrez-Fandiño, Oscar m. Cumbicus-Pineda and Aitor Soroa |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 12:30 | ***Session 2*** |
| 11:00-11:15 | *Lexically Constrained Decoding with Edit Operation Prediction for Controllable Text Simplification*<br>Tatsuya Zetsu, Tomoyuki Kajiwara and Yuki Arase |
| 11:15-11:30 | *(Psycho-)Linguistic Features Meet Transformer Models for Improved Explainable and Controllable Text Simplification*<br>Yu Qiao, Xiaofei Li, Daniel Wiechmann and Elma Kerz |
| 11:30-11:45 | *A Dataset of Word-Complexity Judgements from Deaf and Hard-of-Hearing Adults for Text Simplification*<br>Oliver Alonzo, Sooyeon Lee, Mounica Maddela, Wei Xu and Matt Huenerfauth |

| | |
|---|---|
| 11:45-12:00 | *Eye-tracking based classification of Mandarin Chinese readers with and without dyslexia using neural sequence models*<br>Patrick Haller, Andreas Säuberli, Sarah Kiener, Jinger Pan, Ming Yan and Lena Jäger |
| 12:00-12:15 | *Findings of the TSAR-2022 Shared Task on Multilingual Lexical Simplification*<br>Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North and Marcos Zampieri |
| 12:15-12:30 | *UniHD at TSAR-2022 Shared Task: Is Compute All We Need for Lexical Simplification?*<br>Dennis Aumiller and Michael Gertz |
| 12:30 - 14:00 | ***Lunch Break*** |
| 14:00 - 15:30 | ***Session 3 (Posters)*** |
| 15:30 - 16:00 | ***Coffee Break*** |
| 16:00 - 16:30 | ***Round Table Discussion*** |
| 16:30 - 17:30 | ***Invited Talk 1: Matt Huenerfauth*** |
| 17:45 - 18:45 | ***Invited Talk 2: Sowmya Vajjala*** |
| 18:45 - 19:00 | ***Closing Statements*** |

# W24 - The Seventh Arabic Natural Language Processing Workshop

**Organizers:**
Houda Bouamor, Hend Al-Khalifa, Kareem Darwish, Owen Rambow, Fethi Bougares, Ahmed Abdelali, Nadi Tomeh, Salam Khalifa, Wajdi Zaghouani

http://wanlp2022.arabic-nlp.net/
Venue: Capital Suite 21B
**Thursday, December 8, 2022**

WANLP is the premiere workshop for the Special Interest Group of Arabic NLP (SIGARAB). In this workshop we focus on Arabic, a collection of languages and language varieties that pose challenges for the field of computational linguistics. The challenges are due to many factors, including Arabic's rich morphology and widely varying, largely understudied dialects. As one of the official languages of the United Nations and the native tongue of over 400 million native speakers living in a region of geopolitical significance, Arabic is an attractive object of study from a computational perspective. The fast growth of Arabic on the Internet over the past few years and the vibrant use of Arabic dialects on social media, including issues of global relevance, lend the workshop particular and timely significance. The opportunities that are made possible by working on this language and its dialects cannot be underestimated in their consequence on the Arab World, the Mediterranean Region, and the rest of the World, let alone on the advancement of computational linguistics methods and techniques to address significant inherent linguistic complexities. This workshop aims to provide a forum for researchers in the field of Arabic NLP to share and discuss their ongoing work.

| | |
|---|---|
| 09:00 - 09:05 | ***Opening Remarks*** |
| 09:10 - 10:00 | ***Invited Talk*** |
| 10:00 - 10:30 | ***Session 1 - Information Extraction (in-Person)*** |
| 10:00-10:15 | *CAraNER: The COVID-19 Arabic Named Entity Corpus*<br>Abdulmohsen Al-Thubaity, Sakhar Alkhereyf, Wejdan Alzahrani and Alia Bahanshal |
| 10:15-10:30 | *Joint Coreference Resolution for Zeros and non-Zeros in Arabic*<br>Abdulrahman Aloraini, Sameer Pradhan and Massimo Poesio |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 12:30 | ***Session 2 - NLU/NLG (in-Person)*** |
| 11:00-11:15 | *SAIDS: A Novel Approach for Sentiment Analysis Informed of Dialect and Sarcasm*<br>Abdelrahman Kaseb and Mona Farouk |
| 11:15-11:30 | *AraBART: a Pretrained Arabic Sequence-to-Sequence Model for Abstractive Summarization*<br>Moussa Kamal Eddine, Nadi Tomeh, Nizar Habash, Joseph Le Roux and Michalis Vazirgiannis |
| 11:30-11:45 | *Towards Arabic Sentence Simplification via Classification and Generative Approaches*<br>Nouran Khallaf, Serge Sharoff and Rasha Soliman |
| 11:45-12:00 | *Generating Classical Arabic Poetry using Pre-trained Models*<br>Nehal Elkaref, Mervat Abu-Elkheir, Maryam ElOraby and Mohamed Abdelgaber |
| 12:00-12:15 | *A Benchmark Study of Contrastive Learning for Arabic Social Meaning*<br>Md Tawkat Islam Khondaker, El Moatez Billah Nagoudi, AbdelRahim Elmadany, Muhammad Abdul-Mageed and Laks Lakshmanan, V.S. |

12:15-12:30 *Adversarial Text-to-Speech for low-resource languages*
Ashraf Elneima and Mikołaj Bińkowski

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

14:00 - 14:45 ***Panel discussion: Young Researchers in Arabic NLP***

14:45 - 15:00 ***Best Paper Award Oral Presentation (in-Person)***

15:00 - 15:30 ***Shared Task Papers (in-Person)***

15:00-15:10 *NADI 2022: The Third Nuanced Arabic Dialect Identification Shared Task*
Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor and Nizar Habash

15:10-15:20 *The Shared Task on Gender Rewriting*
Bashar Alhafni, Nizar Habash, Houda Bouamor, Ossama Obeid, Sultan Alrowili, Daliyah AlZeer, Kawla Mohmad Shnqiti, Ahmed Elbakry, Muhammad ElNokrashy, Mohamed Gabr, Abderrahmane Issam, Abdelrahim Qaddoumi, Vijay Shanker and Mahmoud Zyate

15:20-15:30 *Overview of the WANLP 2022 Shared Task on Propaganda Detection in Arabic*
Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino and Preslav Nakov

15:30 - 16:00 ***Coffee Break***

16:00 - 17:00 ***Session 3 - Arabic Dialects (in-Person)***

16:00-16:15 *ArzEn-ST: A Three-way Speech Translation Corpus for Code-Switched Egyptian Arabic-English*
Injy Hamed, Nizar Habash, Slim Abdennadher and Ngoc Thang Vu

16:15-16:30 *Maknuune: A Large Open Palestinian Arabic Lexicon*
Shahd Salah Uddin Dibas, Christian Khairallah, Nizar Habash, Omar Fayez Sadi, Tariq Sairafy, Karmel Sarabta and Abrar Ardah

16:30-16:45 *Developing a Tag-Set and Extracting the Morphological Lexicons to Build a Morphological Analyzer for Egyptian Arabic*
Amany Fashwan and Sameh Alansary

16:45-16:50 *A Weak Supervised Transfer Learning Approach for Sentiment Analysis to the Kuwaiti Dialect*
Fatemah Husain, Hana Al-Ostad and Halima Omar

17:00 - 18:15 ***Main Workshop Posters (in-Person & virtual)***

*Mawqif: A Multi-label Arabic Dataset for Target-specific Stance Detection*
Nora Saleh Alturayeif, Hamzah Abdullah Luqman and Moataz Aly Kamaleldin Ahmed

*Assessing the Linguistic Knowledge in Arabic Pre-trained Language Models Using Minimal Pairs*
Wafa Abdullah Alrajhi, Hend Al-Khalifa and Abdulmalik AlSalman

*Identifying Code-switching in Arabizi*
Safaa Shehadi and Shuly Wintner

*Authorship Verification for Arabic Short Texts Using Arabic Knowledge-Base Model (AraKB)*
Fatimah Alqahtani and Helen Yannakoudakis

*A Semi-supervised Approach for a Better Translation of Sentiment in Dialectical Arabic UGT*
Hadeel Saadany, Constantin Orăsan, Emad Mohamed and Ashraf Tantawy

*Cross-lingual transfer for low-resource Arabic language understanding*
Khadige Abboud, Olga Golovneva and Christopher DiPersio

*Improving POS Tagging for Arabic Dialects on Out-of-Domain Texts*
Noor Abo Mokh, Daniel Dakota and Sandra Kübler

*Domain Adaptation for Arabic Crisis Response*

Reem Alrashdi and Simon O'Keefe

*Weakly and Semi-Supervised Learning for Arabic Text Classification using Monodialectal Language Models*
Reem AlYami and Rabah Al-Zaidy

*Event-Based Knowledge MLM for Arabic Event Detection*
Asma Z Yamani, Amjad K Alsulami and Rabeah A Al-Zaidy

*Establishing a Baseline for Arabic Patents Classification: A Comparison of Twelve Approaches*
Taif Omar Al-Omar, Hend Al-Khalifa and Rawan Al-Matham

*Towards Learning Arabic Morphophonology*
Salam Khalifa, Jordan Kodner and Owen Rambow

*AraDepSu: Detecting Depression and Suicidal Ideation in Arabic Tweets Using Transformers*
Mariam Hassib, Nancy Hossam, Jolie Sameh and Marwan Torki

*End-to-End Speech Translation of Arabic to English Broadcast News*
Fethi Bougares and Salim Jouili

*Arabic Keyphrase Extraction: Enhancing Deep Learning Models with Pre-trained Contextual Embedding and External Features*
Randah Alharbi and Husni Al-Muhtasab

*ArabIE: Joint Entity, Relation and Event Extraction for Arabic*
Niama El Khbir, Nadi Tomeh and Thierry Charnois

*Emoji Sentiment Roles for Sentiment Analysis: A Case Study in Arabic Texts*
Shatha Ali A. Hakami, Robert Hendley and Phillip Smith

*Gulf Arabic Diacritization: Guidelines, Initial Dataset, and Results*
nouf alabbasi, Mohamed Al-Badrashiny, Maryam Aldahmani, Ahmed AlDhanhani, Abdullah Saleh Alhashmi, Fawaghy Ahmed Alhashmi, Khalid Al Hashemi, Rama Emad Alkhobbi, Shamma T Al Maazmi, Mohammed Ali Alyafeai, Mariam M Alzaabi, Mohamed Saqer Alzaabi, Fatma Khalid Badri, Kareem Darwish, Ehab Mansour Diab, Muhammad Morsy Elmallah, Amira Ayman Elnashar, Ashraf Hatim Elneima, MHD Tameem Kabbani, Nour Rabih, Ahmad Saad and Ammar Mamoun Sousou

*Learning From Arabic Corpora But Not Always From Arabic Speakers: A Case Study of the Arabic Wikipedia Editions*
Saied Alshahrani, Esma Wali and Jeanna Matthews

*A Pilot Study on the Collection and Computational Analysis of Linguistic Differences Amongst Men and Women in a Kuwaiti Arabic WhatsApp Dataset*
Hesah Aldihan, Robert Gaizauskas and Susan Fitzmaurice

17:00 - 18:15 ***EMNLP Findings Posters (in-Person & virtual)***

*Improving English-Arabic Transliteration with Phonemic Memories*
Yuanhe Tian, Renze Lou, Xiangyu Pang, Lianxi Wang, Shengyi JIANG and Yan Song

*Sarcasm Detection is Way Too Easy! An Empirical Comparison of Human and Machine Sarcasm Detection*
Ibrahim Abu Farha, Steven Wilson, Silviu Oprea and Walid Magdy

17:00 - 18:15 ***Demos (in-Person & virtual)***

*Beyond Arabic: Software for Perso-Arabic Script Manipulation*
Alexander Gutkin, Cibu Johny, Raiomond Doctor, Brian Roark and Richard Sproat

*Coreference Annotation of an Arabic Corpus using a Virtual World Game*
Wateen Abdullah Aliady, Abdulrahman Aloraini, Christopher Madge, Juntao Yu, Richard Bartle and Massimo Poesio

*NatiQ: An End-to-end Text-to-Speech System for Arabic*

Ahmed Abdelali, Nadir Durrani, Cenk Demiroglu, Fahim Dalvi, Hamdy Mubarak and Kareem Darwish

17:00 - 18:15 ***NADI Shared Task (in-Person & virtual)***

*Optimizing Naive Bayes for Arabic Dialect Identification*
Tommi Jauhiainen, Heidi Jauhiainen and Krister Lindén

*iCompass Working Notes for the Nuanced Arabic Dialect Identification Shared task*
Abir Messaoudi, Chayma Fourati, Hatem Haddad and Moez BenHajhmida

*TF-IDF or Transformers for Arabic Dialect Identification? ITFLOWS participation in the NADI 2022 Shared Task*
Fouad Shammary, Yiyi Chen, Zsolt T Kardkovacs, Mehwish Alam and Haithem Afli

*Domain-Adapted BERT-based Models for Nuanced Arabic Dialect Identification and Tweet Sentiment Analysis*
Giyaseddin Bayrak and ABDUL MAJEED ISSIFU

*Benchmarking transfer learning approaches for sentiment analysis of Arabic dialect*
emna fsih, Sameh Kchaou, Rahma Boujelbane and Lamia Hadrich-Belguith

*SQU-CS @ NADI 2022: Dialectal Arabic Identification using One-vs-One Classification with TF-IDF Weights Computed on Character n-grams*
Abdulrahman Khalifa AAlAbdulsalam

*Ahmed and Khalil at NADI 2022: Transfer Learning and Addressing Class Imbalance for Arabic Dialect Identification and Sentiment Analysis*
Ahmed Oumar and Khalil Mrini

*Arabic Sentiment Analysis by Pretrained Ensemble*
Abdelrahim Qaddoumi

*Dialect & Sentiment Identification in Nuanced Arabic Tweets Using an Ensemble of Prompt-based, Fine-tuned, and Multitask BERT-Based Models*
Reem Abdel-Salam

*On The Arabic Dialects' Identification: Overcoming Challenges of Geographical Similarities Between Arabic dialects and Imbalanced Datasets*
Salma Jamal, Aly M .Kassem, Omar Mohamed and Ali Ashraf

*Arabic dialect identification using machine learning and transformer-based models: Submission to the NADI 2022 Shared Task*
Nouf AlShenaifi and Aqil Azmi

*NLP DI at NADI Shared Task Subtask-1: Sub-word Level Convolutional Neural Models and Pre-trained Binary Classifiers for Dialect Identification*
Vani Kanjirangat, Tanja Samardzic, Ljiljana Dolamic and Fabio Rinaldi

*Word Representation Models for Arabic Dialect Identification*
Mahmoud Sobhy, Ahmed H. Abu El-Atta, Ahmed A. El-Sawy and Hamada Nayel

*Building an Ensemble of Transformer Models for Arabic Dialect Classification and Sentiment Analysis*
Abdullah Salem Khered, Ingy Yasser Hassan Abdou Abdelhalim and Riza Batista-Navarro

*Arabic Dialect Identification and Sentiment Classification using Transformer-based Models*
Joseph Attieh and Fadi Hassan

17:00 - 18:15 ***Gender Rewriting Shared Task Posters (in-Person & virtual)***

*Generative Approach for Gender-Rewriting Task with ArabicT5*
Sultan Alrowili and Vijay Shanker

17:00 - 18:15 ***Propaganda Detection Shared Task Posters (in-Person & virtual)***

*AraProp at WANLP 2022 Shared Task: Leveraging Pre-Trained Language Models for Arabic Propaganda Detection*
Gaurav Singh

*TUB at WANLP22 Shared Task: Using Semantic Similarity for Propaganda Detection in Arabic*
Salar Mohtaj and Sebastian Möller

*SI2M & AIOX Labs at WANLP 2022 Shared Task: Propaganda Detection in Arabic, A Data Augmentation and Name Entity Recognition Approach*
Kamel Gaanoun and Imade Benelallam

*iCompass at WANLP 2022 Shared Task: ARBERT and MARBERT for Multilabel Propaganda Classification of Arabic Tweets*
Bilel - Taboubi, Bechir Brahem and Hatem Haddad

*ChavanKane at WANLP 2022 Shared Task: Large Language Models for Multi-label Propaganda Detection*
Tanmay Chavan and Aditya Manish Kane

*AraBERT Model for Propaganda Detection*
Mohamad Sharara, Wissam Mohamad, Ralph Tawil, Ralph Chobok, Wolf Assi and Antonio Tannoury

*AraBEM at WANLP 2022 Shared Task: Propaganda Detection in Arabic Tweets*
Eshrag Ali Refaee, Basem Ahmed and Motaz Saad

*IITD at WANLP 2022 Shared Task: Multilingual Multi-Granularity Network for Propaganda Detection*
Shubham Mittal and Preslav Nakov

*Pythoneers at WANLP 2022 Shared Task: Monolingual AraBERT for Arabic Propaganda Detection and Span Extraction*
Joseph Attieh and Fadi Hassan

*CNLP-NITS-PP at WANLP 2022 Shared Task: Propaganda Detection in Arabic using Data Augmentation and AraBERT Pre-trained Model*
Sahinur Rahman Laskar, Rahul Singh, Abdullah Faiz Ur Rahman Khilji, Riyanka Manna, Partha Pakray and Sivaji Bandyopadhyay

*NGU CNLP atWANLP 2022 Shared Task: Propaganda Detection in Arabic*
Ahmed Samir Hussein, Abu Bakr Soliman Mohammad, Mohamed Ibrahim, Laila Hesham Afify and Samhaa R. El-Beltagy

17:00 - 18:30 ***Closing Ceremony***

# 10

# Local Guide

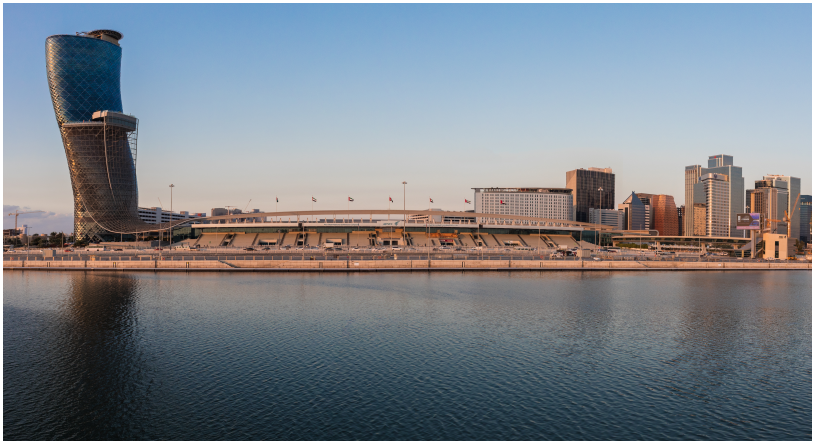
## Abu Dhabi

Abu Dhabi is the capital and the second-most populous city of the United Arab Emirates (UAE) after Dubai. Abu Dhabi is also the capital of the Emirate of Abu Dhabi, one of the seven emirates that comprise the UAE federation. Situated on a set of islands in the Arabian Gulf, Abu Dhabi is a cosmopolitan city with a very large international presence (90% of residents). English is the lingua franca.
To find out more, please visit this page.

## Conference Venue

EMNLP 2022 will take place in the Abu Dhabi National Exhibition Centre (ADNEC) ADNEC is a multi-award-winning venue offering organizers of exhibitions, conferences, and events outstanding facilities spread over a total space of 151,000 square meters. ADNEC is just 15 minutes away from the rapidly expanding Abu Dhabi International Airport which serves over 50 airlines and 100 destinations in 46 countries around the globe. It is only 20 minutes from downtown Abu Dhabi and 90 minutes from Dubai International Airport.

- Address: Abu Dhabi National Exhibitions Company, Khaleej Al Arabi Street, P.O. Box 5546, Abu Dhabi, United Arab Emirates
- Phone: +971 (0) 2 444 6900
- Web: https://adnec.ae



## Transportation

- ADNEC is just a 20-minute car ride from **Abu Dhabi Airport**. From Abu Dhabi Corniche, ADNEC is located just 15 minutes away.
- **Taxis** are available throughout Abu Dhabi and they are quite convenient and affordable. Abu Dhabi's taxi network is operated by TransAD. You can book a taxi in advance by either using the Abu Dhabi Taxi App (AppStore, Google Play) or by calling 600 53 53 53. A number of taxis are available from ADNEC at the adjacent Aloft Hotel entrance. UBER is also available in Abu Dhabi but it is generally a bit more expensive than taxis.
- **Buses** are also available around Abu Dhabi and ADNEC is serviced by bus number 040, seven days a week. For more information on bus routes, services and fares please see the Department of Transport website.

## COVID-19 Regulations

The COVID-19 global situation is continuously changing. For the latest information on Abu Dhabi regulations, please visit this page.

## Accommodation Information

Two hotels have been selected near the conference center that will provide special group rates for attendees.

**Andaz Capital Gate, Abu Dhabi**
Inclusive Rate starting at AED 850
Address: P.O. Box 95165, Abu Dhabi, United Arab Emirates
Tel: +971 2 596 1234
Email: abudhabi.capitalgate@hyatt.com
Website: Boutique 5 star hotel in Abu Dhabi | Andaz Abu Dhabi (hyatt.com)

Located in the iconic Capital Gate building withing the Abu Dhabi National Exhibition Center (ADNEC)
Map, Parking & Transportation
Andaz Capital Gate Abu Dhabi, only 15 minutes away from the airport and most key attractions, adjacent to ADNEC, an impressive 18º leaning design stands tall as the gateway to a luxury lifestyle hotel stay. Arrive a visitor and depart as a local, as the hotel's art gallery, local touch points and cultivated dining experiences will transform your stay.
The 188 stylish guest rooms offer optimum luxury coupled with cutting-edge technology and include digital phone lines, broadband modem points, international direct dialing, and voice mail messaging. The entertainment system offers satellite and interactive television whilst spacious work areas feature high-speed communication lines and Wi-Fi.

**Aloft Abu Dhabi**
Inclusive rates starting at AED 438
Address: Abu Dhabi National Exhibition, AlKhaleej Al Arabi Street, P.O. Box 94943 Abu Dhabi, United Arab Emirates
Tel: +971 2-654 5000
Website: Abu Dhabi 4-Star Hotel in UAE | Aloft Abu Dhabi (marriott.com)

Celebrate style at Aloft Abu Dhabi featuring loft-inspired design in a thriving metropolis. Set in and connected to the ultra-modern Abu Dhabi National Exhibition Centre, we're steps from the diplomatic district and near the airport and city sights. Meet and mingle with friends at our WXYZ Bar, grab a snack from Refuel by Aloft, our pantry open 24 hours a day, or play in our Glow Restaurant & Lounge. Plus, you can always stay connected with complimentary hotel-wide Internet access, and our plug-and-play connectivity station charges all your electronics and links to the 42" Smart TV to maximize work and play. Breeze into one of our Aloft rooms featuring our ultra-comfortable signature bed, an over-sized shower head and more at Aloft Abu Dhabi.

Alternative accommodation options, for example hostels, B&Bs and short-term lodging, can be found via the usual search engines.

Abu Dhabi has an extensive range of modern and traditional hotels featuring internationally recognized brands from budget-conscious to 5-star luxury.

## Childcare

Please check with the specific hotel where you plan to stay for childcare support.

## Dining Options

There are a variety of dining selections located near ADNEC. A short walk or drive away, chances are you will find something to satisfy your tastebuds. From coffee to tapas to cocktails and fine dining. Use a search engine to find a restaurant near you.

https://www.adnec.ae/en/visit/restaurants-listing

There are also a number of cafes (Costa, Starbucks) and two supermarkets (Zoom and Lulu Express) within 5 minute walking from ADNEC.

## Places to Visit

**Emirates Palace**

Located in the heart of Abu Dhabi, Emirates Palace is one of the capital's most well-known landmarks. It is known for its enchanting Arabesque style, award-winning five-star hospitality, and wonderfully unique and authentic local experiences. Directions can be found here.

**Louvre Abu Dhabi**

The iconic Louvre Abu Dhabi is the first universal museum in the Arab World, translating and fostering the spirit of openness between cultures. As one of the premier cultural institutions located in the heart of the Saadiyat Cultural District on https://visitabudhabi.ae/en/where-to-go/islands/saadiyat-island, this art-lovers' dream displays works of historical, cultural, and sociological significance, from ancient times to the contemporary era. Directions can be found here.

**Sheikh Zayed Grand Mosque**

The impressive and inspiring Sheikh Zayed Grand Mosque is one of the world's largest mosques and the only one that captures the unique interactions between Islam and other world cultures. Sheikh Zayed bin Sultan Al Nahyan, the Founder of the UAE, had a very specific vision for this mosque: to incorporate architectural styles from different Muslim civilizations and celebrate cultural diversity by creating a haven that is truly welcoming and inspirational in its foundation. The mosque's architects were British, Italian, and Emirati, with design ideas borrowed from parts of Turkey, Morocco, Pakistan, and Egypt, among other Islamic countries. Directions can be found here.

**Qasr Al Hosn**

Over the centuries, Qasr Al Hosn has been home to the ruling family, acted as the seat of government, housed the National Consultative Council founded by the late Sheikh Zayed Bin Sultan Al Nahyan, Founder of the UAE, as well as being a national archive. Today it stands as the nation's living memorial and a narrator of Abu Dhabi's history. Directions can be found here.

**The National Aquarium Abu Dhabi**

The largest aquarium in the Middle East, The National Aquarium in Al Qana is literally swimming with aquatic wildlife featuring over 46,000 animals from more than 300 unique species. Spread across 10 nautically-themed zones, from the UAE's Natural Treasures, sunken sea wrecks, and Atlantic caves, right through to flooded forests, fiery volcanoes, and a frozen ocean, there are more than 60 attractions that will be sure to delight and excite the whole family. Directions can be found here.

**Yas Island**

This family-friendly entertainment hub is just a 30-minute drive from the city and a 15-minute drive from Abu Dhabi International Airport. Yas Island is home to an F1™ race track, several theme parks, an incredible links golf course, a beautiful beach, stunning hotels, an impressive mall and much more. It's the perfect place to bring the kids, from toddler to teen, for a fun-filled holiday by the sea. Directions can be found here.

**Jubail Mangrove Park**

Jubail Mangrove Park is home to meandering boardwalks that allow you to wander through the mangroves, discovering a nature-rich side to Abu Dhabi and spotting an array of wildlife, from turtles to herons, gazelle, and more. Birdwatchers, nature-lovers, and photographers will be in their element here. Directions can be found here.

**Abu Dhabi's Corniche**

Abu Dhabi's Corniche stretches over eight kilometers of manicured waterfront that includes children's play areas, separate cycle and pedestrian pathways, cafés and restaurants, and a lifeguarded beach. Directions can be found here.

**Abu Dhabi Beaches**

**Soul Beach Saadiyat Island** Tucked away in Saadiyat Island, Soul Beach is Saadiyat Island's new captivating beachfront location, part of the Mamsha Al Saadiyat community.

**Saadiyat Beach** With pristine white sands stretching out along a generous shore, Saadiyat Beach maintains its reputation as one of the most desirable beach locations in the UAE.

**Yas Beach** Set on a majestic stretch of white sand, sun-seekers holidaying at any of Yas Island's hotels and hotel apartments can enjoy complimentary access to the island's crystal clear waters and natural mangrove surrounds.

**Corniche Beach** This beautiful beach isn't just home to turquoise water and soft, white sand, it also has a beautiful seaside boardwalk, with well-kept walkways home to manicured gardens and benches overlooking the picturesque Arabian Gulf.

**Dubai**

Dubai is only 140km (1 hour and 20 minutes by car) away from Abu Dhabi. Dubai is a vibrant, urban, and multicultural city with attractive tourist destinations, such as Burj Khalifa – the tallest building in the world. To find out more, please visit here.

### Sharjah

Sharjah is the third largest Emirate, it is only 165km (1 hour and 50 minutes by car) away from Abu Dhabi. It is a major cultural hub in the UAE, there is also a wide variety of activities awaiting. To find out more, please visit the city's guide here.

## Important Information

### Climate

Abu Dhabi has a northern-hemisphere subtropical, arid climate. November to March is the most appealing time of year, and it is also when infrequent winter rains occur. The temperature range in the winter months is between 56º F (13º C) and 75º F (24º C) with typically bright sunny days which correspond to the best kind of spring weather in the US.

### Clothing

Summer clothing may be worn for most of the year, but during the winter evening temperatures may occasionally call for a jacket or light coat. While dress codes are fairly liberal, consideration should be given not to offend the sensibilities of others. Swimwear should be worn only on beaches or at swimming pools. Visiting shopping malls and other attractions, tourists should wear clothing that is not too tight or revealing. Certain attractions such as mosques or religious sites usually have stricter dress codes, requiring both men and women to cover up bare shoulders, arms and legs, and women to wear headscarves.

### Communications

The international dialing code for incoming calls to landlines in the UAE is +971 and 02 for Abu Dhabi. Calls to and from landlines within Abu Dhabi are free. Direct dialing is possible to over 170 countries. UAE has two mobile networks, Du and Etisalat:

- Both offer temporary SIM cards for tourists and business travelers, including data and calls, and these can be purchased at outlets across the UAE, including at the airport and malls.
- Roaming services are also available for most visitors if they wish to use their existing number and phone.

### Currency & Living Cost

The monetary unit is the Dirham (AED), which is divided into 100 fils. The exchange rate is pegged to the US Dollar at the rate $1 = AED 3.675.
The average costs at an average coffee shop or restaurant are as follows:

- Cup of coffee: 14-20 AED/$3.8-5.5 USD
- Sandwich lunch: 15-30 AED/$5-8.2 USD
- Evening meal: 30-50 AED/$8.2–13.6 USD
- Check out this site to compare living cost with Abu Dhabi.

### Electricity

The electricity supply in Abu Dhabi is 220/240 volts at 50 cycles. Standard British-type 13-amp square three-pin plugs are the norm in most hotels. European or US-made appliances may need a plug adapter.

**Language**

The official language of the UAE is Arabic. However, English is also very widely spoken throughout Abu Dhabi, especially in business, hospitality, retail environments, street signs, taxis, restaurant menus, etc. Urdu and Hindi are also widely spoken. For some basic Arabic phrases used commonly, check out this site. The top eight to learn are:

- Marhaba: Hello
- Ahlan: Welcome
- Ma'asalama: Good Bye
- Shukran: Thank you
- Mabrook: Congratulations
- Yalla: Let's go
- Khalas: enough/done
- Inshallah: God willing, may be, no, and a host of other context dependent meanings

**Security**

As one of the most cosmopolitan and multicultural cities in the world, home to over 200 different nationalities, Abu Dhabi is an advocate for peace and stability, and proud to be a connecting hub between East and West. Abu Dhabi is ranked in the 10 safest cities by Aon Hewitt, with low crime rates, a stable government, and a department of Abu Dhabi Police dedicated entirely to visitors.

**Taxation**

The UAE does not levy income tax on individuals. Value Added Tax is levied on a majority of goods and services.

**Time Zone & Business Hours**

Abu Dhabi is GMT+4. Most businesses are open from 8 am to 6 pm, Monday to Friday, with Saturday and Sunday being official holidays for all government departments. Embassies, consulates, and government offices operate from 7:30 am to 2:30 pm, Monday to Friday.

**Tipping & Gratuities**

Tipping practices are similar to most other parts of the world. Most restaurants include a 10% service charge, but tipping, in general, is at the customer's discretion.

**Water**

The tap water in Abu Dhabi is safe to drink. But locally bottled water is generally served in hotels and restaurants.

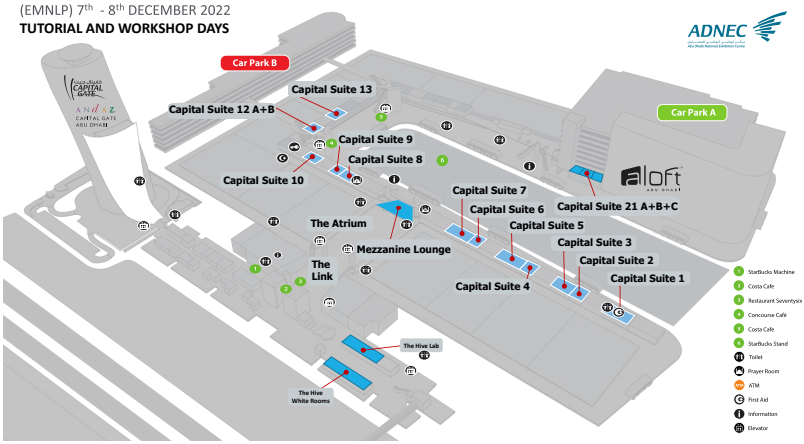
**Emergency**

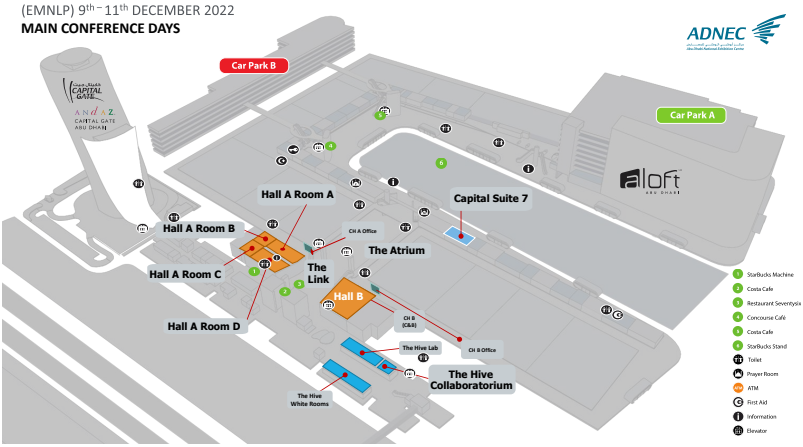
The emergency phone number for Abu Dhabi Police is 999. Whether you need police assistance, an ambulance, or for any other emergency, 999 is the number to call and calls are free. When calling 999, please remember to state your name, the nature of the accident, address of the emergency and how serious the situation is. Please find a detailed list of other emergency numbers at the bottom of this page.
If you're involved in a traffic accident, it's important to contact the police immediately. In case of a minor incident, move your car to the roadside, as there are fines for obstructing traffic. Remember, you cannot file an insurance claim without a police report.
For other enquiries, Abu Dhabi Police operates a dedicated Tourism Police section which will advise and guide you on a range of matters. You can contact them on +971 2 800 2626 and +971 2 512 7777, or visit Emergency Help for Tourists.
In a medical emergency, Abu Dhabi's Sheikh Khalifa Medical City (+971 2 819 0000) and Mediclinic Al Noor Hospital (800 2000) both have Accident and Emergency units. If you're injured in a traffic accident, you will automatically be taken to Sheikh Khalifa Medical City, as it has the best Accident & Emergency treatment facilities.
The Abu Dhabi Government portal provides an updated list of 24-hour pharmacies and medical services, including hospitals, clinics, and medical centres. If you don't have internet access you can call the toll-free number 800 555 (+971 2 666 4442).

# 11 Venue Map

## ADNEC Floor Plan

EMNLP 2022 is held in the Abu Dhabi National Exhibition Centre (ADNEC).
**The Registration Desk** (December 7 to 11, 2022) is in The Link (next to Hall B).
**Workshops and tutorials** (December 7 and 8, 2022) are in Capital Suites 1 to 10, 12A, 12B, 13, 21A, 21B and 21C. Coffee breaks are held in the walkways outside the capital suites. Workshop posters are presented in the Mezzanine Lounge and its adjacent halls.
**The Main Conference** (December 9 to 11, 2022) is held in Hall A (Rooms A, B, C, and D), Hall B, and The Hive Collaboratorium. Demos, posters and exhibitions, as well as coffee breaks are in The Link and The Atrium areas, between and adjacent to Halls A and B.

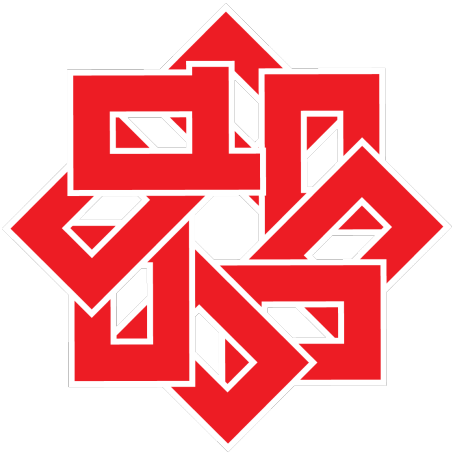



(EMNLP) 9th – 11th DECEMBER 2022
MAIN CONFERENCE DAYS
ADNEC
Car Park B
Car Park A
Hall A Room A
Capital Suite 7
Hall A Room B
The Atrium
Hall A Room C
The Link
Hall B
Hall A Room D
The Hive Collaboratorium
StarBucks Machine
Costa Cafe
Restaurant Seventysix
Concourse Café
Costa Cafe
StarBucks Stand
Toilet
Prayer Room
ATM
First Aid
Information
Elevator

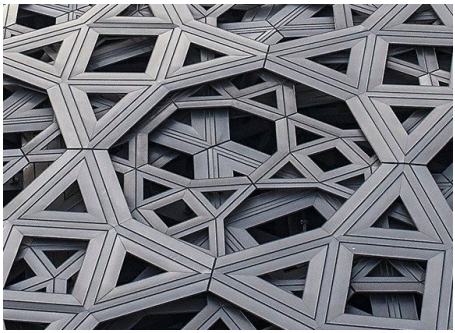
12

# EMNLP 2022 Logo

## The EMNLP 2022 Logo



The EMNLP 2022 logo features an eight-pointed star constructed from rotated repetitions of the ACL's iconic logo. The eight-pointed star is a classic Islamic Arabesque design motif. It is not unique to Abu Dhabi, but quite beloved here. We see it from the design of the central prayer hall in the Grand Mosque to the dome of the Louvre Museum. The logo was designed by Nizar Habash.

# Index

# NYU Abu Dhabi welcomes and educates global citizens, and produces knowledge that promotes human understanding and betters society.

Through distinctive global liberal arts education and graduate programs, we enable students and graduates to achieve intellectual, personal, and professional fulfillment and empower them to make significant societal contributions to Abu Dhabi and the world. Through cutting-edge research, we develop knowledge, foster creativity and innovation, and help solve humanity's shared challenges. Together, we contribute to Abu Dhabi's knowledge-based economy and society, and play a central role in NYU's global mission.

جامعة نيويورك أبوظبي

NYU ABU DHABI

# Come build the future with us

At Amazon, we fundamentally believe that scientific innovation is essential to being the most customer-centric company in the world. It's the company's ability to have an impact at scale that allows us to attract some of the brightest minds in artificial intelligence, machine learning, and related fields.

**Connect with us to learn more about Amazon science, including internships, collaborations, and career opportunities at EMNLP-2022@amazon.com**

amazon | science

**Learn more at amazon.science**

Meta AI

Realizing the potential of AI today and creating the experiences of tomorrow.

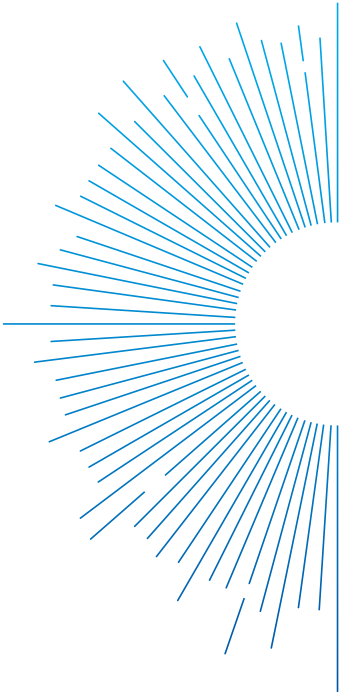
Help us pioneer the future of AI:
www.metacareers.com


Make the difference.
At Bloomberg, we use the power of technology to bring clarity to a complex world. In a career here, you'll help create products that our global customers rely on to make critical financial decisions. We work on purpose.
Come find yours.
bloomberg.com/careers
Bloomberg

# BAIDU NLP

## BAIDU NATURAL LANGUAGE PROCESSING

On a mission to enable machines to understand language and acquire intelligence so as to make the world better, Baidu NLP is dedicated to core NLP technologies, leading technology platforms and innovative products that are set to serve users across the globe and make the complex world simpler.



Baidu is the leading Chinese language internet search provider. Baidu aims to make the complicated world simpler through technology.

Email: nlp@baidu.com
Web: ai.baidu.com

# Megagon Labs

Megagon Labs is an innovation hub within the Recruit Group, conducting top-notch research and building technologies in Mountain View and Tokyo. We are making impacts through the Recruit Group's worldwide services and products by collaborating with its subsidiaries such as Indeed.com and Glassdoor. Our mission is to empower people with better information to make their best decision.

The areas we focus are Natural Language Processing, Machine Learning, Data Management, Data Integration, Human-Computer Interaction, and Intelligent Visual Analytics.

For more information about our lab and hiring, please visit the Megagon Labs booth or www.megagon.ai!

## Accepted Papers at EMNLP 2022

- Main Conference -

**Summarizing Community-based Question-Answer Pairs**
Ting-Yao Hsu, Xiaolan Wang, Yoshihiko Suhara

- Findings -

**Low-resource Interactive Active Labeling for Fine-tuning Language Models**
Seiji Maekawa, Dan Zhang, Hannah Kim, Sajjadur Rahman, Estevam Hruschka

www.megagon.ai

# NYU Abu Dhabi welcomes and educates global citizens, and produces knowledge that promotes human understanding and betters society.

Through distinctive global liberal arts education and graduate programs, we enable students and graduates to achieve intellectual, personal, and professional fulfillment and empower them to make significant societal contributions to Abu Dhabi and the world. Through cutting-edge research, we develop knowledge, foster creativity and innovation, and help solve humanity's shared challenges. Together, we contribute to Abu Dhabi's knowledge-based economy and society, and play a central role in NYU's global mission.

جامعة نيويورك ابوظبي

NYU ABU DHABI

# NOAH’S ARK LAB OF HUAWEI TECHNOLOGIES

The Noah’s Ark Lab is the AI research center for Huawei Technologies, located in Hong Kong, Shenzhen, Beijing, Shanghai, Xi’an, London, Paris, Toronto, Montreal, Edmonton and many more. We welcome talented researchers and engineers to join us to realize their dreams.

The mission of the lab is to make significant contributions to both the company and society by innovating in artificial intelligence, data mining and related fields. Mainly driven by long term and big impact projects, research in the lab also tries to advance the state of the art in numerous fields as well as to harness the products and services of the company, at each stage of the innovation process.

Founded in 2012, the lab has now grown to be a research organization with many significant achievements in both academia and industry. Research areas of the lab mainly include AI Theory, AI System Engineering, Speech and Language Computing, Computer Vision, Decision Making and Reasoning, Recommendation and Search.

## Research Areas


## Application

Interested candidates should send application materials including resume to noahlab@huawei.com

More information about the lab is available at https://www.noahlab.com.hk

HUAWEI | NOAH'S ARK LAB

NOAH'S ARK LAB

## Supporting Partner

مكتب المؤتمرات والمعارض
CONVENTION & EXHIBITION BUREAU

## Diamond Sponsors

جامعة نيويورك أبوظبي
NYU ABU DHABI

جامعة محمد بن زايد للذكاء الاصطناعي
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

## Platinum Sponsors

## Gold Sponsors

## Silver Sponsors

## Bronze Sponsors

Orchestrating a brighter world

Abu Dhabi Culture & Tourism للثقافة والسياحة

مكتب المؤتمرات
CONVENTION BUREAU

amazon | science

Bloomberg
Engineering

Google

جامعة نيويورك أبوظبي
NYU | ABU DHABI

جامعة محمد بن زايد للذكاء الاصطناعي
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Meta