ACL

Association for Computational Linguistics

ACL 2022

22ND – 27TH MAY | 60TH MEETING | DUBLIN

CONFERENCE HANDBOOK (FULL)

Contents

1

Conference Information

Message from the General Chair

Welcome to ACL 2022, the 60th Annual Meeting of the Association for Computational Linguistics! The conference will be held in Dublin, the capital of Ireland, on May 22–27, 2022.

ACL 2022 will be a hybrid conference. After two fully virtual editions, ACL 2020 and ACL 2021, due to the covid-19 pandemic, this year we are gradually coming back to normality, estimating, at the moment of writing this message, that about 50% of the registered participants will be able to attend the conference in-person, enjoying the atmosphere of the CCD congress center, the social events of the conference, and the many opportunities in Dublin. On the other side, virtual attendees will have the possibility to interact almost like they were in Dublin, thanks to a sophisticated virtual conference platform.

There are few important innovations this year. The most relevant is that ACL 2022 adopted a new reviewing process, based on "rolling review" (ARR), with the goal of coordinating and making more efficient the paper reviews of the ACL conferences. This initiative was shared with NAACL 2022, resulting in a coordinated effort. As a side effect of moving to ARR, we have been working on a new version of the software, called ACLPUB2, used to produce both the conference proceedings and the conference schedule. I would like to thank all the people who contributed to those achievements. Finally, this year we celebrate the 60th anniversary of the ACL conference. Thanks to the enthusiastic contributions of many organizations, coordinated by the Diversity and Inclusion co-chairs, we are preparing a very special initiative for our community, which, at the time of writing this message, is still secret and that will be disclosed during the opening of the conference.

I was very lucky to work together with three fantastic Program Chairs: Preslav Nakov, Smaranda Muresan and Aline Villavicencio. I could not thank you more for the dedication and the capacity with which you have organized a very exciting scientific program and for the help in all the phases of the conference organization.

Thanks to the local organizers in Dublin, Andy Way and John Kelleher, and to the PCO, who managed the local organization in a period in which we have had very few certainties, and many more uncertainties.

We are extremely grateful to all sponsors for their continuing and generous support to help our conferences be very successful. Thank you to Chris Callison-Burch, the ACL Sponsorship Director, for managing the relations between the sponsors and ACL 2022.

I am also very grateful to the chairs of the previous years' conferences, who were always ready to help and to provide advice, contributing to the transmission, from year to year, of all the know-how and collective memory. Thanks to all the members of The ACL Executive Committee, they were always supportive, particularly when feedback on delicate issues was needed.

Many thanks to the senior area chairs, the area chairs, the reviewers, our workshop organizers, our tutorial instructors, the authors and presenters of papers, and the invited speakers.

ACL requires a long process, involving a large team of committed people. It is an honor for me to have coordinated such a team of talented people, who kindly volunteered their time to make this conference possible. I would like to thank the members of the organizing committee for their dedication and hard work, often under a tight schedule:

- Workshop Co-Chairs: Elena Cabrio, Sujian Li, Mausam;
- Tutorial Co-Chairs: Naoaki Okazaki, Yves Scherrer, Marcos Zampieri;
- Demo Co-Chairs: Valerio Basile, Zornitsa Kozareva, Sanja Štajner;
- Student Research Workshop Co-Chairs: Samuel Louvan, Brielen Madureira, Andrea Madotto;
- SRW Faculty Advisors: Cecile Paris, Siva Reddy, German Rigau;
- Publication Co-Chairs (also publication co-chairs for NAACL 2022): Danilo Croce, Ryan Cotterell, Jordan Zhang;
- Conference Handbook Chair: Marco Polignano;
- Diversity & Inclusion Co-chairs: Mona Diab, Martha Yifiru Tachbelie;
- Ethic advisor committee: Su Lin Blodgett, Christiane Fellbaum;
- Technical OpenReview Chair: Rodrigo Wilkens;
- Publicity and Social Media Co-chairs: Isabelle Augenstein, Emmanuele Chersoni, Diana Maynard, Soujanya Poria, Joel Tetreault;
- Local Arrangement Committee: Fiona McGillivray, Greg Carew, Laird Smith;
- Student Volunteer Coordinators: Filip Klubicka, Vasudevan Nedumpozhimana, Guodong Xie, Pintu Lohar;
- Internal Communications Chair: Marcely Boito Zanon.

Let me deserve a special thanks to Priscilla Rasmussen. She has been the pillar not only of this year's ACL, but of the ACL conferences for many years. She has offered her invaluable experience to the organizing committee, and her presence has always given us a pleasant sense of security.

Finally, I would like to thank all the participants, both in-person and virtual, who will be the main actors from May 22 to May 27, 2022. I am convinced that we will experience a fantastic conference, scientifically exciting and full of fond memories.

Welcome and hope you all enjoy the conference!

Bernardo Magnini (FBK, Italy)
ACL 2022 General Chair

Message from the Program Chairs

Welcome to the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022). ACL 2022 has a special historical significance, as this is the 60th Anniversary edition. It is also the first hybrid ACL conference after two years of a fully virtual format for ACL in 2020 and 2021 due to the COVID-19 pandemic. Finally, it is the first *ACL conference to fully embrace the ACL Rolling Review (ARR) as a reviewing process. Below, we discuss some of these changes and we highlight the exciting program that we have put together with the help from our community.

Using ARR for Reviewing

In coordination with the NAACL 2022 team and the ACL executive committee, we decided to fully adopt the ACL Rolling Review (ARR) as the only reviewing platform for ACL 2022. ARR is a new review system for *ACL conferences, where reviewing and acceptance of papers to publication venues is done in a two-step process: (*i*) centralized rolling review via ARR, and (*ii*) commitment to a publication venue, e.g., ACL 2022. The purpose of the ACL Rolling Review is to improve the efficiency and the turnaround of reviewing in *ACL conferences while keeping diversity (geographic and otherwise) and editorial freedom.

As ACL 2022 is the first conference to fully adopt the ARR review process, we worked very closely with ARR and we coordinated our efforts with the NAACL 2022 PC chairs. In particular, given the short distance between ACL 2022 and NAACL 2022, we allowed authors to commit their papers to ACL 2022 and simultaneously to submit a revision to ARR in January, which were eligible for NAACL 2022. We also joined ARR as Guest Editors-in-Chief (EiCs) to help with the September–November submissions to ARR, which primarily targeted ACL 2022. We worked together to integrate ARR and some of the conference workflows to ensure scaling up, and to maintain the quality and the timely processing of the submissions for November, and thus to guarantee that all papers submitted by the November 15, 2021 ARR deadline could be considered for ACL 2022 if the authors decided to commit them. This required making sure we had all reviews and meta-reviews ready in time, which we managed to achieve thanks to the combined efforts of the ARR and the ACL 2022 teams. We would also like to note that this is a community effort, and we are grateful for the support of the authors, the reviewers, the Action Editors (AEs), and the Senior Area Chairs (SACs), who have been constructively engaging and helping with ARR and ACL 2022.

Committing to ACL 2022

The commitment form for ACL 2022 asked the authors to provide a link to their paper in ARR: we asked for a link to the latest version of the paper that had reviews and a meta-review. The authors also needed to select an area (including the Special Theme area) they were submitting their paper to (this was needed as ACL 2022 had areas, while ARR did not). Finally, the authors were allowed to submit optional comments to the ACL 2022 Senior Area Chairs (SACs). Note that these comments were only visible to the SACs, and they were not sent to the reviewers or to the Action Editors: the rationale was that responding to reviewers and Action Editors should be handled in a response letter if the authors decided to do a resubmission in ARR, which is a completely different process than committing a paper to ACL 2022. These comments to the SACs were designed mainly to raise concerns about objective misunderstandings by the reviewers and/or by the Action Editor about the technical aspect of the paper that the authors believed might help the SACs in their decision-making process.

Areas While ARR did not have areas, ACL 2022 did: it had 23 areas, including the 22 areas from ACL 2021 plus our Special Theme. Our special theme was on "*Language Diversity: from Low-Resource to Endangered Languages*," to commemorate the 60th anniversary of ACL with the goal of reflecting and stimulating a discussion about how advances in computational linguistics and natural language processing

can be used to promote language diversity from low-resource to endangered languages. We invited papers that discuss and reflect on the "role of the speech and language technologies in sustaining language use" (Bird, 2020) for the large variety of world languages with focus on under-resourced, indigenous, and/or endangered languages. We were interested in the challenges for developing and scaling up the current NLP technologies for the rich diversity of human languages and in the ethical, cultural, and policy implications of such technologies for local communities. We also have a best Theme paper award category.

Acceptance to ACL 2022

As ACL 2022 submissions in ARR, we count all papers from September, October, and November, which we advertised as ACL 2022 months, after removing all re-submissions and also nine papers that selected NAACL 2022 as a preferred venue (a total of 3,360 papers) + the papers from the May–August period that were actually committed to ACL 2022 and that were not resubmissions (a total of 18 papers), for a total of 3,378 papers.

This number is on par with the number of submissions to ACL 2021, which received 3,350 submissions. Subsequently, 1,918 papers were committed to ACL 2022 (i.e., 57%). After the review process, 701 papers (604 long and 97 short) were accepted into the main conference.

Acceptance Rates for the Main Conference

The quality of a conference is often perceived based on the acceptance rate of the papers submitted there, and thus it is important to have an acceptance rate that adequately represents the difficulty of publishing a paper in the conference. Given the adoption of ARR, it is also important to allow for consistency across various conferences. Thus, ACL 2022 (and NAACL 2022) adopted the following two ways of calculating the acceptance rates:

(a) *(Number of accepted papers at ACL 2022) / (Number of papers that selected ACL 2022 as the preferred venue in ARR or were committed to ACL 2022).* For ACL 2022, for the denominator we consider the 3,378 papers as explained above. Thus, the acceptance rate is 701 / 3,378 = 20.75% for the Main conference.

(b) *(Number of accepted papers at ACL 2022) / (Number of papers committed to ACL 2022).* For the denominator, we had 1,918 papers committed to ACL 2022, and thus, the acceptance rate is 701 / 1,918 = 36.54% for the Main conference.

Note that option (a) is closer to the way the acceptance rate was computed at previous *ACL conferences, where submitting and committing a paper was done in one step and papers were rarely withdrawn after the reviews, the meta-reviews, and the corresponding scores were released. However, one issue with this option for ACL 2022 was that indicating a preferred venue was only enabled starting with the October ARR submissions, and it was not available for earlier months. As mentioned above, we removed a small number of papers from our denominator that selected NAACL 2022 as a preferred venue in October and November (a total of 9 papers) and we considered the ARR submissions only for the months of September, October, and November, as these months were advertised in our CFP, plus any papers that were committed to ACL 2022 from earlier months (May–July) and which were also not resubmissions.

Option (b) yields a higher "acceptance rate", as many authors with low reviewing scores chose not to commit their paper to ACL 2022.

Best Paper Awards

From the committed ACL 2022 papers, we selected 32 papers as candidates for the following Best Paper awards, based on nominations by the Senior Area Chairs: Best Research Paper, Best Special Theme Paper, Best Resource Paper, and Best Linguistic Insight Paper. These papers were assessed by the Best Paper Award Committee. The selected best papers will be presented in a dedicated plenary session for Best Paper Awards on May 24, 2022.

Findings of ACL 2022

Given the success of the Findings at EMNLP 2020 and 2021 and ACL-IJCNLP 2021, we also have Findings of ACL 2022 papers, which are papers that were not accepted for publication in the main conference, but nonetheless were assessed by the Program Committee as solid work with sufficient substance, quality, and novelty. A total of 361 papers were offered to be included in the Findings of ACL 2022. Given the two ways of computing acceptance rates described above, this results in a 10.68% acceptance rate in option (a), and 19.82% in option (b). Out of the 361 papers, 30 papers declined the offer, leading to 331 papers to be published in the Findings of ACL 2022. In order to increase the visibility of the Finding of ACL 2022 papers, we offered the authors of these 331 papers the possibility to present their work as a poster at ACL 2022, in addition to making a 6-minute or a 3-minute video to be included in the virtual conference site (for long and for short papers, respectively). The authors of 305 of the 331 papers accepted our invitation to present their work as a poster at ACL 2022.

TACL and Computational Linguistics

Continuing the tradition from previous years, ACL 2022 also features 43 articles that were published at the Transactions of the Association for Computational Linguistics (TACL) and 8 papers from the Computational Linguistics journal.

Keynote and Invited Speakers

Another highlight of our program are the keynotes, which we run in three different formats:

- **a keynote talk** by Angela Friederici (Max Planck Institute for Human Cognitive and Brain Sciences) on "*Language in the Human Brain*";

- **a keynote fire-side chat** on "*The Trajectory of ACL and the Next 60 years*" with Barbara Grosz (Harvard University) and Yejin Choi (University of Washington and Allen Institute for Artificial Intelligence), moderated by Rada Mihalcea (University of Michigan);

- **a keynote panel** on "Supporting Linguistic Diversity" led by Steven Bird (Charles Darwin University), with panelists representing a variety of world languages, including Robert Jimerson, Rochester Institute of Technology (Seneca, USA), Fajri Koto, The University of Melbourne (Minangkabau, Indonesia), Heather Lent, University of Copenhagen (Creole languages), Teresa Lynn, Dublin City University (Irish), Manuel Mager, University of Stuttgart (Wixaritari, Mexico), Perez Ogayo, Carnegie Mellon University (Luo and Kiswahili, Kenya)

We further had two additional invited talk initiatives:

- **Spotlight Talks by Young Research Stars (STIRS)** by Eunsol Choi (University of Texas at Austin), Ryan Cotterell (ETH Zurich), Sebastian Ruder (Google, London), Swabha Swayamdipta (Allen Institute for AI), and Diyi Yang (Georgia Tech);

- **Next Big Ideas Talks** by Marco Baroni (Pompeu Fabra University), Eduard Hovy (The University of Melbourne and Carnegie Mellon University), Heng Ji (UIUC), Mirella Lapata (University of Edinburgh), Hang Li (Bytedance Technology), Dan Roth (University of Pennsylvania and Amazon), and Thamar Solorio (University of Houston).

Thank You

ACL 2022 is the result of a collaborative effort and a supportive community, and we want to acknowledge the efforts of so many people who have made significant efforts into the organization of ACL 2022! First of all, we would like to thank our Program Committee (the full list of names is quite long and it is included in the Program Committee pages of the Proceedings):

- Our awesome 82 Senior Area Chairs who were instrumental in every aspect of the review process, from liaising with ARR, to supporting the implementation of a two-stage reviewing system, recommending Action Editors and reviewers, working on paper acceptance, and nomination of best papers and outstanding reviewers. For all of them, this involved familiarizing themselves with a new protocol to accommodate the integration of ARR reviews and a new system, and for many of them, the scope of their responsibilities was equivalent to chairing a small conference.

- The 363 ARR Action Editors (from the June–November ARR cycles), who had the role of ACL 2022 Area Chairs interacting with reviewers, leading paper review discussions, and writing meta-reviews.

- The 2,323 ARR reviewers (from the June–November ARR cycles), who contributed for the ACL 2022 reviewing cycles, providing valuable feedback to the authors.

- The emergency ARR Action Editors and reviewers, who provided their support at the last minute to ensure a timely reviewing process.

- The amazing ARR team, who collaborated in the challenge of managing and implementing the ARR reviewing needed for the scale of ACL 2022. In particular, we acknowledge Amanda Stent and Goran Glavaš as Guest ARR Editors-in-Chief for ACL 2022, Graham Neubig as Guest ARR Chief Technical Officer for ACL 2022, and Sara Goggi as Guest ARR Editorial Manager for ACL 2022.

ACL 2022 counted on the contributions of many wonderful committees, including:

- Our Best Paper Selection Committee, who selected the best papers and the outstanding papers: Tim Baldwin, Kathleen McKeown, David Chiang, Min-Yen Kan, and Taro Watanabe.

- Our Ethics Advisory Committee, chaired by Christiane Fellbaum and Su Lin Blodgett, for their hard work to ensure that all the accepted papers addressed the ethical issues appropriately, under a very tight schedule and on a new platform.

- Our amazing Publication Chair Danilo Croce, our Handbook Chair Marco Polignano, the Technical OpenReview Chair Rodrigo Wilkens, and the Scheduler Chair Jordan Zhang, who jointly with the NAACL 2022 Publication Chair, Ryan Cotterell, made an enormous contribution to the community by implementing the integration scripts for generating the proceedings, the handbook and the schedule from the OpenReview platform.

- Our Publicity Chairs Isabelle Augenstein, Emmanuele Chersoni, Diana Maynard, Soujanya Poria, and Joel Tetreault, for their work on managing the communications on social media platforms.

- The Internal Communications Chair Marcely Boito Zanon for streamlining the processes.

- The wonderful Technical OpenReview Chair Rodrigo Wilkens, who went above and beyond to ensure that the typical ACL conference functionalities were translated to a new environment.

We would also like to thank many people who helped us with various software used for the conference:

- The ARR Tech team, in particular Sebastin Santy and Yoshitomo Matsubara, who served as Guest ARR Tech Team for ACL 2022.
- The OpenReview team, in particular Nadia L'Bahy, Celeste Martinez Gomez, and Melisa Bok, who helped to implement the integration of ARR as a reviewing platform for ACL 2022.
- The whole Underline team, in particular Sol Rosenberg, Jernej Masnec, Damira Mršić, and Mateo Antonic, who created a virtual site for the conference.

As Program chairs, we had to deal with many tasks, including handling new protocols and situations and a new conference management environment. We would not be able to complete these tasks without the advice from our colleagues, including

- Our fantastic General Chair Bernardo Magnini, who provided invaluable support and feedback throughout the whole process, including collaborating on the efforts to take on the challenge of reengineering the conference reviewing processes and pipeline.
- The Program Co-Chairs of NAACL 2022 Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, and the NAACL 2022 General Chair, Dan Roth, for collaborating in the challenge of coordinated adoption of ARR reviewing in a full scale for ACL 2022 and NAACL 2022.
- The Program Co-Chairs of previous editions of *ACL conferences, in particular the ACL-IJCNLP 2021 PC chairs Roberto Navigli, Fei Xia, and Wenjie Li, as well as the EMNLP 2021 PC chairs Lucia Specia, Scott Wen-tau Yih, and Xuanjing Huang for providing amazing guidance and support, and sharing their experience and answering our many questions, often on short notice.
- The ACL Executive Committee, especially Tim Baldwin (the ACL President), Rada Mihalcea (the ACL Past President), Shiqi Zhao (Secretary), Priscilla Rasmussen (Business Manager), and the members of the ACL executive committee for providing invaluable feedback and for helping us sort through various issues.
- The Computational Linguistics Editor-in-Chief Hwee Tou Ng, the TACL Editors-in-Chief Ani Nenkova and Brian Roark, and the TACL Editorial Assistant Cindy Robinson, for coordinating the Computational Linguistics and the TACL presentations at ACL 2022.

We would also like to thank all the authors who submitted/committed their work to ACL 2022. Although we were only able to accept a small percentage of the submissions, your hard work makes this conference exciting and our community strong. Our huge thanks goes to the *ACL communities for the kind and patient support during a year of major changes in our submission and reviewing processes.

Last, but not least, we thank our students, interns, postdocs, colleagues, and families for being so understanding and supportive during this intense year, and especially when we were swamped by countless conference deadlines and meetings. Our deepest gratitude is to all of you. We hope you will enjoy this 60th Anniversary edition of ACL.

Smaranda Muresan (Columbia University and Amazon AWS AI Labs, USA)
Preslav Nakov (Qatar Computing Research Institute, HBKU)
Aline Villavicencio (University of Sheffield, UK)

ACL 2022 Program Committee Co-Chairs

Message from the Local Chairs

Back in March 2020, just after the first COVID-19 lockdown, we submitted our bid for Dublin to host ACL 2022, conference that you are currently attending. In November 2020, we learned that our bid had been successful, which we were of course delighted to hear. Of course, at that stage – and at many points in between – we have wondered whether we would be able to meet face-to-face at all, and it is great that we are able to host you in the wonderful city of Dublin where we are privileged to live, as well as accommodating many of you online.

ACL is an opportunity to welcome not just our European friends and colleagues, but also those from farther afield. Ireland punches above its weight in the areas of NLP and Machine Learning, principally through the SFI-funded €100 million ADAPT Centre for Digital Content Technology, which comprises experts from 4 local Dublin universities as well as 4 further universities from across the country in a range of disciplines in AI. We have internationally renowned groups in machine translation, information retrieval, speech technology, parsing and grammar Induction, among others, so we believe it is appropriate that ACL is being held in our country for the first time. We are of course grateful to everyone who submitted a paper; whether your work was selected for presentation or not, if no-one had submitted, we wouldn't have had a conference. For those of you whose work was selected for presentation, many thanks for coming to Dublin, or for presenting online.

Along the way, we have been helped greatly by the General Chair Bernardo Magnini, and by Priscilla Rasmussen and others from the ACL executive team, to whom we are extremely thankful. However, by far the biggest thanks are due to Greg Carew and his team in Abbey Conference and Events for their professional support of the conference. You will have met them at registration, and they are available throughout the event to ensure your needs are met. We have been engaging with them for 2 years now on ACL, and for longer as they helped Andy host the MT Summit in 2019. We could not have made a better choice of PCO to assist us with all the requirements involved in hosting the best-regarded conference in our area. This has been a true partnership that has made this journey an enjoyable one.

We are also extremely grateful to Fáilte Ireland for their extremely generous support of this conference, and to our PostDocs Guodong Xie & Pintu Lohar (with Andy at DCU), and Vasudevan Nedumpozhimana & Filip Klubička (with John at TUD) for their huge efforts to recruit and manage the small army of student volunteers. Finally, we really hope that you all enjoy the conference, that you benefit from the excellent programme that has been assembled, and that you go away from here having made new friends. We are fortunate indeed that many of our very best friends are in the computational linguistics community, and we will try our very best to meet as many of you as possible during the event.

Andy Way (Dublin City University, Ireland)
John Kelleher (TU Dublin, Ireland)

Local Chairs, ACL 2022

Organizing Committee

General Chair

Bernardo Magnini , FBK, Italy

Program Chairs

Smaranda Muresan , Columbia University and Amazon AWS AI Labs, USA
Preslav Nakov , Qatar Computing Research Institute, HBKU, Qatar
Aline Villavicencio , University of Sheffield, UK

Local Organization Chairs

John Kelleher , TU Dublin, Ireland
Andy Way , Dublin City University, Ireland

Workshop Chairs

Elena Cabrio , Université Côte d'Azur, France
Sujian Li , Pekin University, China
Mausam , IIT Delhi, India

Tutorial Chairs

Luciana Benotti , National University of Córdoba, Argentina
Naoaki Okazaki , Tokyo Institute of Technology, Japan
Yves Scherrer , University of Helsinki, Finland
Marcos Zampieri , Rochester Institute of Technology, USA

Demo Chairs

Valerio Basile , University of Turin, Italy
Zornitsa Kozareva , Facebook AI Research, USA
Sanja Štajner , Symanto Research, Germany

Student Research Workshop Chairs

Samuel Louvan , FBK, Italy
Andrea Madotto , HKUST, Hong Kong
Brielen Madureira , University of Potsdam, Germany

Student Research Workshop: Faculty Advisors

Cecile Paris , CSIRO, Australia
Siva Reddy , McGill University, Canada
German Rigau , Basque Country University, Spain

Publicity and Social Media Chairs

Isabelle Augenstein , University of Copenhagen, Denmark

Emmanuele Chersoni , The Hong Kong Polytechnic University, Hong Kong
Diana Maynard , University of Sheffield, UK
Soujanya Poria , Singapore University of Technology, Singapore
Joel Tetreault , Dataminr, USA

Publication Chairs

Danilo Croce , University of Rome Tor Vergata, Italy
Ryan Cotterell , ETH Zürich, Switzerland
Jordan Zhang

Handbook Chair

Marco Polignano , University of Bari Aldo Moro, Italy

Technical OpenReview Chair

Rodrigo Wilkens , Université catholique de Louvain, Belgium

Conference App Chair

Pierluigi Cassotti , University of Bari Aldo Moro, Italy

Diversity and Inclusion Chairs

Mona Diab , Facebook AI Research & GWU, USA
Martha Yifiru Tachbelie , Addis Abada University, Ethiopia

Ethic Advisor Committee

Su Lin Blodgett , Microsoft Research Montréal, Canada
Christiane Fellbaum , Princeton University, USA

Student Volunteer Coordinators

Filip Klubička , ADAPT Centre, Ireland
Pintu Lohar , ADAPT Centre, Ireland
Vasudevan Nedumpozhimana , ADAPT Centre, Ireland
Guodong Xie , ADAPT Centre, Ireland

Internal Communications Chair

Marcely Boito Zanon , University of Avignon, France

Best Paper Selection Committee

Tim Baldwin , MBZUAI and The University of Melbourne, Australia
Kathleen McKeown , Columbia University, USA and Amazon AWS AI Labs
David Chiang , University of Notre Dame, USA
Min-Yen Kan , National University of Singapore, Singapore
Taro Watanabe , Nara Institute of Science and Technology, Japan

Guest ARR Editors-in-Chief for ACL 2022

Amanda Stent , Colby College, USA
Goran Glavaš , University of Mannheim, USA

Guest ARR Chief Technical Officer for ACL 2022

Graham Neubig , Carnegie Mellon University, USA

Guest ARR Editorial Manager for ACL 2022

Sara Goggi , CNR-ILC, Italy

Guest ARR Tech Team for ACL 2022

Yoshitomo Matsubara , UC Irvine, USA
Sebastin Santy , University of Washington, USA

Guest OpenReview Team for ACL 2022

Nadia L'Bahy , OpenReview
Celeste Martinez Gomez , OpenReview
Melisa Bok , OpenReview

Underline

Sol Rosenberg , Underline
Jernej Masnec , Underline
Damira Mršić , Underline
Mateo Antonic , Underline
Luka Šimić , Underline

Conference Advisor

Priscilla Rasmussen , ACL

Conference Registration

Nicole Ballard , Yes Events
Terah Shaffer , Yes Events

Local PCO

Greg Carew , Abbey
Fiona McGillivray , Abbey
Laird Smith , Abbey

Program Committee

Program Chairs

Smaranda Muresan , Columbia University and Amazon AWS AI Labs
Preslav Nakov , Qatar Computing Research Institute, HBKU
Aline Villavicencio , University of Sheffield

Computational Social Science and Cultural Analytics

Tanmoy Chakraborty , Indraprastha Institute of Information Technology Delhi
David Jurgens , University of Michigan
Diyi Yang , Georgia Institute of Technology

Dialogue and Interactive Systems

Srinivas Bangalore , Interactions LLC
Yun-Nung Chen , National Taiwan University
David Traum , University of Southern California
Dilek Hakkani-Tur , Amazon Alexa AI
Zhou Yu , Columbia University

Discourse and Pragmatics

Manfred Stede , Universität Potsdam
Junyi Jessy Li , University of Texas, Austin

Ethics and NLP

Saif M. Mohammad , National Research Council Canada
Malvina Nissim , University of Groningen

Generation

Claire Gardent , CNRS
Asli Celikyilmaz , Facebook AI Research
Chenghua Lin , University of Sheffield
Michael Elhadad , Ben Gurion University of the Negev

Information Extraction

Heng Ji , University of Illinois, Urbana-Champaign and Amazon Alexa AI
Marius Pasca , Google Research
Alan Ritter , Georgia Institute of Technology
Veselin Stoyanov , Facebook
Satoshi Sekine , RIKEN

Information Retrieval and Text Mining

Hang Li , Bytedance Technology
Marti Hearst , University of California Berkeley
Jing Jiang , Singapore Management University

Interpretability and Analysis of Models for NLP

Yonatan Belinkov , Technion, Technion
Anders Søgaard , Copenhagen University
Anna Rogers , University of Copenhagen
Hassan Sajjad , Qatar Computing Research Institute, HBKU

Language Grounding to Vision, Robotics and Beyond

William Yang Wang , UC Santa Barbara
Marie-Francine Moens , KU Leuven

Linguistic theories, Cognitive Modeling and Psycholinguistics

Frank Keller , The University of Edinburgh
Afra Alishahi , Tilburg University

Machine Learning for NLP

Mohit Bansal , University of North Carolina at Chapel Hill and Amazon Alexa AI
Nikolaos Aletras , University of Sheffield, University of Sheffield and Amazon
Andre Martins , Instituto Superior Técnico and Unbabel
Andreas Vlachos , Facebook and University of Cambridge
Kristina Toutanova , Google
Shafiq Joty , SalesForce and Nanyang Technological University

Machine Translation and Multilinguality

Taro Watanabe , Nara Institute of Science and Technology
Rico Sennrich , University of Zurich and University of Edinburgh
Francisco Guzmán , Facebook
Philipp Koehn , Facebook and Johns Hopkins University
Kenneth Heafield , The University of Edinburgh
Thamar Solorio , University of Houston

NLP Applications

Joel R. Tetreault , Dataminr
Karin Verspoor , Royal Melbourne Institute of Technology
Jimmy Lin , University of Waterloo
Horacio Saggion , Universitat Pompeu Fabra
Wei Gao , Singapore Management University
Beata Beigman Klebanov , Educational Testing Service

Phonology, Morphology and Word Segmentation

Ryan D Cotterell , ETH Zü"rich
Alexis Palmer , University of Colorado, Boulder

Question Answering

Mohit Iyyer , University of Massachusetts Amherst

Sanda Harabagiu , University of Texas at Dallas
Alessandro Moschitti , Amazon Alexa AI

Resources and Evaluation

Torsten Zesch , University of Duisburg-Essen
Agata Savary , Université Paris-Saclay
Katrin Erk , University of Texas, Austin
Pablo Gamallo , Universidad de Santiago de Compostela
Bonnie L. Webber , The University of Edinburgh

Semantics: Lexical

Carlos Ramisch , Aix Marseille University
Ekaterina Shutova , University of Amsterdam
Ivan Vulić , University of Cambridge and PolyAI Limited

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

Samuel R. Bowman , New York University
Goran Glavaš , University of Mannheim
Valeria de Paiva , Topos Institute
Renata Vieira , Universidade de Evora
Wei Lu , Singapore University of Technology and Design

Sentiment Analysis, Stylistic Analysis, and Argument Mining

Yulan He , The university of Warwick
Iryna Gurevych , TU Darmstadt
Roman Klinger , University of Stuttgart
Bing Liu , University of Illinois at Chicago

Special Theme

Emily M. Bender , University of Washington
laurent besacier , Naver Labs Europe
Steven Bird , Charles Darwin University and International Computer Science Institute

Speech and Multimodality

Grzegorz Chrupała , Tilburg University
Yang Liu , Amazon Alexa AI

Summarization

Kathleen McKeown , Columbia University and Amazon AWS AI Labs
Annie Louis , Google Research
Dragomir Radev , Yale University

Syntax: Tagging, Chunking and Parsing

Barbara Plank , IT University of Copenhagen

Joakim Nivre , Uppsala University

Action Editors

Zeljko Agic, Alan Akbik, Md Shad Akhtar, Firoj Alam, Nikolaos Aletras, Malihe Alikhani, Tanel Alumäe, Sophia Ananiadou, Antonios Anastasopoulos, Mark Anderson, Jacob Andreas, Xiang Ao, Marianna Apidianaki, Yuki Arase, Mikel Artetxe, Ehsaneddin Asgari, Giuseppe Attardi

Niranjan Balasubramanian, Timothy Baldwin, Miguel Ballesteros, David Bamman, Mohamad Hardyman Barawi, Jeremy Barnes, Loic Barrault, Roberto Basili, Ali Basirat, Jasmijn Bastings, Daniel Beck, Iz Beltagy, Luciana Benotti, Steven Bethard, Chandra Bhagavatula, Lidong Bing, Alexandra Birch, Steven Bird, Yonatan Bisk, Eduardo Blanco, Danushka Bollegala, Antoine Bosselut, Florian Boudin, Leonid Boytsov, Chloé Braud, Chris Brew, Wray Buntine

Elena Cabrio, Aoife Cahill, Andrew Caines, Ruken Cakici, Marie Candito, Yanan Cao, Ziqiang Cao, Cornelia Caragea, Xavier Carreras, Paula Carvalho, Andrew Cattle, Daniel Cer, Alessandra Cervone, Tanmoy Chakraborty, Muthu Kumar Chandrasekaran, Angel X Chang, Kai-Wei Chang, Snigdha Chaturvedi, Boxing Chen, Danqi Chen, Yun-Nung Chen, Kuan-Yu Chen, Kehai Chen, Lei Chen, Colin Cherry, Jackie CK Cheung, Hai Leong Chieu, Luis Chiruzzo, Jinho D. Choi, Monojit Choudhury, Khalid Choukri, Grzegorz Chrupała, Oana Cocarascu, Trevor Cohn, John M Conroy, Mathieu Constant, Caio Filippo Corro, Marta Ruiz Costa-jussà, Stefano Cresci, Aron Culotta

Giovanni Da San Martino, Raj Dabre, Walter Daelemans, Daniel Dakota, Dipanjan Das, Johannes Daxenberger, Gaël De Chalendar, Miryam De Lhoneux, Pascal Denis, Leon Derczynski, Barry Devereux, Mona T. Diab, Liang Ding, Georgiana Dinu, Jesse Dodge, Li Dong, Ruihai Dong, Yue Dong, Eduard Dragut, Kevin Duh, Nadir Durrani, Greg Durrett

Liat Ein-Dor, Michael Elhadad, Katrin Erk, Allyson Ettinger

Angela Fan, Anna Feldman, Naomi Feldman, Yang Feng, Yansong Feng, Raquel Fernández, Francis Ferraro, Elisabetta Fersini, Simone Filice, Mark Fishel, Annemarie Friedrich, Pascale Fung

Michel Galley, Matthias Gallé, Zhe Gan, Yang Gao, Marcos Garcia, Sebastian Gehrmann, Alborz Geramifard, Debanjan Ghosh, Goran Glavaš, Kyle Gorman, Jiatao Gu, Qing Gu, Qipeng Guo, Honglei Guo, Francisco Guzmán

Ivan Habernal, Christian Hardmeier, David Harwath, Zhongjun He, Yulan He, Luheng He, Daniel Hershcovich, Julia Hockenmaier, Enamul Hoque, Junjie Hu, Baotian Hu, Xuanjing Huang, Shujian Huang

Dmitry Ilvovsky, Kentaro Inui, Ozan Irsoy, Srini Iyer, Mohit Iyyer

Cassandra L Jacobs, Alon Jacovi, Kokil Jaidka, Hyeju Jang, Yangfeng Ji, Antonio Jimeno Yepes, Shafiq Joty, Preethi Jyothi

Sarvnaz Karimi, Shubhra Kanti Karmaker, Daisuke Kawahara, Daniel Khashabi, Seokhwan Kim, Taeuk Kim, Jin-Dong Kim, Judith Lynn Klavans, Roman Klinger, Hayato Kobayashi, Ekaterina Kochmar, Mamoru Komachi, Grzegorz Kondrak, Parisa Kordjamshidi, Amrith Krishna, Udo Kruschwitz, Marco Kuhlmann, Sumeet Kumar, Jonathan K Kummerfeld

Wai Lam, Zhenzhong Lan, Mark Last, Hady W. Lauw, John Lawrence, Carolin Lawrence, Alessandro Lenci, Lori Levin, Omer Levy, Mike Lewis, Piji Li, Sujian Li, Junyi Jessy Li, Liangyou Li, Wenjie Li, Tianrui Li, Zongxi Li, Juntao Li, Jing Li, Junhui Li, Constantine Lignos, Dekang Lin,

Chenghua Lin, Marco Lippi, Wu Liu, Xuebo Liu, Qun Liu, Yang Liu, Zhiyuan Liu, Pengfei Liu, Yang Liu, Kyle Lo, Wei Lu, Thang Luong, Anh Tuan Luu

Wilson Ma, Craig MacDonald, Nitin Madnani, Andrea Madotto, Navonil Majumder, Prodromos Malakasiotis, Igor Malioutov, Thomas Mandl, Vukosi Marivate, Eugenio Martinez-Camara, Bruno Martins, Yuji Matsumoto, Mausam, David McClosky, Mahnoosh Mehrabani, Ivan Vladimir Meza Ruiz, Margot Mieskes, Makoto Miwa, Daichi Mochihashi, Saif M. Mohammad, Mohamed Morchid, David R Mortensen, Alessandro Moschitti, Lili Mou, Philippe Muller, Kenton Murray

Nona Naderi, Courtney Napoles, Shashi Narayan, Franco Maria Nardini, Tristan Naumann, Mark-Jan Nederhof, Vincent Ng, Dat Quoc Nguyen, Thien Huu Nguyen, Jan Niehues, Qiang Ning

Diarmuid O Seaghdha, Brendan O'Connor, Jose Ochoa-Luna, Kemal Oflazer, Maciej Ogrodniczuk, Alice Oh, Naoaki Okazaki, Manabu Okumura, Matan Orbach, Miles Osborne, Jessica Ouyang

Hamid Palangi, Ankur P Parikh, Joonsuk Park, Seong-Bae Park, Yannick Parmentier, Tommaso Pasini, Rebecca J. Passonneau, Viviana Patti, Haoruo Peng, Nanyun Peng, Gabriele Pergola, Fabio Petroni, Maxime Peyrard, Juan Pino, Emily Pitler, Edoardo Ponti, Simone Paolo Ponzetto, Kashyap Popat, Maja Popovic, Soujanya Poria, Vinodkumar Prabhakaran, Daniel Preotiuc-Pietro, Emily Prud'hommeaux

Tieyun Qian, Xipeng Qiu, Xiaojun Quan

Colin Raffel, Ganesh Ramakrishnan, Siva Reddy, Ines Rehbein, Roi Reichart, Xiang Ren, Yafeng Ren, Sebastian Riedel, Sara Rosenthal, Joseph Le Roux, Alla Rozovskaya, Attapol Rutherford

Mrinmaya Sachan, Benoît Sagot, Hassan Sajjad, Chinnadhurai Sankar, Maarten Sap, Nathan Schneider, Hinrich Schuetze, H. Schwartz, Lane Schwartz, Rico Sennrich, Minjoon Seo, Bei Shi, Tianze Shi, Lei Shu, Melanie Siegel, Kevin Small, Noah Smith, Luca Soldaini, Vivek Srikumar, Shashank Srivastava, Efstathios Stamatatos, Gabriel Stanovsky, Pontus Stenetorp, Amanda Stent, Veselin Stoyanov, Karl Stratos, Emma Strubell, Sara Stymne, Jinsong Su, Yu Su, Saku Sugawara, Jun Suzuki

Dima Taji, Zeerak Talat, Duyu Tang, Amalia Todirascu, Antonio Toral, Paolo Torroni, Kristina Toutanova, Amine Trabelsi, Trang Tran, Chen-Tse Tsai, Junichi Tsujii, Kewei Tu

Stefan Ultes

Olga Vechtomova, Giulia Venturi, Suzan Verberne, Yannick Versley, David Vilares, Serena Villata, Thuy Vu, Ivan Vulić, Yogarshi Vyas

Byron C Wallace, Xiaojun Wan, Xin Eric Wang, Zhiguang Wang, Longyue Wang, Shuai Wang, Jingjing Wang, Leo Wanner, Taro Watanabe, Shinji Watanabe, Bonnie L. Webber, Zhongyu Wei, Michael White, Alina Wróblewska, Lijun Wu

Tong Xiao, Deyi Xiong, Wei Xu, Hainan Xu

Rui Yan, Min Yang, Jin-Ge Yao, Wenpeng Yin, Koichiro Yoshino, Dian Yu, Mo Yu, Jianfei Yu, Kai Yu, Tao Yu, François Yvon

Marcos Zampieri, Fabio Massimo Zanzotto, Luke Zettlemoyer, Yi Zhang, Yue Zhang, Weinan Zhang, Xiangliang Zhang, Zhe Zhang, Xingxing Zhang, Justine Zhang, Xiaoqing Zheng, Michael

Zock

ARR reviewers

Micheal Abaho, Ahmed Abdelali, Mostafa Abdou, Muhammad Abdul-Mageed, Omri Abend, Abdalghani Abujabal, Lasha Abzianidze, Manoj Acharya, Heike Adel, David Ifeoluwa Adelani, Somak Aditya, Vaibhav Adlakha, Stergos D. Afantenos, Sachin Agarwal, Vibhav Agarwal, Rodrigo Agerri, Manex Agirrezabal, Ameeta Agrawal, Priyanka Agrawal, Sweta Agrawal, Gustavo Aguilar, Roee Aharoni, Wasi Uddin Ahmad, Benyamin Ahmadnia, Aman Ahuja, Kabir Ahuja, Chaitanya Ahuja, Xi Ai, Laura Aina, Akiko Aizawa, Alan Akbik, Md Shad Akhtar, Nader Akoury, Ekin Akyürek, Ozge Alacam, Firoj Alam, Mehwish Alam, Chris Alberti, Georgios Alexandridis, David Alfter, Bashar Alhafni, Raquel G. Alhama, Tariq Alhindi, Hamed Alhoori, Hassan Alhuzali, Mohammad Aliannejadi, Afra Alishahi, Tamer Alkhouli, Emily Allaway, Miguel A. Alonso, Sawsan Alqahtani, Emily Alsentzer, Milad Alshomary, Christoph Alt, Tanel Alumäe, Fernando Alva-Manchego, Rami Aly, Maxime Amblard, Prithviraj Ammanabrolu, Reinald Kim Amplayo, Chantal Amrhein, Guozhen An, Aixiu An, Ashish Anand, Sophia Ananiadou, Raviteja Anantha, Antonios Anastasopoulos, Carolyn Jane Anderson, Nicholas Andrews, Ion Androutsopoulos, Gabor Angeli, Diego Antognini, Kaveri Anuranjana, Emilia Apostolova, Jun Araki, Rahul Aralikatte, Eiji Aramaki, Yuki Arase, Arturo Argueta, Mozhdeh Ariannezhad, Ignacio Arroyo-Fernández, Katya Artemova, Yoav Artzi, Masayuki Asahara, Akari Asai, Meysam Asgari, Elliott Ash, Zhenisbek Assylbekov, Duygu Ataman, Dennis Aumiller, Eleftherios Avramidis, Parul Awasthy, Hosein Azarbonyad, Wilker Aziz

Rohit Babbar, Sanghwan Bae, Ebrahim Bagheri, Dzmitry Bahdanau, Ashutosh Baheti, Fan Bai, He Bai, Yu Bai, JinYeong Bak, Vidhisha Balachandran, Mithun Balakrishna, Anusha Balakrishnan, Niranjan Balasubramanian, Ioana Baldini, Livio Baldini Soares, Kalika Bali, Nicolae Banari, Juan M Banda, Pratyay Banerjee, Sameer Bansal, Trapit Bansal, Yu Bao, Hangbo Bao, Forrest Sheng Bao, Junwei Bao, Siqi Bao, Jianzhu Bao, Zuyi Bao, Ankur Bapna, Roy Bar-Haim, Edoardo Barba, Francesco Barbieri, Denilson Barbosa, M Saiful Bari, Ken Barker, Gianni Barlacchi, Jeremy Barnes, Maria Barrett, Valentin Barriere, James Barry, Max Bartolo, Valerio Basile, Pierpaolo Basile, Somnath Basu Roy Chowdhury, John A. Bateman, Riza Batista-Navarro, Anil Batra, Khuyagbaatar Batsuren, Daniel Bauer, Timo Baumann, Rachel Bawden, Kathy Baxter, Tilman Beck, Lee Becker, Lisa Beinborn, Ahmad Beirami, Giannis Bekoulis, Núria Bel, Eric Bell, Gábor Bella, Meriem Beloucif, Iz Beltagy, Eyal Ben-David, Emily M. Bender, Michael Bendersky, Luisa Bentivogli, Adrian Benton, Jonathan Berant, Alexandre Berard, Gábor Berend, Taylor Berg-Kirkpatrick, Toms Bergmanis, Rafael Berlanga, Delphine Bernhard, Dario Bertero, Laurent Besacier, Chandra Bhagavatula, Rishabh Bhardwaj, Aditya Bhargava, Suma Bhat, Parminder Bhatia, Sumit Bhatia, Kasturi Bhattacharjee, Pushpak Bhattacharyya, Satwik Bhattamishra, Shruti Bhosale, Rajarshi Bhowmik, Bin Bi, Wei Bi, Federico Bianchi, Laura Biester, Yi Bin, Lidong Bing, Philippe Blache, Fred Blain, Eduardo Blanco, Terra Blevins, Rexhina Blloshmi, Jelke Bloem, Michael Bloodgood, Valts Blukis, Ben Bogin, Nikolay Bogoychev, Ondrej Bojar, Gemma Boleda, Danushka Bollegala, Marcel Bollmann, Valeriia Bolotova, Daniele Bonadiman, Francis Bond, Claudia Borg, Mihaela Bornea, Aurélien Bossard, Antoine Bosselut, Robert Bossy, Nadjet Bouayad-Agha, Florian Boudin, Zied Bouraoui, Samuel R. Bowman, Jordan Lee Boyd-Graber, Johan Boye, Kristy Elizabeth Boyer, Faeze Brahman, Arthur Brazinskas, Thomas Brochhagen, Samuel Broscheit, Thomas Brovelli, Christopher Bryant, Paweł Budzianowski, Emanuele Bugliarello, Wray Buntine, Joan Byamugisha, Bill Byrne

Sky CH-Wang, Subalalitha CN, Elena Cabrio, Avi Caciularu, Samuel Cahyawijaya, Deng Cai, Pengshan Cai, Yi Cai, Han Cai, Hengyi Cai, Jon Cai, Andrew Caines, Agostina Calabrese, Iacer Calixto, Jose Camacho-Collados, Erik Cambria, Oana-Maria Camburu, Giovanni Campagna, Leonardo Campillos-Llanos, Jon Ander Campos, Daniel F Campos, Marie Candito, Yixin Cao, Yue Cao,

Meng Cao, Steven Cao, Yu Cao, Qingxing Cao, Yuan Cao, Qingqing Cao, Juan Cao, Jie Cao, Yunbo Cao, Kris Cao, Ruisheng Cao, Annalina Caputo, Doina Caragea, Dallas Card, Ronald Cardenas, Rémi Cardon, Danilo Carvalho, Tommaso Caselli, Justine Cassell, Vittorio Castelli, Giuseppe Castellucci, Thiago Castro Ferreira, Paulo Cavalin, Christophe Cerisara, Alessandra Cervone, Arun Tejasvi Chaganty, Soumen Chakrabarti, Tuhin Chakrabarty, Abhisek Chakrabarty, Tanmoy Chakraborty, Bharathi Raja Chakravarthi, Ilias Chalkidis, Jon Chamberlain, Nathanael Chambers, Haw-Shiuan Chang, Jonathan P. Chang, Baobao Chang, Angel X Chang, Serina Chang, Xuankai Chang, WenHan Chao, Lidia S. Chao, Akshay Chaturvedi, Vishrav Chaudhary, Aditi Chaudhary, Wanxiang Che, Yubo Chen, Xiuying Chen, Danqi Chen, Yun Chen, Wenliang Chen, Muhao Chen, Minhua Chen, Wenhu Chen, Guanyi Chen, Yulong Chen, Jifan Chen, Yun-Nung Chen, Mingda Chen, Kehai Chen, Chung-Chi Chen, Yufeng Chen, Xiusi Chen, Qingcai Chen, Xinchi Chen, Long Chen, Yen-Chun Chen, Xilun Chen, Yue Chen, Meng Chen, Qian Chen, Bo Chen, Zhuang Chen, Hanjie Chen, Mei-Hua Chen, Pinzhen Chen, Zhumin Chen, Huimin Chen, Howard Chen, Lei Chen, Maximillian Chen, Lin Chen, John Chen, Tao Chen, Wei-Fan Chen, Shizhan Chen, Zhi Chen, Zhihong Chen, Wenqing Chen, Pei Chen, Jiaze Chen, Sanxing Chen, Luoxin Chen, Chenhua Chen, Wang Chen, Guanhua Chen, Lu Chen, Yunmo Chen, Daoyuan Chen, Bei Chen, Hongshen Chen, Qianglong Chen, Zhuohao Chen, Hao Cheng, Pengxiang Cheng, Lu Cheng, Yu Cheng, Pengyu Cheng, Liying Cheng, Weiwei Cheng, Yong Cheng, Fei Cheng, Minhao Cheng, Jianpeng Cheng, Emmanuele Chersoni, Zewen Chi, Ethan A Chi, Ta-Chung Chi, Yew Ken Chia, David Chiang, Ting-Rui Chiang, Patricia Chiril, Francisco Javier Chiyah-Garcia, Jaemin Cho, Won Ik Cho, Sangwoo Cho, Eleanor Chodroff, Eunsol Choi, Jinho D. Choi, Seungtaek Choi, Jaesik Choi, Shamil Chollampatt, Jaegul Choo, Leshem Choshen, Prafulla Kumar Choubey, Monojit Choudhury, Shammur Absar Chowdhury, Md Faisal Mahbub Chowdhury, Jishnu Ray Chowdhury, Christos Christodoulopoulos, Fenia Christopoulou, Alexandra Chronopoulou, Chenhui Chu, Zewei Chu, Christopher Chu, Tat-Seng Chua, Jin-Woo Chung, Yi-Ling Chung, Kenneth Church, Abu Nowshed Chy, Mark Cieliebak, Manuel Rafael Ciosici, Volkan Cirik, Kevin Clark, Christopher Clark, Elizabeth Clark, Miruna Clinciu, Louis Clouatre, Trevor Cohen, Jeremy R. Cole, Marcus D. Collins, Simone Conia, Mathieu Constant, Danish Contractor, Robin Cooper, Anna Corazza, Luciano Del Corro, Ryan D Cotterell, Josep Crego, Danilo Croce, Paul A. Crook, James Cross, Fermín L. Cruz, Heriberto Cuayahuitl, Yiming Cui, Leyang Cui, Shaobo Cui, Lei Cui, Washington Cunha, Anna Currey, Tonya Custis

Luis Fernando D'Haro, Jennifer D'Souza, Giovanni Da San Martino, Raj Dabre, Deborah A. Dahl, Xinyu Dai, Xiang Dai, Falcon Z Dai, Hongliang Dai, Wenliang Dai, Damai Dai, Yinpei Dai, Siddharth Dalmia, Sandipan Dandapat, Ankit Dangi, Marina Danilevsky, Verna Dankers, Anubrata Das, Rajarshi Das, Sarthak Dash, Pradeep Dasigi, Debajyoti Datta, Hal Daumé Iii, Sam Davidson, Ernest Davis, Brian Davis, Gaël De Chalendar, Christine De Kock, Kordula De Kuthy, Miryam De Lhoneux, Marie-Catherine De Marneffe, Gerard De Melo, José G. C. De Souza, Iria De-Dios-Flores, Steve DeNeefe, Alok Debnath, Mathieu Dehouck, Flor Miriam Plaza Del Arco, Marco Del Tredici, Agustín D. Delgado, Louise Deléger, David Demeter, Çağatay Demiralp, Yang Deng, Yuntian Deng, Zhongfen Deng, Tejaswini Deoskar, Jan Milan Deriu, Franck Dernoncourt, Tim Dettmers, Daniel Deutsch, Sunipa Dev, Joseph Dexter, Kuntal Dey, Bhuwan Dhingra, Luigi Di Caro, Barbara Di Eugenio, Shizhe Diao, Gaël Dias, Liang Ding, Shuoyang Ding, Chenchen Ding, Xiao Ding, Ning Ding, Kaize Ding, Haibo Ding, Stefanie Dipper, Nemanja Djuric, Ngoc Bich Do, Simon Dobnik, Jesse Dodge, Charles Dognin, Miguel Domingo, Lucia Donatelli, Domenic Donato, Li Dong, Qian Qian Dong, MeiXing Dong, Yue Dong, Bonaventure F. P. Dossou, Zi-Yi Dou, Longxu Dou, Doug Downey, A. Seza Doğruöz, Mark Dras, Markus Dreyer, Rotem Dror, Andrew Drozdov, Jinhua Du, Wanyu Du, Jingfei Du, Pan Du, Mengnan Du, Lan Du, Li Du, Xinya Du, Yupei Du, Xiangyu Duan, Junwen Duan, Kumar Avinava Dubey, Pablo Duboue, Philipp Dufter, Jonathan Dunn, Gérard M Dupont, Ondrej Dusek, Ritam Dutt, Subhabrata Dutta, Chris Dyer, Nouha Dziri, Hervé Déjean

Abteen Ebrahimi, Aleksandra Edwards, Steffen Eger, Markus Egg, Koji Eguchi, Yo Ehara, Vladimir

Eidelman, Bryan Eikema, Jacob Eisenstein, Asif Ekbal, Wassim El-Hajj, Aparna Elangovan, Yanai Elazar, Heba Elfardy, Michael Elhadad, AbdelRahim A. Elmadany, Micha Elsner, Denis Emelin, Guy Emerson, Akiko Eriguchi, Liana Ermakova, Patrick Ernst, Carlos Escolano, Arash Eshghi, Ramy Eskander, Cristina España-Bonet, Luis Espinosa-Anke, Kawin Ethayarajh, Allyson Ettinger, Kilian Evang, Ben Eyal

Alexander Fabbri, Marzieh Fadaee, Tiziano Fagni, Farzane Fakhrian, Neele Falk, Tobias Falke, Kai Fan, Feifan Fan, Chuang Fan, Lu Fan, Wei Fang, Yuwei Fang, Yimai Fang, Adam Faulkner, Maryam Fazel-Zarandi, Amir Feder, Hao Fei, Nils Feldhus, Naomi Feldman, Mariano Felice, Zhangyin Feng, Shi Feng, Shaoxiong Feng, Jiazhan Feng, Shi Feng, Xiachong Feng, Manos Fergadiotis, James Ferguson, Patrick Fernandes, Raquel Fernández, Daniel Fernández-González, Elisa Ferracane, Francis Ferraro, Besnik Fetahu, Oluwaseyi Feyisetan, Alejandro Figueroa, Simone Filice, Catherine Finegan-Dollak, Orhan Firat, Nicholas FitzGerald, Margaret M. Fleck, Lucie Flek, Antske Fokkens, Marina Fomicheva, José A.r. Fonollosa, Marco Fonseca, Tommaso Fornaciari, Paula Fortuna, Eric Fosler-Lussier, George Foster, Jennifer Foster, Mary Ellen Foster, Stella Frank, Anette Frank, Thomas François, Alexander Fraser, Kathleen C. Fraser, Marjorie Freedman, Markus Freitag, Dayne Freitag, Lea Frermann, Daniel Fried, Qiankun Fu, Guohong Fu, Tsu-Jui Fu, Jie Fu, Zuohui Fu, Peng Fu, Yoshinari Fujinuma, Atsushi Fujita, Kotaro Funakoshi, Adam Funk, Richard Futrell, Michael Färber

Devi G, Matteo Gabburo, Saadia Gabriel, David Gaddy, Marco Gaido, Andrea Galassi, Mark Gales, Boris Alexandrovich Galitsky, Ygor Gallina, Diana Galvan, Björn Gambäck, Zhe Gan, Yujian Gan, Leilei Gan, Kuzman Ganchev, Sudeep Gandhe, Balaji Ganesan, Rashmi Gangadharaiah, Varun Gangal, Revanth Gangi Reddy, Debasis Ganguly, Yifan Gao, Tianyu Gao, Yanjun Gao, Wei Gao, Jun Gao, Ge Gao, Yang Gao, Yingbo Gao, Shen Gao, Utpal Garain, Cristina Garbacea, Diego Garcia-Olano, Matt Gardner, Sarthak Garg, Siddhant Garg, Dan Garrette, Aina Garí Soler, Kiril Gashteovski, Albert Gatt, Manas Gaur, Eric Gaussier, Dipesh Gautam, Yubin Ge, Sebastian Gehrmann, Michaela Geierhos, Ruiying Geng, Shijie Geng, Xiubo Geng, Xinwei Geng, Ariel Gera, Mor Geva, Hamidreza Ghader, Demian Gholipour Ghalandari, Sarik Ghazarian, Mozhdeh Gheini, Deepanway Ghosal, Deepanway Ghosal, Sayan Ghosh, Debanjan Ghosh, Sourav Ghosh, Soumitra Ghosh, Daniel Gildea, Salvatore Giorgi, Voula Giouli, Adrià de Gispert, Mario Giulianelli, Michael Glass, Goran Glavaš, Alfio Gliozzo, Pranav Goel, Vaibhava Goel, Nazli Goharian, Tejas Gokhale, Elizaveta Goncharova, Hongyu Gong, Heng Gong, Karthik Gopalakrishnan, Philip John Gorinski, Matthew R. Gormley, Koustava Goswami, Akhilesh Deepak Gotmare, Isao Goto, Cyril Goutte, Edward Gow-Smith, Kartik Goyal, Pawan Goyal, Tanya Goyal, Naman Goyal, Mario Graff, Christophe Gravier, Yulia Grishina, Milan Gritta, Loïc Grobol, Dagmar Gromann, Roman Grundkiewicz, Jia-Chen Gu, Yue Gu, Jing Gu, Yi Guan, Saiping Guan, Jian Guan, Marco Guerini, Lin Gui, Tao Gui, Vincent Guigue, Liane Guillou, Camille Guinaudeau, Kalpa Gunaratna, Chulaka Gunasekara, Tunga Gungor, Junliang Guo, Yinpeng Guo, Yuhang Guo, Ruocheng Guo, Zhijiang Guo, Jiaqi Guo, Jiang Guo, Vivek Gupta, Nitish Gupta, Sonal Gupta, Prakhar Gupta, Arshit Gupta, Arpit Gupta, Shashank Gupta, Izzeddin Gur, Suchin Gururangan, Joakim Gustafson, Ximena Gutierrez-Vasques, Carlos Gómez-Rodríguez

Jung-Woo Ha, Nizar Habash, Ivan Habernal, Kais Haddar, Christian Hadiwinoto, Reza Haf, Michael Hahn, Zhen Hai, Huda Hakami, Dilek Hakkani-Tur, Kishaloy Halder, Xianpei Han, Wenjuan Han, Ting Han, Xiaochuang Han, Namgi Han, Rujun Han, Jiawei Han, Jiale Han, Abram Handler, Viktor Hangya, Greg Hanneman, Yaru Hao, Jie Hao, Momchil Hardalov, Mareike Hartmann, Thomas Hartvigsen, Sadid A. Hasan, Peter Hase, Chikara Hashimoto, Nabil Hathout, Robert D. Hawkins, Katsuhiko Hayashi, Yoshihiko Hayashi, Hiroaki Hayashi, Shirley Anugrah Hayati, Devamanyu Hazarika, Luheng He, Zhongjun He, Shizhu He, Junxian He, Jiangen He, Xuanli He, Tianxing He, Hangfeng He, Wanwei He, Keqing He, Liang He, Marti Hearst, Michael Heck, Behnam Hedayatnia, Benjamin Heinzerling, Matthew Henderson, Iris Hendrickx, Leonhard Hennig, Sophie Henning, Daniel Hershcovich, Jonathan Herzig, Jack Hessel, John Hewitt, Ryuichiro

Higashinaka, Swapnil Hingmire, Tsutomu Hirao, Tatsuya Hiraoka, Cuong Hoang, Hieu Hoang, Johannes Hoffart, Valentin Hofmann, Chris Hokamp, Eben Holderness, Nora Hollenstein, Ari Holtzman, Takeshi Homma, Ukyo Honda, Pengfei Hong, Mark Hopkins, Helmut Horacek, Md Mosharaf Hossain, Nabil Hossain, Mohammad Javad Hosseini, Yufang Hou, Lei Hou, Yutai Hou, Feng Hou, Dirk Hovy, David M Howcroft, Estevam Hruschka, Shu-Kai Hsieh, I-Hung Hsu, Chao-Chun Hsu, Chun-Nan Hsu, Wei-Ning Hsu, Phu Mon Htut, Huang Hu, Pengwei Hu, Baotian Hu, Jennifer Hu, Junjie Hu, Guangneng Hu, Minghao Hu, Zhe Hu, Yue Hu, Jinyi Hu, Ziniu Hu, Chi Hu, Po Hu, Wei Hu, Renfen Hu, Linmei Hu, Xinyu Hua, Yiqing Hua, Hen-Hsen Huang, Chieh-Yang Huang, Shujian Huang, Xuancheng Huang, Chenyang Huang, Lifu Huang, Zhongqiang Huang, Xinting Huang, Kuan-Hao Huang, Siyu Huang, Xiaolei Huang, Jie Huang, Fei Huang, Kung-Hsiang Huang, Chao-Wei Huang, Quzhe Huang, Ziming Huang, Jimmy Huang, Heyan Huang, Zhen Huang, Jing Huang, Jiaji Huang, Ruihong Huang, Haoran Huang, Danqing Huang, Minlie Huang, He Huang, Patrick Huber, Kai Hui, Binyuan Hui, Dieuwke Hupkes, Ben Hutchinson, Jena D. Hwang, Sung Ju Hwang, Ali Hürriyetoğlu

Ignacio Iacobacci, Georgiana Ifrim, Oana Ignat, Ryu Iida, Gabriel Ilharco, Filip Ilievski, Nikolai Ilinykh, Irina Illina, Dmitry Ilvovsky, Kenji Imamura, Oana Inel, Naoya Inoue, Radu Tudor Ionescu, Daphne Ippolito, Hitoshi Isahara, Tatsuya Ishigaki, Etsuko Ishii, Tunazzina Islam, Hayate Iso, Dan Iter, Itay Itzhak, Julia Ive, Tomoya Iwakura, Kenichi Iwatsuki, Srini Iyer, Rishabh K Iyer, Gautier Izacard

Aaron Jaech, Sarthak Jain, Masoud Jalili Sabet, Abhik Jana, Hyeju Jang, Tommi Jauhiainen, Sébastien Jean, Sungho Jeon, Minwoo Jeong, Yacine Jernite, Kevin Jesse, Rahul Jha, Harsh Jhamtani, Feng Ji, Zongcheng Ji, Robin Jia, Ruipeng Jia, Renee Jia, Chen Jia, Yuxiang Jia, Ping Jian, Xin Jiang, Zhengbao Jiang, Daxin Jiang, Meng Jiang, Nanjiang Jiang, Zhuoren Jiang, Yong Jiang, Ming Jiang, Nan Jiang, Haoming Jiang, Hongfei Jiang, Zhuolin Jiang, Zhuoxuan Jiang, Wenbin Jiang, Wenxiang Jiao, Zhanming Jie, Antonio Jimeno Yepes, Di Jin, Peng Jin, Zhijing Jin, Hanqi Jin, Hailong Jin, Lisa Jin, Xiaolong Jin, Baoyu Jing, Yohan Jo, Richard Johansson, Melvin Johnson, Gareth J. F. Jones, Erik Jones, Siddhartha Jonnalagadda, Aditya Joshi, Dhanya Jothimani, Shafiq Joty, Xincheng Ju, Jaap Jumelet, Heewoo Jun, David Jurgens, Prathyusha Jwalapuram

Jad Kabbara, Indika Kahanda, Sylvain Kahane, Ivana Kajic, Mihir Kale, Oren Kalinsky, Aikaterini-Lida Kalouli, Ehsan Kamalloo, Hidetaka Kamigaito, Jaap Kamps, Min-Yen Kan, Hiroshi Kanayama, Nikhil Kandpal, Masahiro Kaneko, Dongyeop Kang, Minki Kang, Diptesh Kanojia, Evangelos Kanoulas, Jiun-Yu Kao, Pavan Kapanipathi, Georgi Karadzhov, Alina Karakanta, Giannis Karamanolakis, Siddharth Karamcheti, Mladen Karan, Börje F. Karlsson, Sanjeev Kumar Karn, Jungo Kasai, Omid Kashefi, Yosuke Kashiwagi, Zdeněk Kasner, Nora Kassner, Denys Katerenchuk, Divyansh Kaushik, Pride Kavumba, Anna Kazantseva, Hideto Kazawa, Ashkan Kazemi, Abe Kazemzadeh, Pei Ke, Zixuan Ke, Chris Kedzie, Katherine A. Keith, Yova Kementchedjhieva, Brendan Kennedy, Casey Kennington, Tom Kenter, Daniel J Kershaw, Santosh Kesiraju, Salam Khalifa, Urvashi Khandelwal, Dinesh Khandelwal, Simran Khanuja, Mitesh M Khapra, Eugene Kharitonov, Daniel Khashabi, Mikhail Khodak, Tushar Khot, Johannes Kiesel, Halil Kilicoglu, Yunsu Kim, Seokhwan Kim, Dongkwan Kim, Byeongchang Kim, Jin-Dong Kim, Gunhee Kim, Dong-Jin Kim, Jung-jae Kim, Yoon Kim, Jooyeon Kim, Juyong Kim, Hyunwoo Kim, Gyuwan Kim, Hyounghun Kim, Jihyuk Kim, Kang-Min Kim, Gene Louis Kim, Joo-Kyung Kim, Yeachan Kim, Geonmin Kim, Doo Soon Kim, Tracy Holloway King, Milton King, Christo Kirov, Nikita Kitaev, Hirokazu Kiyomaru, Shun Kiyono, Judith Lynn Klavans, Ayal Klein, Bennett Kleinberg, Jan-Christoph Klie, Mateusz Klimaszewski, Miyoung Ko, Sosuke Kobayashi, Hideo Kobayashi, Thomas H Kober, Jordan Kodner, Svetla Peneva Koeva, Mare Koit, Noriyuki Kojima, Alexander Koller, Keshav Kolluru, Mamoru Komachi, Rik Koncel-Kedziorski, Luyang Kong, Lingkai Kong, Valia Kordoni, Yuta Koreeda, Mandy Barrett Korpusik, Katsunori Kotani, Lili Kotlerman, Fajri Koto, Venelin Kovatchev, Josip Krapac, Sebastian Krause, Elisa Kreiss, Ralf Krestel, Julia Kreutzer, Florian L. Kreyssig, Kalpesh Krishna, Nikhil Krishnaswamy, Reno Kriz, Canasai Kru-

engkrai, Udo Kruschwitz, Germàn Kruszewski, Alexander Ku, Lun-Wei Ku, Marco Kuhlmann, Mayank Kulkarni, Sayali Kulkarni, Vivek Kulkarni, Artur Kulmizev, Devang Kulshreshtha, Shankar Kumar, Vishwajeet Kumar, Sawan Kumar, Varun Kumar, Ashutosh Kumar, Dhruv Kumar, Sachin Kumar, Anoop Kunchukuttan, Souvik Kundu, Shuhei Kurita, Kemal Kurniawan, Sadao Kurohashi, Robin Kurtz, Andrey Kutuzov

Peifeng LI, Matthieu Labeau, Faisal Ladhak, Nikolaos Lagos, Yuxuan Lai, Cheng-I Lai, Viet Dac Lai, Yash Kumar Lal, Divesh Lala, John P. Lalor, Wai Lam, Tsz Kin Lam, Matthew Lamm, Vasileios Lampos, Gerasimos Lampouras, Yunshi Lan, Yanyan Lan, Man Lan, Lukas Lange, Ni Lao, Guy Lapalme, Egoitz Laparra, Mirella Lapata, Gabriella Lapesa, Ekaterina Lapshinova-Koltunski, Stefan Larson, Jey Han Lau, Anne Lauscher, Alberto Lavelli, John Lawrence, Dawn Lawrie, Phong Le, Hung Le, Nayeon Lee, Taesung Lee, Hwanhee Lee, Katherine Lee, Jinhyuk Lee, Hwaran Lee, Mina Lee, Moontae Lee, I-Ta Lee, Dongkyu Lee, Young-Suk Lee, Ji-Ung Lee, Andrew Lee, Seolhwa Lee, Fei-Tzin Lee, Kenton Lee, Kyungjae Lee, Hyunju Lee, Sang-Woo Lee, Roy Ka-Wei Lee, Hung-yi Lee, Artuur Leeuwenberg, Els Lefever, Tao Lei, Wenqiang Lei, Zeyang Lei, Jie Lei, Jochen L. Leidner, Alessandro Lenci, Yichong Leng, Haley Lepp, Piyawat Lertvittayakumjorn, Guy Lev, Lori Levin, Gina-Anne Levow, Bo Li, Yuan-Fang Li, Haoran Li, Junyi Jessy Li, Dongfang Li, Binyang Li, Zhenghua Li, Jicheng Li, Yaoyiran Li, Jing Li, Piji Li, Chenliang Li, Tao Li, Yitong Li, Zuchao Li, Xin Li, Shaohua Li, Yingya Li, Juncheng B Li, Chunyuan Li, Lei Li, Yaliang Li, Juanzi Li, Yitong Li, Manling Li, Xiang Lisa Li, Sheng Li, Xiujun Li, Lucy Li, Huayang Li, Shuangyin Li, Jing Li, Yingjie Li, Peng Li, Fei Li, Dingcheng Li, Xintong Li, Jingye Li, Chen Li, Zhongyang Li, Xiangci Li, Zhoujun Li, Bai Li, Jialu Li, Miao Li, Bei Li, Xinjian Li, Yinqiao Li, Dianqi Li, Chen Li, Yanzeng Li, Wei Li, Yanran Li, Haizhou Li, Baoli Li, Yiyuan Li, Xiang Li, Chenliang Li, Zichao Li, Bowen Li, Xiang Lorraine Li, Quanzhi Li, Ruizhe Li, Qi Li, Hongyu Li, Lin Li, Zhongli Li, Si Li, Xiaonan Li, Irene Li, Zongxi Li, Minglei Li, Jinchao Li, Shang-Wen Li, Paul Pu Liang, Yunlong Liang, Bin Liang, Chen Liang, Xiaobo Liang, Chao-Chun Liang, Lizi Liao, Jindřich Libovický, Chaya Liebeskind, Wang Lijie, Kwan Hui Lim, Gilbert Lim, Ying Lin, Bill Yuchen Lin, Weizhe Lin, Xi Victoria Lin, Zhaojiang Lin, Lucy H. Lin, Zi Lin, Zehao Lin, Yankai Lin, Peiqin Lin, Zhouhan Lin, Junyang Lin, Xiang Lin, Ting-En Lin, Chu-Cheng Lin, Hongfei Lin, Zheng Lin, Chih-Jen Lin, Zhen-Hua Ling, Tal Linzen, Pierre Lison, Johann-Mattis List, Robert Litschko, Patrick William Littell, Marina Litvak, Shujie Liu, Fei Liu, Han Liu, Xinchen Liu, Linqing Liu, Ximeng Liu, Nelson F. Liu, Danni Liu, Zhiyuan Liu, Hui Liu, Xiao Liu, Dairui Liu, Qian Liu, Xuebo Liu, Qianchu Liu, Haokun Liu, Zoey Liu, Fangyu Liu, Ming Liu, Yong Liu, Zemin Liu, Kang Liu, Ling Liu, Dexi Liu, Jiangming Liu, Bin Liu, Xianggen Liu, Haoyan Liu, Liyuan Liu, Xueqing Liu, Junhao Liu, Peng Liu, Yijia Liu, Feifan Liu, Yang Janet Liu, Xudong Liu, Haochen Liu, Bing Liu, Shulin Liu, Siyang Liu, Chi-Liang Liu, Yijin Liu, Jing Liu, Zhengyuan Liu, Jiachang Liu, Lemao Liu, Jian Liu, Weijie Liu, Xiao Liu, Zhun Liu, Xin Liu, Maofu Liu, Zihan Liu, Yingchi Liu, Chen Cecilia Liu, Tianyu Liu, Farhana Ferdousi Liza, Nikola Ljubešić, Kyle Lo, Guodong Long, Yunfei Long, Lucelene Lopes, José David Lopes, Natalia Loukachevitch, Ismini Lourentzou, Sharid Loáiciga, Di Lu, Wei Lu, Yaojie Lu, Yichao Lu, Yao Lu, Weiming Lu, Nurul Lubis, Stephanie M. Lukin, Renqian Luo, Ruotian Luo, Ling Luo, Tianyi Luo, Huaishao Luo, Hongyin Luo, Ping Luo, Anh Tuan Luu, Kelvin Luu, Shangwen Lv, Xin Lv, Teresa Lynn, Michael Lyu, Chenyang Lyu, Samuel Läubli

Jianqiang Ma, Mingyu Derek Ma, Xuezhe Ma, Xinyin Ma, Qianli Ma, Xiaofei Ma, Ji Ma, Xutai Ma, Kaixin Ma, Sean MacAvaney, Aman Madaan, Avinash Madasu, Mounica Maddela, Pranava Madhyastha, Andrea Madotto, Walid Magdy, Manuel Mager, Suchismit Mahapatra, Adyasha Maharana, Debanjan Mahata, Rahmad Mahendra, Ayush Maheshwari, Gaurav Maheshwari, Kyle Mahowald, Wolfgang Maier, Jean Maillard, Olga Majewska, Bodhisattwa Prasad Majumder, Márton Makrai, Prodromos Malakasiotis, Chaitanya Malaviya, Andreas Maletti, Ankur Mali, Eric Malmi, Christopher Malon, Radhika Mamidi, Saab Mansour, Ramesh Manuvinakurike, Emaad Manzoor, Xian-Ling Mao, Yuren Mao, Wenji Mao, Jiaxin Mao, Xin Mao, Vladislav Maraev, Ana Marasovic, Diego Marcheggiani, Daniel Marcu, Piotr Mardziel, Andreas Marfurt, Katerina Mar-

gatina, Benjamin Marie, Zita Marinho, Antonis Maronikolakis, Edison Marrese-Taylor, Hector Martinez Alonso, Pedro Henrique Martins, Yuval Marton, Sameen Maruf, Claudia Marzi, Sandeep Mathias, Prashant Mathur, Puneet Mathur, Sérgio Matos, David Martins De Matos, Yuichiroh Matsubayashi, Takuya Matsuzaki, Yevgen Matusevych, Evgeny Matusov, Kaushal Kumar Maurya, Nickil Maveli, Jonathan May, Stephen Mayhew, Karen Mazidi, Sahisnu Mazumder, Arya D. McCarthy, John Philip McCrae, Matthew B.a. McDermott, Denis Jered McInerney, Alexander Mehler, Shikib Mehri, Nikhil Mehta, Hongyuan Mei, Hardik Meisheri, Clara Isabel Meister, Dheeraj Mekala, Telmo Menezes, Zhao Meng, Rui Meng, Yu Meng, Fandong Meng, Yuanliang Meng, Zaiqiao Meng, Tao Meng, Rakesh R Menon, Samuel Mensah, Wolfgang Menzel, Paola Merlo, William Merrill, Mohsen Mesgar, Florian Metze, Donald Metzler, Marie-Jean Meurs, Haitao Mi, Yisong Miao, Julian Michael, Paul Michel, Lesly Miculicich, Sabrina J Mielke, Margot Mieskes, Todor Mihaylov, Tsvetomila Mihaylova, Elena Mikhalkova, Simon Mille, Tristan Miller, Timothy A Miller, Eleni Miltsakaki, David Mimno, Sewon Min, Bonan Min, Pasquale Minervini, Xu Mingzhou, Hideya Mino, Shachar Mirkin, Seyedabolghasem Mirroshandel, Paramita Mirza, Abhijit Mishra, Swaroop Mishra, Kanishka Misra, Masato Mita, Prasenjit Mitra, Jelena Mitrović, Arpit Mittal, Vibhu O. Mittal, Makoto Miwa, Yusuke Miyao, Takashi Miyazaki, Daichi Mochihashi, Ashutosh Modi, Hans Moen, Aditya Mogadala, Nikita Moghe, Alireza Mohammadshahi, Muqeeth Mohammed, Hosein Mohebbi, Diego Molla, Natawut Monaikul, Nicholas Monath, Ishani Mondal, Joel Ruben Antony Moniz, Syrielle Montariol, Manuel Montes, Seungwhan Moon, Ray Mooney, Nafise Sadat Moosavi, Mehrad Moradshahi, Vlad I Morariu, Erwan Moreau, Jose G Moreno, Mathieu Morey, Gaku Morio, Makoto Morishita, John Xavier Morris, David R Mortensen, Ahmadreza Mosallanezhad, Marius Mosbach, Xiangyang Mou, Lili Mou, Seyed Mahed Mousavi, Maximilian Mozes, Yassine Mrabet, Frank Martin Mtumbuka, Hamdy Mubarak, Pramod Kaushik Mudrakarta, David Mueller, Aaron Mueller, Matteo Muffo, Animesh Mukherjee, Phoebe Mulcaire, Matthew Mulholland, Deepak Muralidharan, Masayasu Muraoka, Elena Musi, Sheshera Mysore, Mathias Müller, Thomas Müller, Mark-Christoph Müller

Seung-Hoon Na, Nona Naderi, Masaaki Nagata, Ajay Nagesh, Saeed Najafi, Tetsuji Nakagawa, Diane Napolitano, Jason Naradowsky, Karthik R Narasimhan, Tahira Naseem, Sudip Kumar Naskar, Alexis Nasr, Vivi Nastase, Anandhavelu Natarajan, Tristan Naumann, Roberto Navigli, Matteo Negri, Graham Neubig, Günter Neumann, Mariana Neves, Denis Newman-Griffis, Thien Huu Nguyen, Hoang Van Nguyen, Dai Quoc Nguyen, Truc-Vien T. Nguyen, Huyen Nguyen, Thanh V Nguyen, Thanh-Tung Nguyen, Hoang-Quoc Nguyen-Son, Jianmo Ni, Garrett Nicolai, Massimo Nicosia, Vlad Niculae, Feng Nie, Yixin Nie, Jan Niehues, Christina Niklaus, Fedor Nikolaev, Giannis Nikolentzos, Vassilina Nikoulina, Qiang Ning, Takashi Ninomiya, Nobal B. Niraula, Kosuke Nishida, Noriki Nishida, Kyosuke Nishida, Masaaki Nishino, Sergiu Nisioi, Tong Niu, Xing Niu, Guanglin Niu, Hiroshi Noji, Tadashi Nomoto, Damien Nouvel, Michal Novák, Pierre Nugues, Claire Nédellec, Aurélie Névéol

Alexander O'Connor, Yusuke Oda, Stephan Oepen, Maciej Ogrodniczuk, Barlas Oguz, Alice Oh, Yoo Rhee Oh, Kiyonori Ohtake, Naoaki Okazaki, Tsuyoshi Okita, Manabu Okumura, Hugo Gonçalo Oliveira, Antoni Oliver, Arturo Oncevay, Yasumasa Onoe, Juri Opitz, Shereen Oraby, John Ortega, Pedro Ortiz Suarez, Yohei Oseki, Malte Ostendorff, Naoki Otani, Myle Ott, Zhijian Ou, Zijing Ou, Hiroki Ouchi, Nedjma Ousidhoum

Maria Leonor Pacheco, Inkit Padhi, Aishwarya Padmakumar, Sukomal Pal, Santanu Pal, Chester Palen-Michel, Alexis Palmer, Endang Wahyu Pamungkas, Liangming Pan, Boyuan Pan, Richard Yuanzhe Pang, Liang Pang, Sheena Panthaplackel, Alexandros Papangelis, Nikolaos Pappas, Emerson Cabrera Paraiso, Letitia Parcalabescu, Natalie Parde, Antonio Pareja-Lora, Cecile Paris, Lucy Park, Youngja Park, ChaeHun Park, Hyunji Hayley Park, Jungsoo Park, Kunwoo Park, Chanjun Park, Ioannis Partalas, Niko Tapio Partanen, Prasanna Parthasarathi, Md Rizwan Parvez, Gabriella Pasi, Tommaso Pasini, Ramakanth Pasunuru, Or Patashnik, Kevin Patel, Raj Patel, Arkil Patel, Roma Patel, Sangameshwar Patil, Braja Patra, Barun Patra, Jasabanta Patro, Siddharth Patwardhan,

Manasi Patwardhan, Debjit Paul, Silviu Paun, John Pavlopoulos, Pavel Pecina, Jiaxin Pei, Stephan Peitz, Viktor Pekar, Hao Peng, Baolin Peng, Haoruo Peng, Xutan Peng, Yifan Peng, Xi Peng, Wei Peng, Siyao Peng, Lis Pereira, Martin Pereira, Julien Perez, Gabriele Pergola, Jan-Thorsten Peter, Ben Peters, Matthew E Peters, Pavel Petrushkov, Sandro Pezzelle, Jonas Pfeiffer, Quan Pham, Minh-Quang Pham, Van-Thuy Phi, Maciej Piasecki, Massimo Piccardi, Karl Pichotta, Mohammad Taher Pilehvar, Tiago Pimentel, Aidan Pine, Juan Pino, Yuval Pinter, Flammie A Pirinen, Benjamin Piwowarski, Lonneke Van Der Plas, Bryan A. Plummer, Brian Plüss, Sylvain Pogodalla, Martin Popel, Octavian Popescu, Andrei Popescu-Belis, Fred Popowich, François Portet, Matt Post, Martin Potthast, Christopher Potts, Amir Pouran Ben Veyseh, Vinodkumar Prabhakaran, Sandhya Prabhakaran, Shrimai Prabhumoye, Aniket Pramanick, Jakob Prange, Animesh Prasad, Archiki Prasad, Judita Preiss, Audi Primadhanty, Victor Prokhorov, Prokopis Prokopidis, Haritz Puerto, Rajkumar Pujari, Matthew Purver, Valentina Pyatkin, Juan Antonio Pérez-Ortiz

Tao Qi, Fanchao Qi, Jianzhong Qi, Peng Qi, Kun Qian, Dong Qian, Yujie Qian, Libo Qin, Yujia Qin, Xipeng Qiu, Long Qiu, Liang Qiu, Lizhen Qu, Xiaoye Qu, Chen Qu

Ella Rabinovich, Gorjan Radevski, Alessandro Raganato, Dinesh Raghu, Vipul Raheja, Afshin Rahimi, Hossein Rajaby Faghihi, Sara Rajaee, Dheeraj Rajagopal, Sanguthevar Rajasekaran, Pavithra Rajendran, Geetanjali Rakshit, Dhananjay Ram, Ori Ram, Taraka Rama, Deepak Ramachandran, Anil Ramakrishna, Ganesh Ramakrishnan, Owen Rambow, Alan Ramponi, Gabriela Ramírez De La Rosa, Tharindu Ranasinghe, Surangika Ranathunga, Priya Rani, Peter A. Rankel, Jinfeng Rao, Yanghui Rao, Ahmad Rashid, Hannah Rashkin, Abhinav Rastogi, Vipul Kumar Rathore, Vikas Raunak, Shauli Ravfogel, Abhilasha Ravichander, Vinit Ravishankar, Anirudh Ravula, Avik Ray, Soumya Ray, Manny Rayner, Julia Rayz, Traian Rebedea, Sravana Reddy, Hanumant Harichandra Redkar, Georg Rehm, Marek Rei, Nils Reimers, Navid Rekabsaz, Xuancheng Ren, Xiang Ren, Shuo Ren, Da Ren, Zhaochun Ren, Shuhuai Ren, Ruiyang Ren, Pengjie Ren, Feiliang Ren, Feiliang Ren, Adi Renduchintala, Mehdi Rezagholizadeh, Saed Rezayi, Leonardo F. R. Ribeiro, Caitlin Laura Richter, Sebastian Riedel, Stefan Riezler, German Rigau, Shruti Rijhwani, Matīss Rikters, Darcey Riley, Laura Rimell, Eric Ringger, Miguel Rios, Anthony Rios, Annette Rios, Brian Roark, Kirk Roberts, Christophe Rodrigues, Pedro Rodriguez, Melissa Roemmele, Lina Maria Rojas-Barahona, Stephen Roller, Roland Roller, Alexey Romanov, Salvatore Romeo, Srikanth Ronanki, Subendhu Rongali, Rudolf Rosa, Aiala Rosá, Michael Roth, Sascha Rothe, Salim Roukos, Dmitri Roussinov, Bryan R. Routledge, Subhro Roy, Aurko Roy, Jos Rozen, Alla Rozovskaya, Dongyu Ru, Raphael Rubino, Sebastian Ruder, Koustav Rudra, Frank Rudzicz, Federico Ruggeri, Thomas Ruprecht, Alexander M Rush, Irene Russo, Phillip Rust, Attapol Rutherford, Max Ryabinin, Maria Ryskina, Andreas Rücklé

C S, Ashish Sabharwal, Mrinmaya Sachan, Fatiha Sadat, Arka Sadhu, Marzieh Saeidi, Niloofar Safi Samghabadi, Kenji Sagae, Horacio Saggion, Swarnadeep Saha, Monjoy Saha, Tulika Saha, Saurav Sahay, Gaurav Sahu, Sunil Kumar Sahu, Hassan Sajjad, Keisuke Sakaguchi, Sakriani Sakti, Elizabeth Salesky, Alexandre Salle, Avneesh Saluja, Tanja Samardzic, Younes Samih, Danae Sanchez Villegas, Chinnadhurai Sankar, Malaikannan Sankarasubbu, Sashank Santhanam, Marina Santini, Bishal Santra, Sebastin Santy, Maarten Sap, Naomi Saphra, Maya Sappelli, Zahra Sarabi, Sheikh Muhammad Sarwar, Felix Sasaki, Shota Sasaki, Ryohei Sasano, Giorgio Satta, Danielle Saunders, Agata Savary, Aleksandar Savkov, Beatrice Savoldi, Apoorv Umang Saxena, Asad B. Sayeed, Thomas Schaaf, Shigehiko Schamoni, Tatjana Scheffler, Christian Scheible, Yves Scherrer, Timo Schick, Marten Van Schijndel, Frank Schilder, Viktor Schlegel, Jonathan Schler, Helmut Schmid, Tyler Schnoebelen, Steven Schockaert, Alexandra Schofield, Sabine Schulte Im Walde, Hannes Schulz, Claudia Schulz, Elliot Schumacher, Anne-Kathrin Schumann, Sebastian Schuster, Tal Schuster, Roy Schwartz, Robert Schwarzenberg, Stefan Schweter, Johannes Schäfer, Djamé Seddah, João Sedoc, Satoshi Sekine, David Semedo, Nasredine Semmar, Sina Semnani, Lütfi Kerem Senel, Rico Sennrich, Minjoon Seo, Yeon Seonwoo, Christophe Servan, Lei Sha, Izhak Shafran, Darsh Jaidip Shah, Kashif Shah, Samira Shaikh, Cory Shain, Jingbo Shang, Chao

Shang, Mingyue Shang, Guokan Shang, Nan Shao, Zhihong Shao, Chenze Shao, Yutong Shao, Ori Shapira, Naomi Tachikawa Shapiro, Amr Sharaf, Ashish Sharma, Arpit Sharma, Vasu Sharma, Serge Sharoff, Rebecca Sharp, Hassan Shavarani, Peter Shaw, Qiaoqiao She, Zaid Sheikh, Artem Shelmanov, Yilin Shen, Qinlan Shen, Tao Shen, Sheng Shen, Shiqi Shen, Yongliang Shen, Hua Shen, Jiaming Shen, Lei Shen, Yikang Shen, Xiaoyu Shen, Qiang Sheng, Emily Sheng, Tom Sherborne, Shuming Shi, Freda Shi, Peng Shi, Weijia Shi, Zhouxing Shi, Weiyan Shi, Yangyang Shi, Tianze Shi, Chuan Shi, Jiaxin Shi, Ning Shi, Xing Shi, Jiatong Shi, Tomohide Shibata, Nobuyuki Shimizu, Anastasia Shimorina, Jamin Shin, Yow-Ting Shiue, Boaz Shmueli, Eyal Shnarch, Linjun Shou, Mohit Shridhar, Akshat Shrivastava, Manish Shrivastava, Lei Shu, Kai Shu, Raphael Shu, Kurt Shuster, Vered Shwartz, Chenglei Si, Mei Si, Aditya Siddhant, A.b. Siddique, Carina Silberer, Miikka Silfverberg, Khalil Sima'an, Patrick Simianer, Kathleen Siminyu, Arabella Jane Sinclair, Sameer Singh, Karan Singla, Koustuv Sinha, Kairit Sirts, Amy Siu, Milena Slavcheva, Noam Slonim, David A. Smith, Felipe Soares, Christine Soh, Yiping Song, Kaiqiang Song, Ruihua Song, Hyun-Je Song, Linfeng Song, Xingyi Song, Kai Song, Haoyu Song, Mingyang Song, Wei Song, Sandeep Soni, Rishi Sonthalia, Claudia Soria, Alexey Sorokin, Daniil Sorokin, William Eduardo Soto Martinez, Sajad Sotudeh, Marlo Souza, Lucia Specia, Matthias Sperber, Vivek Srikumar, Balaji Vasan Srinivasan, Tejas Srinivasan, Shashank Srivastava, Edward P. Stabler, Felix Stahlberg, Ieva Staliunaite, Marija Stanojevic, Gabriel Stanovsky, David Stap, Katherine Stasaski, Manfred Stede, Mark Steedman, Benno Stein, Shane Steinert-Threlkeld, Elias Stengel-Eskin, Amanda Stent, Mark Stevenson, Ian Stewart, Matthew Stone, Kevin Stowe, Karl Stratos, Kristina Striegnitz, Heiner Stuckenschmidt, Nikolaos Stylianou, Sara Stymne, Jinsong Su, Weifeng Su, Yu Su, Keh-Yih Su, Dan Su, Yusheng Su, Shang-Yu Su, Hui Su, Nishant Subramani, Sanjay Subramanian, Lakshmi Subramanian, Katsuhito Sudoh, Saku Sugawara, Hiroaki Sugiyama, Alessandro Suglia, Yoshihiko Suhara, Dianbo Sui, Zhifang Sui, Elior Sulem, Md Arafat Sultan, Huan Sun, Yu Sun, Kai Sun, Zhiqing Sun, Siqi Sun, Zequn Sun, Mingming Sun, Changzhi Sun, Ming Sun, Yifan Sun, Fei Sun, Yawei Sun, Tianxiang Sun, Yibo Sun, Jingyi Sun, Kai Sun, Haitian Sun, Simeng Sun, Si Sun, Jian Sun, Haipeng Sun, Chengjie Sun, Dhanasekar Sundararaman, Mujeen Sung, Hanna Suominen, Mihai Surdeanu, Anshuman Suri, Shiv Surya, Simon Suster, Mirac Suzgun, Jun Suzuki, Masatoshi Suzuki, Swabha Swayamdipta, Benjamin Sznajder, Stan Szpakowicz, Felipe Sánchez-Martínez

Ryuki Tachibana, Oyvind Tafjord, Shabnam Tafreshi, Hiroya Takamura, Ryuichi Takanobu, Sho Takase, Ece Takmaz, Aarne Talman, Derek Tam, George Tambouratzis, Fabio Tamburini, Akihiro Tamura, Xu Tan, Chuanqi Tan, Samson Tan, Fei Tan, Zeqi Tan, Liling Tan, Kumiko Tanaka-Ishii, Siliang Tang, Gongbo Tang, Qingming Tang, Shuai Tang, Yi-Kun Tang, Raphael Tang, Buzhou Tang, Zhiwen Tang, Hao Tang, Ludovic Tanguy, Xavier Tannier, Chongyang Tao, Shiva Taslimipoor, Sandeep Tata, Yuka Tateisi, Michiaki Tatsubori, Marta Tatu, Hillel Taub-Tabib, Yi Tay, Andon Tchechmedjiev, Christoph Teichmann, Selma Tekir, Serra Sinem Tekiroglu, Eric S. Tellez, Irina Temnikova, Zhiyang Teng, Ian Tenney, Hiroki Teranishi, Silvia Terragni, Alberto Testoni, Nithum Thain, Khushboo Thaker, Urmish Thakker, Nandan Thakur, Kilian Theil, Jesse Thomason, Laure Thompson, Sam Thomson, Camilo Thorne, James Thorne, Zhiliang Tian, Ran Tian, Junfeng Tian, Yingtao Tian, Jörg Tiedemann, Tiago Timponi Torrent, Erik Tjong Kim Sang, Gaurav Singh Tomar, Nadi Tomeh, Nicholas Tomlin, Sara Tonelli, Mariya Toneva, MeiHan Tong, Antonio Toral, Kentaro Torisawa, Samia Touileb, Julien Tourille, Quan Hung Tran, Dietrich Trautmann, Marcos Vinicius Treviso, Hai-Long Trieu, Alina Trifan, Enrica Troiano, Tuan Quoc Truong, Chen-Tse Tsai, Bo-Hsiang Tseng, Masaaki Tsuchida, Yoshimasa Tsuruoka, Zhaopeng Tu, Kewei Tu, Lifu Tu, Mei Tu, Iulia Raluca Turc, Martin Tutek, Francis M. Tyers, Andre Tättar

Rutuja Ubale, Ana Sabina Uban, Takuma Udagawa, Umair Ul Hassan, Stefan Ultes, Shyam Upadhyay, L. Alfonso Ureña, Ricardo Usbeck

Keyon Vafa, Sowmya Vajjala, Jannis Vamvas, Tim Van De Cruys, Benjamin Van Durme, Emiel Van Miltenburg, Rik Van Noord, Keith N VanderLinden, Lucy Vanderwende, David Vandyke, Na-

talia Vanetik, Daniel Varab, Siddharth Varia, Lucy Vasserman, Julien Velcin, Alakananda Vempala, Sriram Venkatapathy, Giulia Venturi, Suzan Verberne, Gaurav Verma, Rakesh M Verma, Giorgos Vernikos, Yannick Versley, Karin Verspoor, Anvesh Rao Vijjini, David Vilar, Jesús Vilares, Serena Villata, Aline Villavicencio, Éric Villemonte De La Clergerie, Veronika Vincze, Krishnapriya Vishnubhotla, Ngoc Phuoc An Vo, Rob Voigt, Elena Voita, Soroush Vosoughi, Thang Vu, Tu Vu, Thuy Vu, Thuy-Trang Vu, Xuan-Son Vu, Yogarshi Vyas, Ekaterina Vylomova

Henning Wachsmuth, Takashi Wada, Joachim Wagner, Byron C Wallace, Stephen Wan, Mengting Wan, Mingyu Wan, Yao Wan, Yu Wan, Wei Wang, Xing Wang, Hai Wang, Yizhong Wang, Rui Wang, Alex Wang, Xiaolin Wang, Qingyun Wang, Xinyi Wang, Shuohang Wang, Jin Wang, Yan Wang, Dingmin Wang, Baoxun Wang, Yile Wang, Xin Wang, Zhongqing Wang, Guoyin Wang, Xiaozhi Wang, Cunxiang Wang, Fei Wang, Zhen Wang, Haohan Wang, Jingkang Wang, Bingqing Wang, Ping Wang, Guangrun Wang, Wenya Wang, Zirui Wang, Chao Wang, Wei Wang, Lucy Lu Wang, Yaqing Wang, Hongfei Wang, Jin Wang, Haoyu Wang, Hao Wang, Bin Wang, Zijian Wang, Hanrui Wang, Liang Wang, Shuo Wang, Xiaojie Wang, Tong Wang, Chenguang Wang, Wen Wang, Qiang Wang, Hua Wang, Yifei Wang, Boxin Wang, Hao Wang, Qifan Wang, Yue Wang, Lidan Wang, Changhan Wang, Lingzhi Wang, Yue Wang, Pinghui Wang, Zhichun Wang, Wenhui Wang, Quan Wang, Jingang Wang, Daling Wang, Han Wang, Rui Wang, Yijue Wang, Yong Wang, Yiran Wang, Tong Wang, Yequan Wang, Ke Wang, Sinong Wang, Baoxin Wang, Runze Wang, Bailin Wang, Yujing Wang, Shi Wang, Jue Wang, Hong Wang, Wenbo Wang, Xuezhi Wang, Weiyue Wang, Liwen Wang, Shaonan Wang, Yingyao Wang, Ziqi Wang, Chengyu Wang, Leo Wanner, Nigel G. Ward, Alex Warstadt, Christian Wartena, Koki Washio, Noah Weber, Leon Weber, Ingmar Weber, Kellie Webster, Julie Weeds, Jason Wei, Xiangpeng Wei, Junqiu Wei, Penghui Wei, Wei Wei, Xiaochi Wei, Johnny Wei, Shira Wein, David Weir, Ralph M. Weischedel, Charles Welch, Orion Weller, Haoyang Wen, Lijie Wen, Rongxiang Weng, Peter West, Taesun Whang, Michael White, Michael Wiegand, Sarah Wiegreffe, Adam Wiemerslage, Derry Wijaya, Gijs Wijnholds, Ethan Wilcox, Rodrigo Wilkens, Jennifer Williams, Jake Ryland Williams, Steven R. Wilson, Shomir Wilson, Genta Indra Winata, Shuly Wintner, Sam Wiseman, Guillaume Wisniewski, Magdalena Wolska, Derek F. Wong, Tak-Lam Wong, Dina Wonsever, Zach Wood-Doughty, Chien-Sheng Wu, Lijun Wu, Chuhan Wu, Lingfei Wu, Youzheng Wu, Fangzhao Wu, Shijie Wu, Chun-Kai Wu, Xianchao Wu, Yuting Wu, Zhiyong Wu, Wei Wu, Yuanbin Wu, Zhen Wu, Zeqiu Wu, Stephen Wu, Junshuang Wu, Yanan Wu, Qianhui Wu, Zhonghai Wu, Xiaobao Wu, Dayong Wu, Jian Wu, Di Wu, Shuangzhi Wu, Bowen Wu, Tongshuang Wu, Lianwei Wu, Bo Wu, Sixing Wu, Yu Wu, Yunfang Wu, Qingyang Wu, Joern Wuebker

Patrick Xia, Congying Xia, Qingrong Xia, Jingbo Xia, Mengzhou Xia, Rui Xia, Yikun Xian, Jiannan Xiang, Rong Xiang, Lin Xiao, Yanghua Xiao, Chaojun Xiao, Huiru Xiao, Liqiang Xiao, Wen Xiao, Min Xiao, Chunyang Xiao, Tong Xiao, Jinghui Xiao, Boyi Xie, Jun Xie, Yuqiang Xie, Qianqian Xie, Tianbao Xie, Ruobing Xie, Ji Xin, Frank Xing, Deyi Xiong, Wenhan Xiong, Jiacheng Xu, Can Xu, Frank F. Xu, Canwen Xu, Kun Xu, Ruochen Xu, Peng Xu, Ruifeng Xu, Jinan Xu, Peng Xu, Hongfei Xu, Weiran Xu, Chen Xu, Hu Xu, Boyan Xu, Jing Xu, Lu Xu, Yan Xu, Qiongkai Xu, Qiantong Xu, Jingjing Xu, Jia Xu, Wang Xu, Zenglin Xu, Dongkuan Xu, Weiwen Xu, Wenduan Xu, Zhen Xu, Benfeng Xu, Zhiyang Xu, Weijia Xu, Xinnuo Xu, Shusheng Xu, Yang Xu, Yumo Xu, Jitao Xu, Jun Xu, Runxin Xu

Shuntaro Yada, Vikas Yadav, Yadollah Yaghoobzadeh, Ikuya Yamada, Ivan P. Yamshchikov, Yu Yan, Hanqi Yan, Jun Yan, Yuanmeng Yan, Lingyong Yan, Min Yang, Wei Yang, Zhilin Yang, Yujiu Yang, Chenghao Yang, Mingming Yang, Zhen Yang, Baosong Yang, Songlin Yang, Jie Yang, Yinfei Yang, Sen Yang, Yilin Yang, Zhichao Yang, Linyi Yang, Haiqin Yang, Muyun Yang, Ziyi Yang, Zhao Yang, Tsung-Yen Yang, Fan Yang, Changbing Yang, Ziqing Yang, Sen Yang, Jun Yang, Ruosong Yang, Wenmian Yang, Shunyu Yao, Jianmin Yao, Wenlin Yao, Ziyu Yao, Liang Yao, Mark Yatskar, Qinyuan Ye, Deming Ye, Reyyan Yeniterzi, Jinyoung Yeo, Xiaoyuan Yi, Seid Muhie Yimam, Pengcheng Yin, Yichun Yin, Xuwang Yin, Qingyu Yin, Da Yin, Sho Yokoi, Zheng

Xin Yong, Kang Min Yoo, Seunghyun Yoon, Masashi Yoshikawa, Steve Young, Safoora Yousefi, Tiezheng Yu, Kai Yu, Dian Yu, Juntao Yu, Yue Yu, Mo Yu, Tao Yu, Dong Yu, Wenhao Yu, Dian Yu, Heng Yu, Bowen Yu, Xiaodong Yu, Changlong Yu, Chen Yu, Bei Yu, Jifan Yu, Hong Yu, Jianhua Yuan, Zheng Yuan, Yu Yuan, Nicholas Jing Yuan, Caixia Yuan, Xiang Yue, Hyokun Yun

Annie Zaenen, Wajdi Zaghouani, Marcos Zampieri, Marcely Zanon Boito, Alessandra Zarcone, Sina Zarrieß, Vicky Zayats, Rabih Zbib, Albin Zehe, Rowan Zellers, Yury Zemlyanskiy, Daojian Zeng, Jiali Zeng, Qi Zeng, Weixin Zeng, Fengzhu Zeng, Shuang Zeng, Xingshan Zeng, Jichuan Zeng, Zhiyuan Zeng, Thomas Zenkel, Deniz Zeyrek, Hanwen Zha, Fangzhou Zhai, Haolan Zhan, Runzhe Zhan, Li-Ming Zhan, Zhuosheng Zhang, Yuhao Zhang, Weinan Zhang, Richong Zhang, Peng Zhang, Jinchao Zhang, Qi Zhang, Yu Zhang, Yue Zhang, Rui Zhang, Danqing Zhang, Ruixiang Zhang, Meishan Zhang, Zhihao Zhang, Biao Zhang, Jiajun Zhang, Yuan Zhang, Guanhua Zhang, Yao Zhang, Denghui Zhang, Hao Zhang, Ningyu Zhang, Wei Zhang, Xinyuan Zhang, Hainan Zhang, Zhirui Zhang, Yan Zhang, Wen Zhang, Dongxu Zhang, Jianguo Zhang, Michael JQ Zhang, Yichi Zhang, Xingxing Zhang, Tongtao Zhang, Ke Zhang, Shuo Zhang, Jingqing Zhang, Xiaotong Zhang, Zeyu Zhang, Xinsong Zhang, Yunyi Zhang, Hongming Zhang, Meng Zhang, Kun Zhang, Sheng Zhang, Jieyu Zhang, Licheng Zhang, Chuheng Zhang, Lei Zhang, Chen Zhang, Shiyue Zhang, Haoyu Zhang, Yusen Zhang, Xinliang Frederick Zhang, Hao Zhang, Delvin Ce Zhang, Chiyu Zhang, Xiao Zhang, Yan Zhang, Zheng Zhang, Wen Zhang, Bowen Zhang, Bowen Zhang, Haibo Zhang, Yu Zhang, Xuanyu Zhang, Hu Zhang, Tong Zhang, Xuchao Zhang, Wei Emma Zhang, Ziqi Zhang, Xiang Zhang, Dong Zhang, Tianlin Zhang, Dawei Zhang, Xuanwei Zhang, Dongyu Zhang, Haisong Zhang, Yuanzhe Zhang, Dongdong Zhang, Yuhui Zhang, Min Zhang, Lei Zhang, Chen Zhang, Li Zhang, Shujian Zhang, Mike Zhang, Zhisong Zhang, Longyin Zhang, Zhengyan Zhang, Tiancheng Zhao, Dongyan Zhao, Jieyu Zhao, Zhenjie Zhao, Yanpeng Zhao, Tianyu Zhao, Sanqiang Zhao, Guangxiang Zhao, Chao Zhao, Kai Zhao, Mengjie Zhao, Yang Zhao, Chen Zhao, Yilun Zhao, Tiejun Zhao, Yao Zhao, Zhou Zhao, Xiaoqing Zheng, Renjie Zheng, Yinhe Zheng, Bo Zheng, Zaixiang Zheng, Changmeng Zheng, Chujie Zheng, Zexuan Zhong, Ming Zhong, Victor Zhong, Wanjun Zhong, Peixiang Zhong, Guangyou Zhou, Ben Zhou, Zhihan Zhou, Qingyu Zhou, Chunting Zhou, Yichao Zhou, Yaqian Zhou, Jingbo Zhou, Shuyan Zhou, Yilun Zhou, Yichu Zhou, Long Zhou, Junpei Zhou, Xiang Zhou, Pei Zhou, Yi Zhou, Xiangyang Zhou, Junsheng Zhou, Yucheng Zhou, Jiawei Zhou, Wenxuan Zhou, Jie Zhou, Jie Zhou, Giulio Zhou, Dong Zhou, Deyu Zhou, Zhengyu Zhou, Wangchunshu Zhou, Meng Zhou, Li Zhou, Kenny Q. Zhu, Junnan Zhu, Muhua Zhu, Hao Zhu, Qingfu Zhu, Su Zhu, Lixing Zhu, Zining Zhu, Conghui Zhu, Qinglin Zhu, Wei Zhu, Xiaoyan Zhu, Jun Zhu, Yilun Zhu, Jian Zhu, Yong Zhu, Qi Zhu, Yimeng Zhuang, Fuzhen Zhuang, Caleb Ziems, Roger Zimmermann, Heike Zinsmeister, Ayah Zirikly, Shi Zong, Bowei Zou, Yanyan Zou, Amal Zouaq, Arkaitz Zubiaga, Pierre Zweigenbaum

Erion Çano

Robert Östling

Lilja Øvrelid

Gözde Gül Şahin

Outstanding Action Editors

Antonios Anastasopoulos, David Bamman, Steven Bethard, Leonid Boytsov, Paula Carvalho, Snigdha Chaturvedi, Raj Dabre, Daniel Dakota, Johannes Daxenberger, Leon Derczynski, Greg Durrett, Michael Elhadad, Allyson Ettinger, Goran Glavaš, David Harwath, Shubhra Kanti Karmaker, Daniel Khashabi, Mamoru Komachi, Carolin Lawrence, John Lawrence, Constantine Lig-

nos, Saif M. Mohammad, Philippe Muller, Rebecca J. Passonneau, Emily Prud'hommeaux, Mrinmaya Sachan, Lane Schwartz, Kevin Small, Efstathios Stamatatos, Amanda Stent, Amalia Todirascu, Junichi Tsujii, Suzan Verberne, Antonio Jimeno Yepes, François Yvon, Luke Zettlemoyer, Justine Zhang

Outstanding Reviewers

Nader Akoury, Gianni Barlacchi, Rachel Bawden, Gábor Bella, Delphine Bernhard, Shruti Bhosale, Michael Bloodgood, Ondrej Bojar, Iacer Calixto, Rémi Cardon, Thiago Castro Ferreira, Tuhin Chakrabarty, Verna Dankers, Yupei Du, Micha Elsner, Antske Fokkens, Stella Frank, Alexander Fraser, Dayne Freitag, Daniel Fried, Dan Garrette, Philip John Gorinski, Dagmar Gromann, Liane Guillou, Jack Hessel, Nanjiang Jiang, Gareth J. F. Jones, Min-Yen Kan, Anna Kazantseva, Fajri Koto, Julia Kreutzer, Kalpesh Krishna, Dawn Lawrie, Andrew Lee, Jordan Lee Boyd-Graber, Gina-Anne Levow, Xiang Lisa Li, Patrick William Littell, Kaixin Ma, Vladislav Maraev, Alexander Mehler, Florian Metze, Julian Michael, Paul Michel, Elena Musi, Sheshera Mysore, Denis Newman-Griffis, Tong Niu, Michal Novák, Siddharth Patwardhan, Karl Pichotta, Yuval Pinter, Peng Qi, Surangika Ranathunga, Vikas Raunak, Pedro Rodriguez, Sebastian Ruder, Alexander M. Rush, Elizabeth Salesky, Thomas Schaaf, Yves Scherrer, Viktor Schlegel, Elliot Schumacher, Ian Stewart, Naomi Tachikawa Shapiro, Emiel van Miltenburg, Peter West, Adam Wiemerslage, Jitao Xu, Yue Yu, Yury Zemlyanskiy

2 Anti-Harassment Policy

ACL 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 Meal Info

Tea, coffee and pastries or biscuits will be provided early morning, mid-morning and mid-afternoon. Lunch is not provided, but there are plenty of cafes, restaurants and shops in the city centre, which is around 20 minutes' walk from the CCD. A meal will be provided during the Welcome Reception on Sunday evening (starting at 18:30) and the Conference Dinner on Monday evening (starting at 19:30).

4 Social Events

We want to ensure ACL 2022 is a meeting to remember; not only as a result of the amazing speakers and conference venue, but due to the lively Social Programme. What visitors enjoy the most about Dublin are the Dubliners themselves and the hospitality and warm welcome offered to our visitors. Ireland's wonderful capital is a place to build lasting memories, savour great experiences and celebrate the unexpected. We will ensure all our delegates experience the best we have to offer. Here's a taste of what's planned!

Welcome Reception - Sunday 22nd, May, 2022

Venue: **The Convention Centre, Dublin**
Time: **18.30 - 21.00**
Dress code: **Smart Casual**

Enjoy drinks, a light buffet and panoramic views of the River Liffey, Dublin city centre and the Wicklow mountains as you unwind after day one. The Reception will allow for optimum networking and socialising with friends and colleagues. With the Convention Centre based in the heart of Dublin's Docklands, and just a few minutes' walk from the city centre, you are spoilt for choice on places to visit after the reception. Whether you are seeking a quiet drink or a lively music session, you will find your perfect spot nearby. A short walk, cycle or tram ride will connect you to the world famous Temple Bar. A hive of activity, there is no shortage of bars in the neighbourhood with Ceol agus Craic (Music and Fun) aplenty!

Conference Dinner - Monday 23th, May, 2022

Venue: **Guinness Storehouse**
Time: **19.30 - 23.00**
Dress code: **Casual**

To top it all off, the Conference Dinner will be held in the famous Guinness Storehouse. Guests will have exclusive access to Ireland's No. 1 visitor attraction for a memorable night of Irish food and entertainment. The Guinness Brewery in Dublin is Europe's largest stout producing brewery and home to the

Guinness Storehouse. Opened in 1904, the Storehouse was an operational plant for fermenting and storing Guinness. Today it houses an exhibition dedicated to the Guinness story. Visitors will discover what goes into the making of a pint of Guinness - the ingredients, the brewing process, the time, the craft and the passion. The exhibition shows how the famous brew has been marketed and how it is today sold in over 150 countries. Coaches will depart from outside the CCD and the Clayton Cardiff Lane HQ Hotel from 19.00 to transport delegates to the Guinness Storehouse. Coaches will depart from the Guinness Storehouse from 22.00 onwards.

5 Keynotes

Keynote Talk: Language in the Human Brain

Angela D. Friederici
Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Monday, May 23, 2022 - Room: **Auditorium** - Time: **9:30-10:30**

Abstract: Language is considered to be a uniquely human faculty. The different aspects of the language system, namely phonology, semantics and syntax have long been discussed with respect to their speciesspecificity. Syntax as the ability to process hierarchical structures appears to be specific to humans. The available neuroscientific data allow us to define the functional language network which involves Broca's area in the inferior frontal cortex and the posterior superior temporal cortex. Within this network, the posterior part of Broca's area plays a special role as it supports the processing of hierarchical syntactic structures, in particular the linguistic computation Merge which is at the root of every language. This part of Broca's area is connected to the posterior temporal cortex via a dorsally located white matter fiber tract hereby providing to structural basis for the functional interplay of these regions. It has been shown that the maturation of this white matter pathway is directly correlated with the ability to process syntactically complex sentences during human development. Moreover, this dorsal pathway appears to be weak in the prelinguistic infant and in the non-human primate. These findings suggest that the dorsal pathway plays a

crucial role in the emergence of syntax in human language.

Bio: Angela D. Friederici is a cognitive neuroscientist in the domain of language. She is director at the Max Planck Institute for Human Cognitive and Brain Sciences (MPI CBS) in Leipzig, Germany and the Founding director of this institution founded in 1994.
She graduated in linguistics and psychology at the University of Bonn (Germany) and spent a postdoctoral year at MIT (USA). She was a research fellow at the Max Planck Institute in Nijmegen (NL), at the University Rene Descartes, Paris (F) and University of California, San Diego (USA). Prior to joining the Max Planck Society as a director, she was professor for Cognitive Sciences at the Free University Berlin.
Friederici is honorary professor at the University of Leipzig (Psychology), the University of Potsdam (Linguistics) and the Charité Universitätsmedizin Berlin (Neurology) and she holds a Doctor honoris causa from the University of Mons, Belgium. Between 2014 and 2020 she was Vice President for the Human Sciences Section of the Max Planck Society.
Her main field of research is the neurobiology of language. She published about 500 scientific papers on this topic in major international journals. She received a number of scientific awards: 1987 Heisenberg Fellowship of the German Research Foundation, 1990 Alfried Krupp Award of the Alfried Krupp von Bohlen and Halbach-Stiftung, 1997 Gottfried Wilhelm Leibniz Prize of the German Research Foundation, and 2011 Carl Friedrich Gauss Medal of the Brunswick Scientific Society. She is member of the Berlin-Brandenburg Academy of Sciences and Humanities, member of the national German Academy of Sciences 'Leopoldina' and member of the Academia Europaea.

Keynote Fire-Side Chat with Barbara Grosz and Yejin Choi on "The Trajectory of ACL and the Next 60 years"

For the 60th Anniversary of ACL 2022, we will feature a keynote fire-side chat on "***The Trajectory of ACL and the Next 60 years***" with two keynote talks in dialogue: Barbara Grosz and Yejin Choi followed by a moderated discussion lead by Rada Mihalcea.

Remarks on What the Past Can Tell the Future

Barbara J. Grosz
Harvard University SEAS

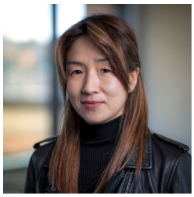
Tuesday, May 24, 2022 - Room: **Auditorium** - Time: **16:45-18:30**

Abstract: Research in computational linguistics and spoken language systems has made astonishing progress in the last decade. Even so, the challenge remains of achieving human-level fluent dialogue conversational capabilities beyond narrowly defined domains and tasks. Findings of earlier ACL times research on dialogue hold some lessons for breaking the "dialogue boundary" in computational linguistics yet again, if ways can be found to integrate them into deep-learning language models. These models raise some of the most serious ethical challenges of current computing research and technologies. Expanding their powers in this direction will raise more. In discussing these topics, I will raise questions for Prof. Choi and our subsequent discussion.

Bio: Barbara J. Grosz is Higgins Research Professor of Natural Sciences in the Paulson School of Engineering and Applied Sciences at Harvard University. Her contributions to AI include fundamental advances in natural-language dialogue processing and in theories of multi-agent collaboration as well as innovative uses of models developed in this research to improve healthcare coordination and science education. She co-founded Harvard's Embedded EthiCS program, which integrates teaching of ethical reasoning into core computer science courses. A member of the National Academy of Engineering, the American Philosophical Society, and the American Academy of Arts and Sciences, she is a fellow of several scientific societies and recipient of the 2009 ACM/AAAI Allen Newell Award, the 2015 IJCAI Award for Research Excellence, and the 2017 Association for Computational Linguistics Lifetime Achievement Award.

2082: An ACL Odyssey
The Dark Matter of Intelligence and Language

Yejin Choi
Paul G. Allen School of Computer Science & Engineering at the University of Washington



Tuesday, May 24, 2022 - Room: **Auditorium** - Time: **16:45-18:30**

Abstract: In this talk, I will wander around reflections on the past of ACL and speculations on the future of ACL. This talk will be purposefully imaginative and accidentally controversial, by emphasizing on the importance of deciphering the dark matter of intelligence, by arguing for embracing all the ambiguous aspects of language at all pipelines of language processing, by highlighting the counterintuitive continuum across language, knowledge, and reasoning, and by pitching the renewed importance of formalisms, algorithms, and structural inferences in the modern deep learning era. Looking back, at the 50'th ACL, I couldn't possibly imagine that I would be one day giving this very talk. For that reason, I will also share my personal anecdotes on the lasting inspirations from the previous lifetime achievement award speeches, how I believe talent is made, not born, and the implication of that belief for promoting diversity and equity.

Bio: Yejin Choi is Brett Helsel Professor at the Paul G. Allen School of Computer Science & Engineering at the University of Washington and a senior research manager at AI2 overseeing the project Mosaic. Her research investigates commonsense knowledge and reasoning, neuro-symbolic integration, neural language generation and degeneration, multimodal representation learning, and AI for social good. She is a co-recipient of the ACL Test of Time award in 2021, the CVPR Longuet-Higgins Prize in 2021, a NeurIPS Outstanding Paper Award in 2021, the AAAI Outstanding Paper Award in 2020, the Borg Early Career Award in 2018, the inaugural Alexa Prize Challenge in 2017, IEEE AI's 10 to Watch in 2016, and the ICCV Marr Prize in 2013.

Moderator: Rada Mihalcea

Computer Science and Engineering, University of Michigan

Bio: Rada Mihalcea is the Janice M. Jenkins Collegiate Professor of Computer Science and Engineering at the University of Michigan and the Director of the Michigan Artificial Intelligence Lab. Her research interests are in computational linguistics, with a focus on lexical semantics, multilingual natural language processing, and computational social sciences. She was a program co-chair for EMNLP 2009 and ACL 2011, and a general chair for NAACL 2015 and *SEM 2019. She currently serves as ACL Past President. She is the recipient of a Presidential Early Career Award for Scientists and Engineers awarded by President Obama (2009), she is an ACM Fellow (2019) and a AAAI Fellow (2021). In 2013, she was made an honorary citizen of her hometown of Cluj-Napoca, Romania.

Keynote Panel: "Supporting Linguistic Diversity"

Chair: **Steven Bird**
Charles Darwin University, Australia






Wednesday, May 25, 2022 - Room: **Auditorium** - Time: **9:00-10:15**

Panelists and languages represented:

- Robert Jimerson, Rochester Institute of Technology (Seneca, USA)
- Fajri Koto, The University of Melbourne (Minangkabau, Indonesia)
- Heather Lent, University of Copenhagen (Creole languages)
- Teresa Lynn, Dublin City University (Irish)
- Manuel Mager, University of Stuttgart (Wixaritari, Mexico)
- Perez Ogayo, Carnegie Mellon University (Luo and Kiswahili, Kenya)

How do the tools and techniques of computational linguistics serve the full diversity of the world's languages? In particular, how do they serve the people who are still speaking thousands of local languages, often in highly multilingual, post-colonial situations? This 60th meeting of the ACL features a special theme track on language diversity with the goal of "reflecting and stimulating discussion about how the advances in computational linguistics and natural language processing can be used for promoting language diversity". This keynote talk-panel will showcase the special theme and identify key learnings from the conference. We hope this session will help to shape the future agenda for speech and language technologies in support of global linguistic diversity. The session will be organised around a series of questions under three headings.

Diverse Contexts. What is the situation of local languages where panel members are working? Are there multiple languages with distinct functions and ideologies? What are the local aspirations for the future of these languages. How are people advocating for language technology on the ground? How did the work begin? What does success look like?

Understanding Risks. Do the people who provide language data fully understand the ways their data might be used in future, including ways that might not be in their interest? What benefit are local participants promised in return for their participation, and do they actually receive these benefits? Are there harms that come with language standardisation? What principles of doing no harm can we adopt?

New Challenges. How can we provide benefits of text technologies without assuming language standardisation, official orthography, and monolingual usage? When working with local communities, do we always require data in exchange for technologies, or is a non-extractive NLP possible? How do we decolonise speech and language technology? At the beginning of the International Decade of Indigenous Languages 2022–2032, we ask: how do we respond as a community, and how can our field be more accessible to indigenous participation?

Steven Bird, Charles Darwin University, Australia
Bio: Steven Bird has spent much of his research career pursuing scalable computational methods for capturing, enriching, and analysing data from endangered languages, drawing on fieldwork in West Africa, South America, and Melanesia. Over the past 5 years he has shifted to working from the ground up with remote Aboriginal communities, supporting language learning and development in an Aboriginal ranger program, school, and arts centre.

Robert Jimerson, Rochester Institute of Technology
Bio: Robert Jimerson is a PhD candidate in the Thomas Golisano College of Computing and Information Science at the Rochester Institute of Technology. One of Robert's areas of research is using deep learning algorithms in automatic speech recognition of low-resource Indigenous languages. Robert is a member of the Seneca Nation and a speaker of the Seneca language, a North American Indigenous language that is a part of the Hodinöhšöni/Rotinohshonni (Iroquoian) family of languages.

Fajri Koto, University of Melbourne
Bio: Fajri Koto is an Australia Awards awardee and a fourth-year PhD student at the School of Computing and Information Systems, the University of Melbourne, with research interests in NLP for Indonesian languages, text summarization, generation and discourse analysis. Fajri is a native speaker of Minangkabau and Indonesian, and his ongoing research is on NLP systems for ten Indonesian local languages including Acehnese, Ngaju, Madurese, Bataknese, Buginese, Banjarese, Sundanese, Balinese, Javanese, and Minangkabau. Previously, Fajri joined Amazon and Samsung R&D Institute as a research scientist.

Heather Lent, University of Copenhagen
Bio: Heather Lent is a Ph.D. fellow at the University of Copenhagen with the Datalogisk Institut in Denmark. Her primary research interests are focused on transfer learning in NLP for low-resource languages, and in particular Creole languages. Heather's philosophy for this work is to involve Creole language speakers in her research, as this diverse set of languages proves that there is no "one size fits all" approach to language technology. In the past, Heather has also engaged in research in semantic parsing.

Teresa Lynn, Dublin City University, Ireland
Bio: Teresa Lynn is a Research Fellow at the ADAPT Centre in Dublin City University. Teresa's main interests lie in developing tools and resources for Irish language technology. She is the principal investigator on the GaelTech project, funded by the Irish Department of the Gaeltacht, which covers various research topics in Irish language technology. She is also a core member of the European Language Equality project and Ireland's National Anchor Point for the ELRC (European Language Resource Coordination), overseeing national data collection for Irish machine translation. Her research covers treebank development, syntactic parsing, social media NLP and multiword expressions.

Manuel Mager (Turatemai), University of Stuttgart
Bio: Manuel Mager is a Ph.D. candidate at the University of Stuttgart (Institute for Natural Language Processing), Germany. His work is focused on Natural Language Processing for low resource languages, morphological analysis and translation of polysynthetic languages, and code-switching. He is co-organizer of the AmericasNLP workshop and member of the Wixarika community. His main aim is to include indigenous languages of the Americas into the current NLP community and democratize the advancements in the field to all languages of the world.

Perez Ogayo, Carnegie Mellon University
Bio: Perez Ogayo is a master's student at Carnegie Mellon University in the Language Technologies Institute (LTI) where she is focusing on low resource natural language processing. Her interests in NLP are in machine translation, speech synthesis and recognition and NLP for endangered languages. She is a researcher at Masakhane working on Luo and Kiswahili.

6

Tutorials: Sunday, May 22, 2022

Overview

07:30 - 18:00	***Registration***	
09:30 - 11:00	***Morning tutorials – Session 1***	
	Tutorial 2 – Towards Reproducible Machine Learning Research in Natural Language Processing Ana Lucic, Maurits Bleeker, Samarth Bhargav, Jessica Zosa Forde, Koustuv Sinha, Jesse Dodge, Alexandra Luccioni, and Robert Stojnic	*Liffey Hall 1*
	Tutorial 3 – Knowledge-Augmented Methods for Natural Language Processing Chenguang Zhu, Yichong Xu, Xiang Ren, Bill Yuchen Lin, Meng Jiang, and Wenhao Yu	*Wicklow Hall 2*
	Tutorial 5 – Learning with Limited Text Data Diyi Yang, Ankur P Parikh, and Colin Raffel	*The Liffey A*
	Tutorial 8 – Natural Language Processing for Multilingual Task-Oriented Dialogue Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Edoardo Ponti, Ivan Vulić	*Liffey Hall 2*
11:00 - 11:30	***Coffee break***	
11:30 - 13:00	***Morning tutorials – Session 2***	
	Tutorial 2 – Towards Reproducible Machine Learning Research in Natural Language Processing Ana Lucic, Maurits Bleeker, Samarth Bhargav, Jessica Zosa Forde, Koustuv Sinha, Jesse Dodge, Alexandra Luccioni, and Robert Stojnic	*Liffey Hall 1*
	Tutorial 3 – Knowledge-Augmented Methods for Natural Language Processing	*Wicklow Hall 2*

	Chenguang Zhu, Yichong Xu, Xiang Ren, Bill Yuchen Lin, Meng Jiang, and Wenhao Yu	
	Tutorial 5 – Learning with Limited Text Data Diyi Yang, Ankur P Parikh, and Colin Raffel	*The Liffey A*
	Tutorial 8 – Natural Language Processing for Multilingual Task-Oriented Dialogue Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Edoardo Ponti, and Ivan Vulić	*Liffey Hall 2*
13:00 - 14:30	***Lunch break***	
14:30 - 16:00	***Afternoon tutorials – Session 1***	
	Tutorial 1 – A Gentle Introduction to Deep Nets and Opportunities for the Future Kenneth Church, Valia Kordoni, Gary Marcus, Ernest Davis, Yan-jun Ma, and Zeyu Chen	*Liffey Hall 1*
	Tutorial 4 – Non-Autoregressive Sequence Generation Jiatao Gu and Xu Tan	*Liffey Hall 2*
	Tutorial 6 – Zero- and Few-Shot NLP with Pretrained Language Models Iz Beltagy, Arman Cohan, Robert L. Logan IV, Sewon Min, and Sameer Singh	*The Liffey A*
	Tutorial 7 – Vision-Language Pretraining: Current Trends and the Future Aishwarya Agrawal, Damien Teney, and Aida Nematzadeh	*Wicklow Hall 2*
16:00 - 16:30	***Coffee break***	
16:30 - 18:00	***Afternoon tutorials – Session 2***	
	Tutorial 1 – A Gentle Introduction to Deep Nets and Opportunities for the Future Kenneth Church, Valia Kordoni, Gary Marcus, Ernest Davis, Yan-jun Ma, and Zeyu Chen	*Liffey Hall 1*
	Tutorial 4 – Non-Autoregressive Sequence Generation Jiatao Gu and Xu Tan	*Liffey Hall 2*
	Tutorial 6 – Zero- and Few-Shot NLP with Pretrained Language Models Iz Beltagy, Arman Cohan, Robert L. Logan IV, Sewon Min, and Sameer Singh	*The Liffey A*
	Tutorial 7 – Vision-Language Pretraining: Current Trends and the Future Aishwarya Agrawal, Damien Teney, and Aida Nematzadeh	*Wicklow Hall 2*

Message from the Tutorial Co-Chairs

Welcome to the Tutorials Session of ACL 2022.

The ACL tutorials session is organized to give conference attendees a comprehensive introduction by expert researchers to some topics of importance drawn from our rapidly growing and changing research field.

This year, as has been the tradition over the past few years, the call, submission, reviewing and selection of tutorials were coordinated jointly for multiple conferences: ACL, NAACL, COLING and EMNLP. We formed a review committee of 34 members, including the ACL tutorial chairs (Luciana Benotti (then), Naoaki Okazaki, and Marcos Zampieri), the NAACL tutorial chairs (Cecilia O. Alm, Yulia Tsetkov, and Miguel Ballesteros), the COLING tutorial chairs (Heng Ji, Hsin-Hsi Chen, and Lucia Donatelli), the EMNLP tutorial chairs (Samhaa R. El-Beltagy and Xipeng Qiu), and 23 external reviewers (see Program Committee of the Tutorial Abstracts volume of the proceedings for the full list). A reviewing process was organised so that each proposal 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. A total of 47 tutorial submissions were received, of which 8 were selected for presentation at ACL.

We solicited two types of tutorials, namely cutting-edge themes and introductory themes. The 8 tutorials for ACL include 2 introductory tutorials and 6 cutting-edge tutorials. The introductory tutorials are dedicated to deep neural networks and reproducibility in NLP. The cutting-edge discussions address knowledge-augmented methods, non-autoregressive sequence generation, learning with limited data, zero- and few-shot learning with pretrained language models, vision-language pretraining, and multilingual task-oriented dialogue.

We would like to thank the tutorial authors for their contributions and flexibility while organising the conference in the hybrid mode. We are also grateful to the 23 external reviewers for their generous help in the decision process. Our thanks go to the conference organizers for effective collaboration, and in particular to the general chair Bernardo Magnini, the publication chair Danilo Croce, the handbook chair Marco Polignano, and the authors of `aclpub2`. Finally, special thanks go to Luciana Benotti, who worked hard as a tutorial chair of ACL especially maintaining the reviewing process (including the administrative work with OpenReview) but later resigned from this position when she was elected to the NAACL executive board as the NAACL chair for 2022.

We hope you enjoy the tutorials.

ACL 2022 Tutorial Co-chairs
Luciana Benotti (until Jan 2022)
Naoaki Okazaki
Yves Scherrer
Marcos Zampieri

T1 - A Gentle Introduction to Deep Nets and Opportunities for the Future

Kenneth Church, Valia Kordoni, Gary Marcus, Ernest Davis, Yanjun Ma, and Zeyu Chen

Introductory

Sunday, May 22, 2022 - 14:30-18:00 (Liffey Hall 1)

https://github.com/kwchurch/ACL2022_deepnets_tutorial

The first half of this tutorial will make deep nets more accessible to a broader audience, following "Deep Nets for Poets" and "A Gentle Introduction to Fine-Tuning." We will also introduce GFT (general fine tuning), a little language for fine tuning deep nets with short (one line) programs that are as easy to code as regression in statistics packages such as R using glm (general linear models). Based on the success of these methods on a number of benchmarks, one might come away with the impression that deep nets are all we need. However, we believe the glass is half-full: while there is much that can be done with deep nets, there is always more to do. The second half of this tutorial will discuss some of these opportunities.

Suggested readings:

- Deep Nets for Poets (Church et al., 2021)
- A Gentle Introduction to Fine-Tuning (Church et al., 2021)

Kenneth Church, Fellow, Baidu, USA
email: Kennneth.Ward.Church@gmail.com and kennethchurch@baidu.com
website: https://scholar.google.com/citations?user=E6aqGvYAAAAJ
Kenneth Church has worked on many topics in computational linguistics including: web search, language modeling, text analysis, spelling correction, word-sense disambiguation, terminology, translation, lexicography, compression, speech (recognition, synthesis and diarization), OCR, and more. He was an early advocate of empirical methods, and was a founder of EMNLP. He earned his undergraduate and graduate degrees from MIT, and has worked at AT&T, Microsoft, Hopkins, IBM and Baidu. He was the president of ACL in 2012, and SIGDAT (the group that organizes EMNLP) from 1993 until 2011. He became an AT&T Fellow in 2001 and ACL Fellow in 2015.

Valia Kordoni, Professor, Humboldt-Universitaet zu Berlin, Germany
email: evangelia.kordoni@anglistik.hu-berlin.de
webiste: http://www.lt-innovate.org/directory/contact/kordoni-valia

Valia Kordoni is a faculty member of the Department of English at Humboldt University Berlin. She is an active researcher in Language Technology (LT), Data Science and Artificial Intelligence (AI). Her research interests include multilingual Robust Natural Language Analytics, Computational Semantics, Discourse and Human Cognition Modeling, as well as Machine Learning for the automated acquisition of knowledge, especially concerning multiword units and their impact in Natural Language Processing, spoken and written. She has been the president of the ACL (Association for Computational Linguistics) SIGLEX's (Special Interest Group on Lexicon) MWE (Multiword Expressions) Group. She was the Local Chair of ACL 2016 - The 54th Annual Meeting of the Association for Computational Linguistics. She has coordinated and contributed to many projects funded by the EU, the DFG (Germany), the BMBF (Germany), the DAAD (Germany), as well as the NSF (USA), the latest of those being "TraMOOC: Translation for Massive Open Online Courses", a EU-funded Horizon 2020 collaborative project aiming at providing reliable Neural Machine Translation for Massive Open Online Courses (MOOCs).

Gary Marcus , CEO, Robust.AI
email: gary.marcus@icloud.com
website: http://garymarcus.com/index.html
Gary Marcus is a scientist, best-selling author, and entrepreneur. He is Founder and CEO of Robust.AI, and was Founder and CEO of Geometric Intelligence, a machine learning company acquired by Uber in 2016. He is the author of five books, including *The Algebraic Mind, Kluge, The Birth of the Mind,* and *The New York Times* best seller *Guitar Zero*, as well as editor of *The Future of the Brain* and *The Norton Psychology Reader*. He has published extensively in fields ranging from human and animal behavior to neuroscience, genetics, linguistics, evolutionary psychology and artificial intelligence, often in leading journals such as Science and Nature, and is perhaps the youngest Professor Emeritus at NYU. His newest book, co-authored with Ernest Davis, *Rebooting AI: Building Artificial Intelligence We Can Trust* aims to shake up the field of artificial intelligence.

Ernest Davis , Professor, New York University
email: davise@cims.nyu.edu
website: https://cs.nyu.edu/~davise/
Ernest Davis' research focuses on commonsense reasoning, particularly spatial and physical reasoning. He earned his bachelor's degree in math at MIT and his doctorate at Yale in computer science. In addition to *Rebooting AI: Building Artificial Intelligence We Can Trust* with Gary Marcus, he is the suthor of *Representations of Commonsense Knowledge*, the leading textbook in the area, and the co-editor with Philip Davis of *Mathematics, Substance and Surmise*. He has written articles and book reviews for a general readership on a wide range of topics ranging from history of science to cognitive psychology, to children's literature. He has also authored a collection of light verse, entitled *Verses for the Information Age*.

Yanjun Ma , Senior Director of Deep Learning Platform, Baidu
email: mayanjun02@baidu.com
website: https://scholar.google.com/citations?user=Q0wdSd8AAAAJ
Yanjun Ma is a senior director of deep learning platform at Baidu, overseeing the development of open-source deep learning platform PaddlePaddle. His research covers natural language processing and deep learning, which is widely used in Baidu products. In 2015, Dr. Yanjun Ma received National Technology Advancement Award.

Zeyu Chen, Staff Engineer of Deep Learning Platform, Baidu
email: chenzeyu01@baidu.com
website: https://github.com/ZeyuChen
website: https://scholar.google.com/citations?user=LCzu9MEAAAAJ
Zeyu Chen is a staff engineer of deep learning platform at Baidu, leading the R&D efforts in building open-source deep learning libraries for natural language processing and speech on top of PaddlePaddle, including PaddleNLP, PaddleSpeech and PaddleHub.

T2 - Towards Reproducible Machine Learning Research in Natural Language Processing

Ana Lucic, Maurits Bleeker, Samarth Bhargav, Jessica Zosa Forde, Koustuv Sinha, Jesse Dodge, Sasha Luccioni, and Robert Stojnic

Introductory

Sunday, May 22, 2022 - 09:30-13:00 (Liffey Hall 1)

https://acl-reproducibility-tutorial.github.io

While recent progress in the field of ML has been significant, the reproducibility of these cutting-edge results is often lacking, with many submissions lacking the necessary information in order to ensure subsequent reproducibility. Despite proposals such as the Reproducibility Checklist and reproducibility criteria at several major conferences, the reflex for carrying out research with reproducibility in mind is lacking in the broader ML community. We propose this tutorial as a gentle introduction to ensuring reproducible research in ML, with a specific emphasis on computational linguistics and NLP. We also provide a framework for using reproducibility as a teaching tool in university-level computer science programs.

Ana Lucic, PhD Candidate, University of Amsterdam
email: a.lucic@uva.nl
website: https://a-lucic.github.io
Ana Lucic is a PhD Candidate at the University of Amsterdam. Her work primarily focuses on developing and evaluating methods for explainable machine learning (ML). She co-developed a graduate-level course called *Fairness, Accountability, Confidentiality and Transparency in Artificial Intelligence (FACT-AI)* that is centered around reproducing existing FACT-AI algorithms. Her email is a.lucic@uva.nl.

Maurits Bleeker, PhD Candidate, University of Amsterdam
email : m.j.r.bleeker@uva.nl
website: https://mauritsbleeker.github.io
Maurits Bleeker is PhD Candidate at the University of Amsterdam who co-developed the FACT-AI course. His work is primarily on multi-modal information retrieval (IR) and representation learning. His email is m.j.r.bleeker@uva.nl

Samarth Bhargav, PhD Candidate, University of Amsterdam
email: s.bhargav@uva.nl
website: http://samarthbhargav.github.io/
Samarth Bhargav is a PhD Candidate at the University of Amsterdam. His work mainly focuses on recommender systems and conversational AI. His email is s.bhargav@uva.nl.

Jessica Zosa Forde, PhD Candidate, Brown University
email: jessica_forde@brown.edu
website: https://jzf2101.github.io/
Jessica Zosa Forde is a PhD Candidate at Brown University, studying the science of deep learning. She is a co-organizer of the ML Reproducibility Challenge. Prior to starting her PhD, Jessica was a core maintainer at Project Jupyter, which maintains open source projects for reproducibile science such as the Jupyter Notebook.
Koustuv Sinha, PhD Candidate, McGill University
email: koustuv.sinha@mail.mcgill.ca
website: https://cs.mcgill.ca/~ksinha4
Koustuv Sinha is a PhD Candidate at McGill University/Mila. He has organized the annual ML Reproducibility Challenge five times since 2018 (at ICLR 2018, ICLR 2019, NeurIPS 2019, MLRC 2020, MLRC 2021) and serves as an associate editor of ReScience, a journal promoting reproducibility reports in various fields of science. His email is koustuv.sinha@mail.mcgill.ca

Jesse Dodge, Research Scientist, AllenNLP, Allen Institute of AI
email: dodgejesse@gmail.com
website: https://jessedodge.github.io/
Jesse Dodge is a Research Scientist at AllenNLP, Allen Institute for AI. Jesse created the NLP Reproducibility Checklist, has been an organizer of the ML Reproducibility Challenge (MLRC) 2020 and 2021, will be a Reproducibility Chair at NAACL 2022, and has published numerous papers in top NLP conferences on reproducibility. His email is jessed@allenai.org

Sasha Luccioni, Research Scientist, HuggingFace
email: sasha.luccioni@huggingface.co
website: https://www.sashaluccioni.com/
Sasha Luccioni is a Research Scientist at HuggingFace. She is working on projects that aim to encourage responsible research and development in ML, spanning from data to models. She has been an organizer of the ML Reproducibility Challenge since 2021 and is also an Ethics Co-Chair at NeurIPS 2022. Her email is sasha.luccioni@huggingface.co

Robert Stojnic, Papers with Code, Meta AI
email: rstojnic@fb.com
website: https://paperswithcode.com/
Robert Stojnic is the co-creator of Papers with Code and an Engineering Manager at Meta AI (Facebook AI Research). He is a co-organizator for ML Reproducibility Challenge. His email is rstojnic@fb.com

T3 - Knowledge-Augmented Methods for Natural Language Processing

Chenguang Zhu, Yichong Xu, Xiang Ren, Bill Yuchen Lin, Meng Jiang, and Wenhao Yu

Cutting-edge

Sunday, May 22, 2022 - 09:30-13:00 (Wicklow Hall 2)

https://github.com/zcgzcgzcg1/ACL2022_KnowledgeNLP_Tutorial/

Knowledge in natural language processing (NLP) has been a rising trend especially after the advent of large scale pre-trained models. NLP models with attention to knowledge can i) access unlimited amount of external information; ii) delegate the task of storing knowledge from its parameter space to knowledge sources; iii) obtain up-to-date information; iv) make prediction results more explainable via selected knowledge. In this tutorial, we will introduce the key steps in integrating knowledge into NLP, including knowledge grounding from text, knowledge representation and fusing. In addition, we will introduce recent state-of-the-art applications in fusing knowledge into language understanding, language generation and commonsense reasoning.

Chenguang Zhu, Principal Research Manager, Microsoft Cognitive Services Research Group
Website: https://www.microsoft.com/en-us/research/people/chezhu/
Chenguang Zhu is a Principal Research Manager in Microsoft Cognitive Services Research Group, where he leads the Knowledge & Language Team. His research in NLP covers knowledge graph, text summarization and task-oriented dialogue. He has led teams to achieve first places in multiple NLP competitions, including CommonsenseQA, CommonGen, FEVER, CoQA, ARC and SQuAD v1.0. He holds a Ph.D. degree in Computer Science from Stanford University. He has given talks at Stanford University, Carnegie Mellon University and University of Notre Dame. He has previously been TA for Coursera online class "Automata", giving teaching sessions to 100K international students.

Yichong Xu, Senior Researcher, Microsoft Cognitive Services Research Group
Website: https://xycking.wixsite.com/yichongxu
Yichong Xu is a Senior Researcher in Knowledge & Language Team in Microsoft Cognitive Services Research Group. His research in NLP focuses on using external knowledge to help natural language processing, including question answering, commonsense reasoning, and text summarization. Dr. Xu received his Ph.D. in Machine Learning from Carnegie Mellon University. During his time at CMU, he has been TA for large classes ($>$ 200 students) on machine learning and convex optimization. Dr. Xu has given talks at CMU AI Seminar, as well as in many international conferences including ACL, NeurIPS, ICML, etc.

Xiang Ren, Assistant Professor, University of Southern California
Website: https://shanzhenren.github.io/
Xiang Ren is an assistant professor at the USC Computer Science Department, a Research Team Leader at USC ISI, and the PI of the Intelligence and Knowledge Discovery (INK) Lab at USC. Priorly, he received his Ph.D. in Computer Science from the University of Illinois Urbana-Champaign. Dr. Ren works on knowledge acquisition and reasoning in natural language processing, with focuses on developing human-centered and label-efficient computational methods for building trustworthy NLP systems. Ren publishes over 100 research papers and delivered over 10 tutorials at the top conferences in natural language process, data mining, and artificial intelligence. He received NSF CAREER Award, The Web Conference Best Paper runner-up, ACM SIGKDD Doctoral Dissertation Award, and several research awards from Google, Amazon, JP Morgan, Sony, and Snapchat. He was named Forbes' Asia 30 Under 30 in 2019.

Bill Yuchen Lin, Ph.D Student, University of Southern California
Website: https://yuchenlin.xyz/
Bill Yuchen Lin is a Ph.D. candidate at USC. His research goal is to teach machines to think, talk, and act with commonsense knowledge and commonsense reasoning ability as humans do. Towards this ultimate goal, he has been developing knowledge-augmented reasoning methods (e.g., KagNet, MHGRN, DrFact) and constructing benchmark datasets (e.g., CommonGen, RiddleSense, X-CSR) that require commonsense knowledge and complex reasoning for both NLU and NLG. He initiated an online compendium of commonsense reasoning research, which serves as a portal site[1] for the community.

Meng Jiang, Assistant Professor, University of Notre Dame
Website: http://www.meng-jiang.com/
Meng Jiang is an assistant professor at the Department of Computer Science and Engineering in the University of Notre Dame. He obtained his bachelor degree and PhD from Tsinghua University. His research interests include data mining, machine learning, and natural language processing. He has published more than 100 peer-reviewed papers of these topics. He is the recipient of Notre Dame International Faculty Research Award. The honors and awards he received include best paper finalist in KDD 2014, best paper award in KDD-DLG workshop 2020, and ACM SIGSOFT Distinguished Paper Award in ICSE 2021. He received NSF CRII award in 2019 and CAREER award in 2022.

Wenhao Yu, Ph.D. Student, University of Notre Dame
Website: https://wyu97.github.io/
Wenhao Yu is a Ph.D. student in the Department of Computer Science and Engineering at the University of Notre Dame. His research lies in controllable knowledge-driven natural language processing, particularly in natural language generation. His research has been published in top-ranked NLP and data mining conferences such as ACL, EMNLP, KDD and WWW.

[1] https://commonsense.run/

T4 - Non-Autoregressive Sequence Generation

Jiatao Gu and Xu Tan
Cutting-edge
Sunday, May 22, 2022 - 14:30-18:00 (Liffey Hall 2)

https://nar-tutorial.github.io/acl2022/

Non-autoregressive sequence generation (NAR) attempts to generate the entire or partial output sequences in parallel to speed up the generation process and avoid potential issues (e.g., label bias, exposure bias) in autoregressive generation. While it has received much research attention and has been applied in many sequence generation tasks in natural language and speech, naive NAR models still face many challenges to close the performance gap between state-of-the-art autoregressive models because of a lack of modeling power. In this tutorial, we will provide a thorough introduction and review of non-autoregressive sequence generation, in four sections: 1) Background, which covers the motivation of NAR generation, the problem definition, the evaluation protocol, and the comparison with standard autoregressive generation approaches. 2) Method, which includes different aspects: model architecture, objective function, training data, learning paradigm, and additional inference tricks. 3) Application, which covers different tasks in text and speech generation, and some advanced topics in applications. 4) Conclusion, in which we describe several research challenges and discuss the potential future research directions. We hope this tutorial can serve both academic researchers and industry practitioners working on non-autoregressive sequence generation.

Jiatao Gu, Research Scientist, Facebook AI Research (FAIR)
email: jgu@fb.com
website: https://jiataogu.me/
Dr. Jiatao Gu is a Research Scientist at Facebook AI Research (FAIR). Jiatao received his Ph.D. degree in 2018 from the University of Hong Kong and B.Eng from Tsinghua University in 2014. His research interests cover representation learning and generative models and their applications on NLP, speech, computer vision, and multi-modal learning. Particularly, his research focuses on developing efficient learning and inference algorithms and applying them successfully to neural machine translation and 3D-aware image synthesis. He has over 40 papers published at top-tier conferences and journals, including ACL, EMNLP, NeurIPS, ICLR, and TACL. Jiao has also served as an area chair for several top conferences. Jiatao has rich research experience on the topic of non-autoregressive sequence generation. He published the first of its kind paper for non-autoregressive neural machine translation in 2018 and has led the following exploration and extensions.

Xu Tan, Senior Researcher, Microsoft Research Asia (MSRA)
email: xuta@microsoft.com
website: https://www.microsoft.com/en-us/research/people/xuta/
Xu Tan is a Senior Researcher at Microsoft Research Asia (MSRA). His research interests cover deep learning and its applications in language/speech/music, including neural machine translation, text to speech, automatic speech recognition, pre-training, music generation, etc. The machine translation systems have achieved human parity on Chinese-English news translation in 2018 and won several champions on WMT machine translation competition in 2019. He has designed several popular language/speech/music models, and systems (e.g., MASS, FastSpeech, and Muzic) and has transferred many research works to the products in Microsoft (e.g., Azure, Bing). He has rich research experiences on non-autoregressive sequence generation and has designed several models such as FastCorrect 1/2, FastSpeech 1/2.

T5 - Learning with Limited Text Data

Diyi Yang, Ankur P Parikh, and Colin Raffel

Cutting-edge

Sunday, May 22, 2022 - 09:30-13:00 (The Liffey A)

https://github.com/diyiy/ACL2022_Limited_Data_Learning_Tutorial

Natural Language Processing (NLP) has achieved great progress in the past decade on the basis of neural models, which often make use of large amounts of labeled data to achieve state-of-the-art performance. The dependence on labeled data prevents NLP models from being applied to low-resource settings and languages because of the time, money, and expertise that is often required to label massive amounts of textual data. Consequently, the ability to learn with limited labeled data is crucial for deploying neural systems to real-world NLP applications. Recently, numerous approaches have been explored to alleviate the need for labeled data in NLP such as data augmentation and semi-supervised learning. This tutorial aims to provide a systematic and up-to-date overview of these methods in order to help researchers and practitioners understand the landscape of approaches and the challenges associated with learning from limited labeled data, an emerging topic in the computational linguistics community. We will consider applications to a wide variety of NLP tasks (including text classification, generation, and structured prediction) and will highlight current challenges and future directions.

Diyi Yang, Assistant Professor, Georgia Institute of Technology
email: diyi.yang@cc.gatech.edu
website: https://faculty.cc.gatech.edu/ dyang888/
Diyi Yang is an assistant professor at the School of Interactive Computing, Georgia Tech. Her research focuses on learning with limited and noisy text data, user-centric language generation, and computational social science. Diyi has organized four workshops at NLP conferences: Widening NLP Workshops at NAACL 2018 and ACL 2019, Casual Inference workshop at EMNLP 2021, and NLG Evaluation workshop at EMNLP 2021. Diyi has served as area chairs and senior area chairs for ACL, NAACL and EMNLP.

Ankur Parikh, Senior Research Scientist, Google Research
email: aparikh@google.com
website: http://www.ankurparikh.com/
Ankur Parikh is a senior research scientist at Google NYC and adjunct assistant professor at NYU. His research interests are in natural language processing and machine learning with a recent focus on high precision text generation. Ankur received his PhD from Carnegie Mellon in 2015 and has received a best paper runner up award at EMNLP 2014 and a best paper in translational bioinformatics at ISMB 2011. He has taught natural language processing at NYU since 2017.

Colin Raffel, Assistant Professor, University of North Carolina, Chapel Hill and Hugging Face
email: craffel@gmail.com
website: https://colinraffel.com/
Colin Raffel is an assistant professor of Computer Science at the University of North Carolina, Chapel Hill. His research is focused on machine learning algorithms for learning from limited labeled data, including semi-supervised, unsupervised, and transfer learning methods. His best-known work on the topics related to this tutorial include the T5 model and the MixMatch,ReMixMatch, FixMatch series of semi-supervised learning algorithms. He gave a tutorial at the 2017 International Society for Music Information Retrieval Conference[2] and has taught machine learning courses at UNC, Columbia University, and Google's TechExchange program.

[2] https://colinraffel.com/talks/ismir2017leveraging.pdf

T6 - Zero- and Few-Shot NLP with Pretrained Language Models

Iz Beltagy, Arman Cohan, Robert L. Logan IV, Sewon Min, and Sameer Singh

Cutting-edge

Sunday, May 22, 2022 - 14:30-18:00 (The Liffey A)

https://github.com/allenai/acl2022-zerofewshot-tutorial

The ability to efficiently learn from little-to-no data is critical to applying NLP to tasks where data collection is costly or otherwise difficult. This is a challenging setting both academically and practically—particularly because training neutral models typically require large amount of labeled data. More recently, advances in pretraining on unlabelled data have brought up the potential of better zero-shot or few-shot learning (Devlin et al., 2019; Brown et al., 2020). In particular, over the past year, a great deal of research has been conducted to better learn from limited data using large-scale language models. In this tutorial, we aim at bringing interested NLP researchers up to speed about the recent and ongoing techniques for zero- and few-shot learning with pretrained language models. Additionally, our goal is to reveal new research opportunities to the audience, which will hopefully bring us closer to address existing challenges in this domain.

Suggested readings:

- Language Models are Few-Shot Learners (Brown et al., 2020)
- It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners (Schick and Schütze, 2021)
- Multitask Prompted Training Enables Zero-Shot Task Generalization (Sanh et al., 2021)
- FLEX: Unifying Evaluation for Few-Shot NLP (Bragg et al., 2021)

Iz Beltagy, Research Scientist, Allen Institute for AI
email: beltagy@allenai.org
website: beltagy.net
Iz Beltagy is a Research Scientist at AI2 focusing on language modeling, transfer learning, summarization, explainability and efficiency. His research has been recognized with a best paper honorary mention at ACL 2020 and an outstanding paper award at AKBC 2021. He was a co-instructor of the tutorial on "Beyond Paragraphs: NLP for Long Sequences" (NAACL-HLT 2021). He worked as a Teaching Assistant at the University of Texas at Austin teaching computer science.

Arman Cohan, Research Scientist, Allen Institute for AI, and Affiliate Assistant Professor, University of Washington
email: armanc@allenai.org
website: armancohan.com
Arman Cohan is a Research Scientist at AI2 and an Affiliate Assistant Professor at University of Washington, focusing on representation learning and transfer learning methods, as well as NLP applications in specialized domains and scientific text. His research has been recognized with a best paper award at EMNLP 2017, an honorable mention at COLING 2018, and Harold N. Glassman Distinguished Doctoral Dissertation award in 2019. He was a co-instructor of the tutorial on "Beyond Paragraphs: NLP for Long Sequences" (NAACL-HLT 2021).

Robert L. Logan IV, Ph.D. Student, University of California, Irvine
email: rlogan@uci.edu
website: rloganiv.github.io
Robert L. Logan IV is a Ph.D. student at the University of California, Irvine, advised by Sameer Singh and Padhraic Smyth. His research focuses on problems at the intersection of information extraction and language modeling, and encompasses recently published work on language model prompting that is relevant to this proposal. He has presented invited talks at the SoCal NLP Symposium (2019), the CHASE-CI Workshop (2019), and the UCI Center for Machine Learning Seminar (2021).

Sewon Min, Ph.D. Student, University of Washington
email: sewon@cs.washington.edu
website: shmsw25.github.io
Sewon Min is a Ph.D. student in the Paul G. Allen School of Computer Science & Engineering at the University of Washington, advised by Hannaneh Hajishirzi and Luke Zettlemoyer. Her research focuses on natural language understanding, question answering, and knowledge representation. She was a co-instructor of the tutorial on "Beyond Paragraphs: NLP for Long Sequences" (NAACL-HLT 2021), and has co-organized multiple workshops at ACL, EMNLP, NeurIPS and AKBC, including a workshop on Machine Reading for Question Answering, a competition on Efficient Open-domain Question Answering, a workshop on Representation Learning for NLP, workshop on Semiparametric Methods in NLP.

Sameer Singh, Associate Professor, University of California, Irvine and Allen AI Fellow, Allen Institute for AI
email: sameer@uci.edu
website: sameersingh.org
Sameer Singh is an Associate Professor of Computer Science at the University of California, Irvine and an Allen AI Fellow at the Allen Institute for AI. He is working on large-scale and interpretable machine learning models for NLP. His work has received paper awards at ACL 2020, AKBC 2020, EMNLP 2019, ACL 2018, and KDD 2016. Sameer has presented a number of tutorials, many relevant to this proposal, such as "Deep Adversarial Learning" (NAACL 2019), "Mining Knowledge Graphs from Text" (WSDM 2018 and AAAI 2017), "Interpretability and Explanations" (NeurIPS 2020 and EMNLP 2020), and "Robustness in NLP" (EMNLP 2021). Sameer has also received teaching awards at UCI.

T7 - Vision-Language Pretraining: Current Trends and the Future

Aishwarya Agrawal, Damien Teney, and Aida Nematzadeh

Cutting-edge

Sunday, May 22, 2022 - 14:30-18:00 (Wicklow Hall 2)

https://vlp-tutorial-acl2022.github.io/

In the last few years, there has been an increased interest in building multimodal (vision-language) models that are pretrained on larger but noisier datasets where the two modalities (e.g., image and text) loosely correspond to each other (e.g., Lu et al., 2019; Radford et al., 2021). Given a task (such as visual question answering), these models are then often fine-tuned on task-specific supervised datasets. (e.g., Lu et al., 2019; Chen et al.,2020; Tan and Bansal, 2019; Li et al., 2020a,b). In addition to the larger pretraining datasets, the transformer architecture (Vaswani et al., 2017) and in particular self-attention applied to two modalities are responsible for the impressive performance of the recent pretrained models on downstream tasks (Hendricks et al., 2021).
In this tutorial, we focus on recent vision-language pretraining paradigms. Our goal is to first provide the background on image–language datasets, benchmarks, and modeling innovations before the multimodal pretraining area. Next we discuss the different family of models used for vision-language pretraining, highlighting their strengths and shortcomings. Finally, we discuss the limits of vision-language pretraining through statistical learning, and the need for alternative approaches such as causal representation learning.

Aishwarya Agrawal, Assistant Professor, University of Montreal and Mila, and Research Scientist, DeepMind
email: aishwarya.agrawal@mila.quebec
website: https://www.iro.umontreal.ca/~agrawal/
Aishwarya Agrawal is an Assistant Professor in the Department of Computer Science and Operations Research at the University of Montreal. She is also a Canada CIFAR AI Chair and a core academic member of Mila – Quebec AI Institute. She also spends one day a week at DeepMind as a Research Scientist. Aishwarya's research interests lie at the intersection of computer vision, deep learning and natural language processing. Aishwarya is one of the two lead authors on the VQA paper that introduced the task and the VQA v1.0 dataset. She has played an active role in releasing the dataset to the public. She is, in particular, keen about building vision-language models that generalize to out-of-distribution datasets. She used to co-organize the annual VQA challenge and workshop, and has given numerous invited talks (see https://www.iro.umontreal.ca/~agrawal/index.html#talks).

Damien Teney, Research Scientist, Idiap Research Institute
email: damien.teney@idiap.ch

website: http://damienteney.info
Damien Teney is a research scientist heading the machine learning group at the Idiap Research Institute in Switzerland. He is known for his work at the intersection of computer vision, machine learning, and natural language processing. He was part of the team that won the Visual Question Answering Challenge at CVPR 2017, which introduced the bottom-up/top-down attention mechanisms that are now ubiquitous for vision and language. His current research focuses on out-of-distribution generalization and learning methods inspired by causal reasoning. He has given multiple introductory talks on these topics and is a regular invited speaker at workshops and seminars on vision and language (e.g., https://visualqa.org/workshop.html and http://qi-wu.me/accv_v2l/invited_people.html).

Aida Nematzadeh, Staff Research Scientist, DeepMind
email: nematzadeh@deepmind.com
website: http://www.aidanematzadeh.me
Aida Nematzadeh is a staff research scientist at DeepMind. Her research interests are in the intersection of computational linguistics, cognitive science, and machine learning. Her recent work has focused on multimodal learning and evaluation and analysis of neural representations. She co-instructed a tutorial on "Language Learning and Processing in People and Machines" at NAACL 2019, and has given numerous invited talks (see http://aidanematzadeh.me/talks.html).

T8 - Natural Language Processing for Multilingual Task-Oriented Dialogue

Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Edoardo Ponti, and Ivan Vulić

Cutting-edge

Sunday, May 22, 2022 - 09:30-13:00 (Liffey Hall 2)

https://github.com/cambridgeltl/ACL2022_tutorial_multilingual_dialogue

Recent advances in deep learning have also enabled fast progress in the research of task-oriented dialogue (ToD) systems. However, the majority of ToD systems are developed for English and merely a handful of other widely spoken languages, e.g., Chinese and German. This hugely limits the global reach and, consequently, transformative socioeconomic potential of such systems. In this tutorial, we will thus discuss and demonstrate the importance of (building) multilingual ToD systems, and then provide a systematic overview of current research gaps, challenges and initiatives related to multilingual ToD systems, with a particular focus on their connections to current research and challenges in multilingual and low-resource NLP. The tutorial will aim to provide answers or shed new light to the following questions: a) Why are multilingual dialogue systems so hard to build: what makes multilinguality for dialogue more challenging than for other NLP applications and tasks? b) What are the best existing methods and datasets for multilingual and cross-lingual (task-oriented) dialog systems? How are (multilingual) ToD systems usually evaluated? c) What are the promising future directions for multilingual ToD research: where can one draw inspiration from related NLP areas and tasks?

Evgeniia Razumovskaia, PhD Student, University of Cambridge
email: er563@cam.ac.uk
website: https://evgeniiaraz.github.io
Evgeniia Razumovskaia is a PhD student in the Language Technology Lab at the University of Cambridge, advised by Prof. Anna Korhonen and Dr. Ivan Vulic. She works on dialogue systems, focusing on efficient few-shot methods for multilingual dialogue.

Goran Glavaš, Full Professor (Chair for Natural Language Processing) and member of the Center for Artificial Intelligence and Data Science (CAIDAS), University of Würzburg
email: goran@informatik.uni-mannheim.de
website: https://sites.google.com/view/goranglavas
Goran Glavaš is a Full Professor (Chair for Natural Language Processing) and member of the Center for Artificial Intelligence and Data Science (CAIDAS) at the University of Würzburg. His research focuses on multilingual representation learning and cross-lingual transfer (primarily for low-resource languages), fair

and sustainable NLP, and NLP applications for social sciences and humanities. He has given tutorials at ACL 2019 and EMNLP 2019, organized workshops TextGraphs and SustainNLP, and served as reviewer and (senior) area chair for a number of *ACL events. He currently serves as an Editor-in-Chief for the ACL Rolling Review.

Olga Majewska, Research Scientist, Amazon Alexa, and Affiliated Researcher, University of Cambridge
email: om304@cam.ac.uk
website: https://om304.github.io
Olga Majewska has recently joined Amazon Alexa in Cambridge, UK, while remaining an affiliated researcher at the Language Technology Lab, University of Cambridge, where she earned her PhD in computational linguistics in 2021. Her interests lie, among others, in multilingual expansion of conversational AI and development of efficient protocols for generation of task-oriented dialogue evaluation data for under-resourced languages.

Edoardo Maria Ponti, Visiting Postdoctoral Scholar, University of Stanford, and Postdoctoral Fellow, MILA Montreal
email: edoardo-maria.ponti@mila.quebec
website: https://ducdauge.github.io
Edoardo Maria Ponti is a Visiting Postdoctoral Scholar at the University of Stanford and a Postdoctoral Fellow at MILA Montreal. He works on sample efficiency and modularity in neural networks, with applications to multilingual NLP. In 2020, he obtained a PhD in computational linguistics from the University of Cambridge, St John's College. Previously, he interned as an AI/ML researcher at Apple in Cupertino. His research earned him a Google Research Faculty Award and an ERC Proof of Concept grant. He received 2 Best Paper Awards at EMNLP 2021 and RepL4NLP 2019.

Ivan Vulić, Senior Research Associate, University of Cambridge, and Senior Scientist, PolyAI
email: iv250@cam.ac.uk
website: https://sites.google.com/site/ivanvulic
Ivan Vulić is a Senior Research Associate 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., EMNLP 2017, NAACL-HLT 2018, ESSLLI 2018, ACL 2019, EMNLP 2019, AILC Lectures 2021) and workshops in areas relevant to this proposal (e.g., SIGTYP, DeeLIO, RepL4NLP, PC of *SEM 2021). For his contributions to NLP and IR, he obtained the 2021 Karen Spärck Jones award.

7

Main Conference

Main Conference Program (Overview)

Main Conference Program (Overview): Day 1

8:00- **Registration**

8:30-9:30 *Opening Remarks and Presidential Address (Room: Auditorium)*

9:30-10:30 *Keynote 1 - Angela D. Friederici: "Language in the Human Brain" (Room: Auditorium)*

10:30-11:00 Coffee break (Room: Forum)

11:00-12:30 **Session 1**

Machine Learning for NLP 1 *Room: The Liffey B*	**Machine Translation and Multilinguality 1** *Room: The Liffey A*
Resources and Evaluation 1 *Room: Liffey Hall 2*	**Semantics 1** *Room: Wicklow Hall 2a*
Information Extraction 1 *Room: Wicklow Hall 2b*	**Language Grounding, Speech and Multimodality 1** *Room: Liffey Hall 1*
Question Answering 1 *Room: Wicklow Hall 1*	**Poster+Demo Session 1** *Room: Forum*

12:30-14:00 Lunch break

14:00-15:00 **Session 2**

Dialogue and Interactive Systems 1 *Room: The Liffey B*	**Ethics in NLP** *Room: The Liffey A*
Special Theme 1 *Room: Liffey Hall 2*	**Discourse and Pragmatics** *Room: Wicklow Hall 2a*
Sentiment Analysis, Stylistic Analysis, and Argument Mining 1 *Room: Wicklow Hall 2b*	**Information Retrieval and Text Mining** *Room: Liffey Hall 1*
Phonology, Morphology and Word Segmentation *Room: Wicklow Hall 1*	**Poster+Demo Session 2** *Room: Forum*

15:00-15:15 Mini break (Room: Forum)

15:15-16:30 *Spotlight Talks by Young Rising Stars (STIRS) (Room: Auditorium)*

16:30-17:00 Coffee break (Room: Forum)

17:00-18:00 **Session 3**

Machine Learning for NLP 2 *Room: The Liffey B*	**Machine Translation and Multilinguality 2** *Room: The Liffey A*
Interpretability and Analysis of Models for NLP 1 *Room: Liffey Hall 2*	**NLP Applications 1** *Room: Wicklow Hall 2a*
Syntax: Tagging, Chunking and Parsing *Room: Wicklow Hall 2b*	**Student Research Workshop** *Room: Liffey Hall 1*
Generation 1 *Room: Wicklow Hall 1*	**Poster+Demo Session 3** *Room: Forum*

19:30-23:00 *Social Event: Guinness Storehouse*

Main Conference Program (Overview): Day 2

7:30-8:30 *Virtual poster session 1 (GatherTown)*

8:00- **Registration**

9:00-10:30 *Next Big Ideas Talks (Room: Auditorium)*

10:30-11:00 Coffee break (Room: Forum)

11:00-12:30 **Session 4**

Machine Learning for NLP 3 *Room: The Liffey B*	**Interpretability and Analysis of Models for NLP 2** *Room: The Liffey A*
Special Theme 2 *Room: Liffey Hall 2*	**NLP Applications 2** *Room: Wicklow Hall 2a*
Dialogue and Interactive Systems 2 *Room: Wicklow Hall 2b*	**Resources and Evaluation 2** *Room: Liffey Hall 1*
Summarization 1 *Room: Wicklow Hall 1*	**Poster+Demo Session 4** *Room: Forum*

12:30-13:30 Lunch break

13:30-15:00 *ACL Business Meeting + Panel on the Future of Reviewing in NLP*

15:00-15:15 Mini break (Room: Forum)

15:15-16:15 **Session 5**

Machine Learning for NLP 4 *Room: The Liffey B*	**Machine Translation and Multilinguality 3** *Room: The Liffey A*
Resources and Evaluation 3 *Room: Liffey Hall 2*	**Semantics 2** *Room: Wicklow Hall 2a*
Information Extraction 2 *Room: Wicklow Hall 2b*	**Language Grounding, Speech and Multimodality 2** *Room: Liffey Hall 1*
Question Answering 2 *Room: Wicklow Hall 1*	**Poster+Demo Session 5** *Room: Forum*

16:15-16:45 Coffee break (Room Forum)

16:45-18:30 *Keynote 2: Fire-Side Chat with Barbara Grosz and Yejin Choi: "The Trajectory of ACL and the Next 60 Years" (Room: Auditorium)*

19:00-20:00 *Virtual poster session 2 (GatherTown)*

Main Conference Program (Overview): Day 3

7:30-8:30 *Virtual poster session 3 (GatherTown)*

8:00- **Registration**

9:00-10:15 *Keynote 3: Panel on "Supporting Linguistic Diversity" (chaired by Steven Bird) (Room: Auditorium)*

10:15-10:45 Coffee break (Room: Forum)

10:45-12:15 **Session 6**

Dialogue and Interactive Systems 3 *Room: The Liffey B*	**Interpretability and Analysis of Models for NLP 3** *Room: The Liffey A*
NLP Applications 3 *Room: Liffey Hall 2*	**Semantics 3** *Room: Wicklow Hall 2a*
Linguistic Theories, Cognitive Modeling and Psycholinguistics *Room: Wicklow Hall 2b*	**Computational Social Science and Cultural Analytics** *Room: Liffey Hall 1*
Generation 2 *Room: Wicklow Hall 1*	**Poster+Demo Session 6** *Room: Forum*

12:15-13:30 Lunch break

13:30-14:30 **Session 7**

Machine Learning for NLP 5 *Room: The Liffey B*	**Machine Translation and Multilinguality 4** **Room: The Liffey A**
Resources and Evaluation 4 *Room: Liffey Hall 2*	**Summarization 2** *Room: Wicklow Hall 2a*
Sentiment Analysis, Stylistic Analysis, and Argument Mining 2 *Room: Wicklow Hall 2b*	**Language Grounding, Speech and Multimodality 3** *Room: Liffey Hall 1*
Question Answering 3 *Room: Wicklow Hall 1*	**Poster+Demo Session 7** *Room: Forum*

14:30-14:45 Mini break (Room: Forum)

14:45-16:00 *Best Paper Awards (Room: Auditorium)*

16:00-16:30 Coffee break (Room: Forum)

16:30-18:00 *ACL Awards (Room: Auditorium)*

18:00-18:30 **Closing Session (Room: Auditorium)**

19:00-20:00 *Virtual poster session 4 (GatherTown)*

19:00 Socials and Company-Sponsored Activities

The list of contributions presented in the Poster and Demo sessions are available in the **"Posters and Demos Guide".

Main Conference: Monday, May 23, 2022

Opening Remarks and Presidential Address
08:30-09:30 - **Auditorium** (Auditorium)

Keynote 1 - Angela D. Friederici: "Language in the Human Brain"
09:30-10:30 - **Auditorium** (Auditorium)

Coffee Break
10:30-11:00 - **Auditorium** (Forum)

Session 1 - 11:00-12:30

Question Answering 1

11:00-12:30 (Wicklow Hall 1)

11:00-11:15 (Wicklow Hall 1)

Turning Tables: Generating Examples from Semi-structured Tables for Endowing Language Models with Reasoning Skills

Ori Yoran, Alon Talmor and Jonathan Berant

Models pre-trained with a language modeling objective possess ample world knowledge and language skills, but are known to struggle in tasks that require reasoning. In this work, we propose to leverage semi-structured tables, and automatically generate at scale question-paragraph pairs, where answering the question requires reasoning over multiple facts in the paragraph. We add a pre-training step over this synthetic data, which includes examples that require 16 different reasoning skills such as number comparison, conjunction, and fact composition. To improve data efficiency, we sample examples from reasoning skills where the model currently errs. We evaluate our approach on three reasoning-focused reading comprehension datasets, and show that our model, PReasM, substantially outperforms T5, a popular pre-trained encoder-decoder model. Moreover, sampling examples based on model errors leads to faster training and higher performance.

11:15-11:30 (Wicklow Hall 1)

ConditionalQA: A Complex Reading Comprehension Dataset with Conditional Answers

Haitian Sun, William W. Cohen and Ruslan Salakhutdinov

We describe a Question Answering (QA) dataset that contains complex questions with conditional answers, i.e. the answers are only applicable when certain conditions apply. We call this dataset ConditionalQA. In addition to conditional answers, the dataset also features: (1) long context documents with information that is related in logically complex ways; (2) multi-hop questions that require compositional logical reasoning; (3) a combination of extractive questions, yes/no questions, questions with multiple answers, and not-answerable questions; (4) questions asked without knowing the answers. We show that ConditionalQA is challenging for many of the existing QA models, especially in selecting answer conditions. We believe that this dataset will motivate further research in answering complex questions over long documents.

11:30-11:45 (Wicklow Hall 1)

Retrieval-guided Counterfactual Generation for QA

Bhargavi Paranjape, Matthew Lamm and Ian Tenney

Deep NLP models have been shown to be brittle to input perturbations. Recent work has shown that data augmentation using counterfactuals — i.e. minimally perturbed inputs — can help ameliorate this weakness. We focus on the task of creating counterfactuals for question answering, which presents unique challenges related to world knowledge, semantic diversity, and answerability. To address these challenges, we develop a Retrieve-Generate-Filter(RGF) technique to create counterfactual evaluation and training data with minimal human supervision. Using an open-domain QA framework and question generation model trained on original task data, we create counterfactuals that are fluent, semantically diverse, and automatically labeled. Data augmentation with RGF counterfactuals improves performance on out-of-domain and challenging evaluation sets over and above existing methods, in both the reading comprehension and open-domain QA settings. Moreover,

we find that RGF data leads to significant improvements in a model's robustness to local perturbations.

11:45-12:00 (Wicklow Hall 1)

Simulating Bandit Learning from User Feedback for Extractive Question Answering

Ge Gao, Eunsol Choi and Yoav Artzi

We study learning from user feedback for extractive question answering by simulating feedback using supervised data. We cast the problem as contextual bandit learning, and analyze the characteristics of several learning scenarios with focus on reducing data annotation. We show that systems initially trained on few examples can dramatically improve given feedback from users on model-predicted answers, and that one can use existing datasets to deploy systems in new domains without any annotation effort, but instead improving the system on-the-fly via user feedback.

12:00-12:15 (Wicklow Hall 1)

Open Domain Question Answering with A Unified Knowledge Interface

Kaixin Ma, Hao Cheng, Xiaodong Liu, Eric Nyberg and Jianfeng Gao

The retriever-reader framework is popular for open-domain question answering (ODQA) due to its ability to use explicit knowledge. Although prior work has sought to increase the knowledge coverage by incorporating structured knowledge beyond text, accessing heterogeneous knowledge sources through a unified interface remains an open question. While data-to-text generation has the potential to serve as a universal interface for data and text, its feasibility for downstream tasks remains largely unknown. In this work, we bridge this gap and use the data-to-text method as a means for encoding structured knowledge for open-domain question answering. Specifically, we propose a verbalizer-retriever-reader framework for ODQA over data and text where verbalized tables from Wikipedia and graphs from Wikidata are used as augmented knowledge sources. We show that our Unified Data and Text QA, UDT-QA, can effectively benefit from the expanded knowledge index, leading to large gains over text-only baselines. Notably, our approach sets the single-model state-of-the-art on Natural Questions. Furthermore, our analyses indicate that verbalized knowledge is preferred for answer reasoning for both adapted and hot-swap settings.

12:15-12:30 (Wicklow Hall 1)

[TACL] **Break, Perturb, Build: Automatic Perturbation of Reasoning Paths Through Question Decomposition**

Mor Geva, Tomer Wolfson and Jonathan Berant

Semantics 1

11:00-12:30 (Wicklow Hall 2a)

11:00-11:15 (Wicklow Hall 2a)

Just Rank: Rethinking Evaluation with Word and Sentence Similarities

Bin Wang, C.-c. Jay Kuo and Haizhou Li

Word and sentence embeddings are useful feature representations in natural language processing. However, intrinsic evaluation for embeddings lags far behind, and there has been no significant update since the past decade. Word and sentence similarity tasks have become the de facto evaluation method. It leads models to overfit to such evaluations, negatively impacting embedding models' development. This paper first points out the problems using semantic similarity as the gold standard for word and sentence embedding evaluations. Further, we propose a new intrinsic evaluation method called EvalRank, which shows a much stronger correlation with downstream tasks. Extensive experiments are conducted based on 60+ models and popular datasets to certify our judgments. Finally, the practical evaluation toolkit is released for future benchmarking purposes.

11:15-11:30 (Wicklow Hall 2a)

Improving Word Translation via Two-Stage Contrastive Learning

Yaoyiran Li, Fangyu Liu, Nigel Collier, Anna Korhonen and Ivan Vulić

Word translation or bilingual lexicon induction (BLI) is a key cross-lingual task, aiming to bridge the lexical gap between different languages. In this work, we propose a robust and effective two-stage contrastive learning framework for the BLI task. At Stage C1, we propose to refine standard cross-lingual linear maps between static word embeddings (WEs) via a contrastive learning objective; we also show how to integrate it into the self-learning procedure for even more refined cross-lingual maps. In Stage C2, we conduct BLI-oriented contrastive fine-tuning of mBERT, unlocking its word translation capability. We also show that static WEs induced from the 'C2-tuned' mBERT complement static WEs from Stage C1. Comprehensive experiments on standard BLI datasets for diverse languages and different experimental setups demonstrate substantial gains achieved by our framework. While the BLI method from Stage C1 already yields substantial gains over all state-of-the-art BLI methods in our comparison, even stronger improvements are met with the full two-stage framework: e.g., we report gains for 112/112 BLI setups, spanning 28 language pairs.

11:30-11:45 (Wicklow Hall 2a)

Leveraging Similar Users for Personalized Language Modeling with Limited Data

Charles Welch, Chenxi Gu, Jonathan K Kummerfeld, Veronica Perez-Rosas and Rada Mihalcea
Personalized language models are designed and trained to capture language patterns specific to individual users. This makes them more accurate at predicting what a user will write. However, when a new user joins a platform and not enough text is available, it is harder to build effective personalized language models. We propose a solution for this problem, using a model trained on users that are similar to a new user. In this paper, we explore strategies for finding the similarity between new users and existing ones and methods for using the data from existing users who are a good match. We further explore the trade-off between available data for new users and how well their language can be modeled.

11:45-12:00 (Wicklow Hall 2a)
Predicate-Argument Based Bi-Encoder for Paraphrase Identification
Qiwei Peng, David Weir, Julie Weeds and Yekun Chai
Paraphrase identification involves identifying whether a pair of sentences express the same or similar meanings. While cross-encoders have achieved high performances across several benchmarks, bi-encoders such as SBERT have been widely applied to sentence pair tasks. They exhibit substantially lower computation complexity and are better suited to symmetric tasks. In this work, we adopt a bi-encoder approach to the paraphrase identification task, and investigate the impact of explicitly incorporating predicate-argument information into SBERT through weighted aggregation. Experiments on six paraphrase identification datasets demonstrate that, with a minimal increase in parameters, the proposed model is able to outperform SBERT/SRoBERTa significantly. Further, ablation studies reveal that the predicate-argument based component plays a significant role in the performance gain.

12:00-12:15 (Wicklow Hall 2a)
[TACL] **It's not Rocket Science: Interpreting Figurative Language in Narratives**
Tuhin Chakrabarty, Yejin Choi and Vered Shwartz

12:15-12:25 (Wicklow Hall 2a)
Exploiting Language Model Prompts Using Similarity Measures: A Case Study on the Word-in-Context Task
Mohsen Tabasi, Kiamehr Rezaee and Mohammad Taher Pilehvar
As a recent development in few-shot learning, prompt-based techniques have demonstrated promising potential in a variety of natural language processing tasks. However, despite proving competitive on most tasks in the GLUE and SuperGLUE benchmarks, existing prompt-based techniques fail on the semantic distinction task of the Word-in-Context (WiC) dataset. Specifically, none of the existing few-shot approaches (including the in-context learning of GPT-3) can attain a performance that is meaningfully different from the random baseline. Trying to fill this gap, we propose a new prompting technique, based on similarity metrics, which boosts few-shot performance to the level of fully supervised methods. Our simple adaptation shows that the failure of existing prompt-based techniques in semantic distinction is due to their improper configuration, rather than lack of relevant knowledge in the representations. We also show that this approach can be effectively extended to other downstream tasks for which a single prompt is sufficient.

Machine Learning for NLP 1

11:00-12:30 (The Liffey B)

11:00-11:15 (The Liffey B)
Label Semantic Aware Pre-training for Few-shot Text Classification
Aaron Mueller, Jason Krone, Salvatore Romeo, Saab Mansour, Elman Mansimov, Yi Zhang and Dan Roth
In text classification tasks, useful information is encoded in the label names. Label semantic aware systems have leveraged this information for improved text classification performance during fine-tuning and prediction. However, use of label-semantics during pre-training has not been extensively explored. We therefore propose Label Semantic Aware Pre-training (LSAP) to improve the generalization and data efficiency of text classification systems. LSAP incorporates label semantics into pre-trained generative models (T5 in our case) by performing secondary pre-training on labeled sentences from a variety of domains. As domain-general pre-training requires large amounts of data, we develop a filtering and labeling pipeline to automatically create sentence-label pairs from unlabeled text. We perform experiments on intent (ATIS, Snips, TOPv2) and topic classification (AG News, Yahoo! Answers). LSAP obtains significant accuracy improvements over state-of-the-art models for few-shot text classification while maintaining performance comparable to state of the art in high-resource settings.

11:15-11:30 (The Liffey B)
mLUKE: The Power of Entity Representations in Multilingual Pretrained Language Models
Ryokan Ri, Ikuya Yamada and Yoshimasa Tsuruoka
Recent studies have shown that multilingual pretrained language models can be effectively improved with cross-lingual alignment information from Wikipedia entities. However, existing methods only exploit entity information in pretraining and do not explicitly use entities in downstream tasks. In this study, we explore the effectiveness of leveraging entity representations for downstream cross-lingual tasks. We train a multilingual language model with 24 languages with entity representations and show the model consistently outperforms word-based pretrained models in various cross-lingual transfer tasks. We also analyze the model and the key insight is that incorporating entity representations into the input allows us to extract more language-agnostic features. We also evaluate the model with a multilingual cloze prompt task with the mLAMA dataset. We show that entity-based prompt elicits correct factual knowledge more likely than using only word representations.

11:30-11:45 (The Liffey B)
Continual Sequence Generation with Adaptive Compositional Modules
Yanzhe Zhang, Xuezhi Wang and Diyi Yang
Continual learning is essential for real-world deployment when there is a need to quickly adapt the model to new tasks without forgetting knowledge of old tasks. Existing work on continual sequence generation either always reuses existing parameters to learn new tasks, which is vulnerable to catastrophic forgetting on dissimilar tasks, or blindly adds new parameters for every new task, which could prevent knowledge sharing between similar tasks. To get the best of both worlds, in this work, we propose continual sequence generation with adaptive compositional modules to adaptively add modules in transformer architectures and compose both old and new modules for new tasks. We also incorporate pseudo experience replay to facilitate knowledge transfer in those shared modules. Experiment results on various sequences

of generation tasks show that our framework can adaptively add modules or reuse modules based on task similarity, outperforming state-of-the-art baselines in terms of both performance and parameter efficiency. We make our code public at https://github.com/GT-SALT/Adaptive-Compositional-Modules.

11:45-12:00 (The Liffey B)
Coherence boosting: When your pretrained language model is not paying enough attention
Nikolay Malkin, Zhen Wang and Nebojsa Jojic
Long-range semantic coherence remains a challenge in automatic language generation and understanding. We demonstrate that large language models have insufficiently learned the effect of distant words on next-token prediction. We present coherence boosting, an inference procedure that increases a LM's focus on a long context. We show the benefits of coherence boosting with pretrained models by distributional analyses of generated ordinary text and dialog responses. It is also found that coherence boosting with state-of-the-art models for various zero-shot NLP tasks yields performance gains with no additional training.

12:00-12:15 (The Liffey B)
Compression of Generative Pre-trained Language Models via Quantization
Chaofan Tao, Lu Hou, Wei Zhang, Lifeng Shang, Xin Jiang, Qun Liu, Ping Luo and Ngai Wong
The increasing size of generative Pre-trained Language Models (PLMs) have greatly increased the demand for model compression. Despite various methods to compress BERT or its variants, there are few attempts to compress generative PLMs, and the underlying difficulty remains unclear. In this paper, we compress generative PLMs by quantization. We find that previous quantization methods fail on generative tasks due to the homogeneous word embeddings caused by reduced capacity and the varied distribution of weights. Correspondingly, we propose a token-level contrastive distillation to learn distinguishable word embeddings, and a module-wise dynamic scaling to make quantizers adaptive to different modules. Empirical results on various tasks show that our proposed method outperforms the state-of-the-art compression methods on generative PLMs by a clear margin. With comparable performance with the full-precision models, we achieve 14.4x and 13.4x compression rate on GPT-2 and BART, respectively.

12:15-12:25 (The Liffey B)
Unsupervised multiple-choice question generation for out-of-domain Q&A fine-tuning
Guillaume Le Berre, Christophe Cerisara, Philippe Langlais and Guy Lapalme
Pre-trained models have shown very good performances on a number of question answering benchmarks especially when fine-tuned on multiple question answering datasets at once. In this work, we propose an approach for generating a fine-tuning dataset thanks to a rule-based algorithm that generates questions and answers from unannotated sentences. We show that the state-of-the-art model UnifiedQA can greatly benefit from such a system on a multiple-choice benchmark about physics, biology and chemistry it has never been trained on. We further show that improved performances may be obtained by selecting the most challenging distractors (wrong answers), with a dedicated ranker based on a pretrained RoBERTa model.

Machine Translation and Multilinguality 1

11:00-12:30 (The Liffey A)

11:00-11:15 (The Liffey A)
From Simultaneous to Streaming Machine Translation by Leveraging Streaming History
Javier Iranzo Sanchez, Jorge Civera and Alfons Juan-Císcar
Simultaneous machine translation has recently gained traction thanks to significant quality improvements and the advent of streaming applications. Simultaneous translation systems need to find a trade-off between translation quality and response time, and with this purpose multiple latency measures have been proposed. However, latency evaluations for simultaneous translation are estimated at the sentence level, not taking into account the sequential nature of a streaming scenario. Indeed, these sentence-level latency measures are not well suited for continuous stream translation, resulting in figures that are not coherent with the simultaneous translation policy of the system being assessed. This work proposes a stream-level adaptation of the current latency measures based on a re-segmentation approach applied to the output translation, that is successfully evaluated on streaming conditions for a reference IWSLT task

11:15-11:30 (The Liffey A)
Multilingual Document-Level Translation Enables Zero-Shot Transfer From Sentences to Documents
Biao Zhang, Ankur Bapna, Melvin Johnson, Ali Dabirmoghaddam, Naveen Arivazhagan and Orhan Firat
Document-level neural machine translation (DocNMT) achieves coherent translations by incorporating cross-sentence context. However, for most language pairs there's a shortage of parallel documents, although parallel sentences are readily available. In this paper, we study whether and how contextual modeling in DocNMT is transferable via multilingual modeling. We focus on the scenario of zero-shot transfer from teacher languages with document level data to student languages with no documents but sentence level data, and for the first time treat document-level translation as a transfer learning problem. Using simple concatenation-based DocNMT, we explore the effect of 3 factors on the transfer: the number of teacher languages with document level data, the balance between document and sentence level data at training, and the data condition of parallel documents (genuine vs. back-translated). Our experiments on Europarl-7 and IWSLT-10 show the feasibility of multilingual transfer for DocNMT, particularly on document-specific metrics. We observe that more teacher languages and adequate data balance both contribute to better transfer quality. Surprisingly, the transfer is less sensitive to the data condition, where multilingual DocNMT delivers decent performance with either back-translated or genuine document pairs.

11:30-11:45 (The Liffey A)
Under the Morphosyntactic Lens: A Multifaceted Evaluation of Gender Bias in Speech Translation
Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri and Marco Turchi
Gender bias is largely recognized as a problematic phenomenon affecting language technologies, with recent studies underscoring that it might surface differently across languages. However, most of current evaluation practices adopt a word-level focus on a narrow set of occupational nouns under synthetic conditions. Such protocols overlook key features of grammatical gender languages, which are characterized by morphosyntactic chains of gender agreement, marked on a variety of lexical items and parts-of-speech (POS). To overcome this limitation, we

enrich the natural, gender-sensitive MuST-SHE corpus (Bentivogli et al., 2020) with two new linguistic annotation layers (POS and agreement chains), and explore to what extent different lexical categories and agreement phenomena are impacted by gender skews. Focusing on speech translation, we conduct a multifaceted evaluation on three language directions (English-French/Italian/Spanish), with models trained on varying amounts of data and different word segmentation techniques. By shedding light on model behaviours, gender bias, and its detection at several levels of granularity, our findings emphasize the value of dedicated analyses beyond aggregated overall results.

11:45-12:00 (The Liffey A)
[TACL] **ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models**
Noah Constant, Linting Xue, Aditya Barua, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts and Colin Raffel

12:00-12:15 (The Liffey A)
[TACL] **The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation**
Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán and Angela Fan

12:15-12:30 (The Liffey A)
Learning to Generalize to More: Continuous Semantic Augmentation for Neural Machine Translation
Xiangpeng Wei, Heng Yu, Yue Hu, Rongxiang Weng, Weihua Luo and Rong Jin
The principal task in supervised neural machine translation (NMT) is to learn to generate target sentences conditioned on the source inputs from a set of parallel sentence pairs, and thus produce a model capable of generalizing to unseen instances. However, it is commonly observed that the generalization performance of the model is highly influenced by the amount of parallel data used in training. Although data augmentation is widely used to enrich the training data, conventional methods with discrete manipulations fail to generate diverse and faithful training samples. In this paper, we present a novel data augmentation paradigm termed Continuous Semantic Augmentation (CsaNMT), which augments each training instance with an adjacency semantic region that could cover adequate variants of literal expression under the same meaning. We conduct extensive experiments on both rich-resource and low-resource settings involving various language pairs, including WMT14 English→{German,French}, NIST Chinese→English and multiple low-resource IWSLT translation tasks. The provided empirical evidences show that CsaNMT sets a new level of performance among existing augmentation techniques, improving on the state-of-the-art by a large margin. The core codes are contained in Appendix E.

Resources and Evaluation 1

11:00-12:30 (Liffey Hall 2)

11:00-11:15 (Liffey Hall 2)
Quantified Reproducibility Assessment of NLP Results
Anya Belz, Maja Popovic and Simon Mille
This paper describes and tests a method for carrying out quantified reproducibility assessment (QRA) that is based on concepts and definitions from metrology. QRA produces a single score estimating the degree of reproducibility of a given system and evaluation measure, on the basis of the scores from, and differences between, different reproductions. We test QRA on 18 different system and evaluation measure combinations (involving diverse NLP tasks and types of evaluation), for each of which we have the original results and one to seven reproduction results. The proposed QRA method produces degree-of-reproducibility scores that are comparable across multiple reproductions not only of the same, but also of different, original studies. We find that the proposed method facilitates insights into causes of variation between reproductions, and as a result, allows conclusions to be drawn about what aspects of system and/or evaluation design need to be changed in order to improve reproducibility.

11:15-11:30 (Liffey Hall 2)
BenchIE: A Framework for Multi-Faceted Fact-Based Open Information Extraction Evaluation
Kiril Gashteovski, Mingying Yu, Bhushan Kotnis, Carolin Lawrence, Mathias Niepert and Goran Glavaš
Intrinsic evaluations of OIE systems are carried out either manually—with human evaluators judging the correctness of extractions—or automatically, on standardized benchmarks. The latter, while much more cost-effective, is less reliable, primarily because of the incompleteness of the existing OIE benchmarks: the ground truth extractions do not include all acceptable variants of the same fact, leading to unreliable assessment of the models' performance. Moreover, the existing OIE benchmarks are available for English only. In this work, we introduce BenchIE: a benchmark and evaluation framework for comprehensive evaluation of OIE systems for English, Chinese, and German. In contrast to existing OIE benchmarks, BenchIE is fact-based, i.e., it takes into account informational equivalence of extractions: our gold standard consists of *fact synsets*, clusters in which we exhaustively list all acceptable surface forms of the same fact. Moreover, having in mind common downstream applications for OIE, we make BenchIE multi-faceted; i.e., we create benchmark variants that focus on different facets of OIE evaluation, e.g., compactness or minimality of extractions. We benchmark several state-of-the-art OIE systems using BenchIE and demonstrate that these systems are significantly less effective than indicated by existing OIE benchmarks. We make BenchIE (data and evaluation code) publicly available.

11:30-11:45 (Liffey Hall 2)
RoMe: A Robust Metric for Evaluating Natural Language Generation
Md Rashad Al Hasan Rony, Liubov Kovriguina, Debanjan Chaudhuri, Ricardo Usbeck and Jens Lehmann
Evaluating Natural Language Generation (NLG) systems is a challenging task. Firstly, the metric should ensure that the generated hypothesis reflects the reference's semantics. Secondly, it should consider the grammatical quality of the generated sentence. Thirdly, it should be robust enough to handle various surface forms of the generated sentence. Thus, an effective evaluation metric has to be multifaceted. In this paper, we propose an automatic evaluation metric incorporating several core aspects of natural language understanding (language competence, syntactic and semantic variation). Our proposed metric, RoMe, is trained on language features such as semantic similarity combined with tree edit distance and grammatical acceptability, using a self-supervised neural network to assess the overall quality of the generated sentence. Moreover, we perform an extensive robustness analysis of the state-of-the-art methods and RoMe. Empirical results suggest that RoMe

has a stronger correlation to human judgment over state-of-the-art metrics in evaluating system-generated sentences across several NLG tasks.

11:45-12:00 (Liffey Hall 2)
Active Evaluation: Efficient NLG Evaluation with Few Pairwise Comparisons
Akash Kumar Mohankumar and Mitesh M Khapra
Recent studies have shown the advantages of evaluating NLG systems using pairwise comparisons as opposed to direct assessment. Given k systems, a naive approach for identifying the top-ranked system would be to uniformly obtain pairwise comparisons from all $\binom{k}{2}$ pairs of systems. However, this can be very expensive as the number of human annotations required would grow quadratically with k. In this work, we introduce Active Evaluation, a framework to efficiently identify the top-ranked system by actively choosing system pairs for comparison using dueling bandit algorithms. We perform extensive experiments with 13 dueling bandits algorithms on 13 NLG evaluation datasets spanning 5 tasks and show that the number of human annotations can be reduced by 80

12:00-12:15 (Liffey Hall 2)
Human Evaluation and Correlation with Automatic Metrics in Consultation Note Generation
Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz and Aleksandar Savkov
In recent years, machine learning models have rapidly become better at generating clinical consultation notes; yet, there is little work on how to properly evaluate the generated consultation notes to understand the impact they may have on both the clinician using them and the patient's clinical safety. To address this we present an extensive human evaluation study of consultation notes where 5 clinicians (i) listen to 57 mock consultations, (ii) write their own notes, (iii) post-edit a number of automatically generated notes, and (iv) extract all the errors, both quantitative and qualitative. We then carry out a correlation study with 18 automatic quality metrics and the human judgements. We find that a simple, character-based Levenshtein distance metric performs on par if not better than common model-based metrics like BertScore. All our findings and annotations are open-sourced.

12:15-12:25 (Liffey Hall 2)
PriMock57: A Dataset Of Primary Care Mock Consultations
Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac and Aleksandar Savkov
Recent advances in Automatic Speech Recognition (ASR) have made it possible to reliably produce automatic transcripts of clinician-patient conversations. However, access to clinical datasets is heavily restricted due to patient privacy, thus slowing down normal research practices. We detail the development of a public access, high quality dataset comprising of 57 mocked primary care consultations, including audio recordings, their manual utterance-level transcriptions, and the associated consultation notes. Our work illustrates how the dataset can be used as a benchmark for conversational medical ASR as well as consultation note generation from transcripts.

Information Extraction 1

11:00-12:25 (Wicklow Hall 2b)

11:00-11:15 (Wicklow Hall 2b)
A Meta-framework for Spatiotemporal Quantity Extraction from Text
Qiang Ning, Ben Zhou, Hao Wu, Haoruo Peng, Chuchu Fan and Matt Gardner
News events are often associated with quantities (e.g., the number of COVID-19 patients or the number of arrests in a protest), and it is often important to extract their type, time, and location from unstructured text in order to analyze these quantity events. This paper thus formulates the NLP problem of spatiotemporal quantity extraction, and proposes the first meta-framework for solving it. This meta-framework contains a formalism that decomposes the problem into several information extraction tasks, a shareable crowdsourcing pipeline, and transformer-based baseline models. We demonstrate the meta-framework in three domains—the COVID-19 pandemic, Black Lives Matter protests, and 2020 California wildfires—to show that the formalism is general and extensible, the crowdsourcing pipeline facilitates fast and high-quality data annotation, and the baseline system can handle spatiotemporal quantity extraction well enough to be practically useful. We release all resources for future research on this topic at https://github.com/steqe.

11:15-11:30 (Wicklow Hall 2b)
[TACL] **VILA: Improving Structured Content Extraction from Scientific PDFs Using Visual Layout Groups**
Zejiang Shen, Kyle Lo, Lucy Wang, Bailey Kuehl, Daniel Weld and Doug Downey

11:30-11:45 (Wicklow Hall 2b)
[TACL] **Predicting Document Coverage for Relation Extraction**
Sneha Singhania, Simon Razniewski and Gerhard Weikum

11:45-12:00 (Wicklow Hall 2b)
Continual Few-shot Relation Learning via Embedding Space Regularization and Data Augmentation
Chengwei Qin and Shafiq Joty
Existing continual relation learning (CRL) methods rely on plenty of labeled training data for learning a new task, which can be hard to acquire in real scenario as getting large and representative labeled data is often expensive and time-consuming. It is therefore necessary for the model to learn novel relational patterns with very few labeled data while avoiding catastrophic forgetting of previous task knowledge. In this paper, we formulate this challenging yet practical problem as continual few-shot relation learning (CFRL). Based on the finding that learning for new emerging few-shot tasks often results in feature distributions that are incompatible with previous tasks' learned distributions, we propose a novel method based on embedding space regularization and data augmentation. Our method generalizes to new few-shot tasks and avoids catastrophic forgetting of previous tasks by enforcing extra constraints on the relational embeddings and by adding extra relevant data in a self-supervised manner. With extensive experiments we demonstrate that our method can significantly outperform previous state-of-the-art methods in CFRL task settings.

12:00-12:15 (Wicklow Hall 2b)
Distantly Supervised Named Entity Recognition via Confidence-Based Multi-Class Positive and Unlabeled Learning
Kang Zhou, Yuepei Li and Qi Li
In this paper, we study the named entity recognition (NER) problem under distant supervision. Due to the incompleteness of the external dictionaries and/or knowledge bases, such distantly annotated training data usually suffer from a high false negative rate. To this end, we formulate the Distantly Supervised NER (DS-NER) problem via Multi-class Positive and Unlabeled (MPU) learning and propose a theoretically and practically novel CONFidence-based MPU (Conf-MPU) approach. To handle the incomplete annotations, Conf-MPU consists of two steps. First, a confidence score is estimated for each token of being an entity token. Then, the proposed Conf-MPU risk estimation is applied to train a multi-class classifier for the NER task. Thorough experiments on two benchmark datasets labeled by various external knowledge demonstrate the superiority of the proposed Conf-MPU over existing DS-NER methods. Our code is available at Github.

12:15-12:25 (Wicklow Hall 2b)
Simple and Effective Knowledge-Driven Query Expansion for QA-Based Product Attribute Extraction
Keiji Shinzato, Naoki Yoshinaga, Yandi Xia and Wei-Te Chen
A key challenge in attribute value extraction (AVE) from e-commerce sites is how to handle a large number of attributes for diverse products. Although this challenge is partially addressed by a question answering (QA) approach which finds a value in product data for a given query (attribute), it does not work effectively for rare and ambiguous queries. We thus propose simple knowledge-driven query expansion based on possible answers (values) of a query (attribute) for QA-based AVE. We retrieve values of a query (attribute) from the training data to expand the query. We train a model with two tricks, knowledge dropout and knowledge token mixing, which mimic the imperfection of the value knowledge in testing. Experimental results on our cleaned version of AliExpress dataset show that our method improves the performance of AVE (+6.08 macro F1), especially for rare and ambiguous attributes (+7.82 and +6.86 macro F1, respectively).

Language Grounding, Speech and Multimodality 1

11:00-12:30 (Liffey Hall 1)

11:00-11:15 (Liffey Hall 1)
Analyzing Generalization of Vision and Language Navigation to Unseen Outdoor Areas
Raphael Schumann and Stefan Riezler
Vision and language navigation (VLN) is a challenging visually-grounded language understanding task. Given a natural language navigation instruction, a visual agent interacts with a graph-based environment equipped with panorama images and tries to follow the described route. Most prior work has been conducted in indoor scenarios where best results were obtained for navigation on routes that are similar to the training routes, with sharp drops in performance when testing on unseen environments. We focus on VLN in outdoor scenarios and find that in contrast to indoor VLN, most of the gain in outdoor VLN on unseen data is due to features like junction type embedding or heading delta that are specific to the respective environment graph, while image information plays a very minor role in generalizing VLN to unseen outdoor areas. These findings show a bias to specifics of graph representations of urban environments, demanding that VLN tasks grow in scale and diversity of geographical environments.

11:15-11:30 (Liffey Hall 1)
Language-Agnostic Meta-Learning for Low-Resource Text-to-Speech with Articulatory Features
Florian Lux and Thang Vu
While neural text-to-speech systems perform remarkably well in high-resource scenarios, they cannot be applied to the majority of the over 6,000 spoken languages in the world due to a lack of appropriate training data. In this work, we use embeddings derived from articulatory vectors rather than embeddings derived from phoneme identities to learn phoneme representations that hold across languages. In conjunction with language agnostic meta learning, this enables us to fine-tune a high-quality text-to-speech model on just 30 minutes of data in a previously unseen language spoken by a previously unseen speaker.

11:30-11:45 (Liffey Hall 1)
Cross-Modal Discrete Representation Learning
Alexander H. Liu, SouYoung Jin, Cheng-I Lai, Andrew Rouditchenko, Aude Oliva and James R. Glass
In contrast to recent advances focusing on high-level representation learning across modalities, in this work we present a self-supervised learning framework that is able to learn a representation that captures finer levels of granularity across different modalities such as concepts or events represented by visual objects or spoken words. Our framework relies on a discretized embedding space created via vector quantization that is shared across different modalities. Beyond the shared embedding space, we propose a Cross-Modal Code Matching objective that forces the representations from different views (modalities) to have a similar distribution over the discrete embedding space such that cross-modal objects/actions localization can be performed without direct supervision. We show that the proposed discretized multi-modal fine-grained representation (e.g., pixel/word/frame) can complement high-level summary representations (e.g., video/sentence/waveform) for improved performance on cross-modal retrieval tasks. We also observe that the discretized representation uses individual clusters to represent the same semantic concept across modalities.

11:45-12:00 (Liffey Hall 1)
Leveraging Visual Knowledge in Language Tasks: An Empirical Study on Intermediate Pre-training for Cross-Modal Knowledge Transfer
Woojeong Jin, Dong-Ho Lee, Chenguang Zhu, Jay Pujara and Xiang Ren
Pre-trained language models are still far from human performance in tasks that need understanding of properties (e.g. appearance, measurable quantity) and affordances of everyday objects in the real world since the text lacks such information due to reporting bias. In this work, we study whether integrating visual knowledge into a language model can fill the gap. We investigate two types of knowledge transfer: (1) *text knowledge transfer using image captions that may contain enriched visual knowledge and (2) cross-modal knowledge transfer using both images and captions with vision-language training objectives. On 5 downstream tasks that may need visual knowledge to solve the problem, we perform extensive empirical comparisons over the presented objectives. Our experiments show that visual knowledge transfer can improve performance in both low-resource and fully supervised settings.*

12:00-12:15 (Liffey Hall 1)
Image Retrieval from Contextual Descriptions
Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti and Siva Reddy
The ability to integrate context, including perceptual and temporal cues, plays a pivotal role in grounding the meaning of a linguistic utterance. In order to measure to what extent current vision-and-language models master this ability, we devise a new multimodal challenge, Image Retrieval from Contextual Descriptions (ImageCoDe). In particular, models are tasked with retrieving the correct image from a set of 10 minimally contrastive candidates based on a contextual description. As such, each description contains only the details that help distinguish between images. Because of this, descriptions tend to be complex in terms of syntax and discourse and require drawing pragmatic inferences. Images are sourced from both static pictures and video frames. We benchmark several state-of-the-art models, including both cross-encoders such as ViLBERT and bi-encoders such as CLIP, on ImageCoDe. Our results reveal that these models dramatically lag behind human performance: the best variant achieves an accuracy of 20.9 on video frames and 59.4 on static pictures, compared with 90.8 in humans. Furthermore, we experiment with new model variants that are better equipped to incorporate visual and temporal context into their representations, which achieve modest gains. Our hope is that ImageCoDE will foster progress in grounded language understanding by encouraging models to focus on fine-grained visual differences.

12:15-12:25 (Liffey Hall 1)
Sample, Translate, Recombine: Leveraging Audio Alignments for Data Augmentation in End-to-end Speech Translation
Tsz Kin Lam, Shigehiko Schamoni and Stefan Riezler
End-to-end speech translation relies on data that pair source-language speech inputs with corresponding translations into a target language. Such data are notoriously scarce, making synthetic data augmentation by back-translation or knowledge distillation a necessary ingredient of end-to-end training. In this paper, we present a novel approach to data augmentation that leverages audio alignments, linguistic properties, and translation. First, we augment a transcription by sampling from a suffix memory that stores text and audio data. Second, we translate the augmented transcript. Finally, we recombine concatenated audio segments and the generated translation. Our method delivers consistent improvements of up to 0.9 and 1.1 BLEU points on top of augmentation with knowledge distillation on five language pairs on CoVoST 2 and on two language pairs on Europarl-ST, respectively.

Poster Session 1: Interpretability and Analysis of Models for NLP

11:00-12:30 (Forum)

11:00-12:30 (Forum)
#1 Rewire-then-Probe: A Contrastive Recipe for Probing Biomedical Knowledge of Pre-trained Language Models
Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Anne Collins and Nigel Collier
Knowledge probing is crucial for understanding the knowledge transfer mechanism behind the pre-trained language models (PLMs). Despite the growing progress of probing knowledge for PLMs in the general domain, specialised areas such as the biomedical domain are vastly under-explored. To facilitate this, we release a well-curated biomedical knowledge probing benchmark, MedLAMA, constructed based on the Unified Medical Language System (UMLS) Metathesaurus. We test a wide spectrum of state-of-the-art PLMs and probing approaches on our benchmark, reaching at most 3% of acc@10. While highlighting various sources of domain-specific challenges that amount to this underwhelming performance, we illustrate that the underlying PLMs have a higher potential for probing tasks. To achieve this, we propose Contrastive-Probe, a novel self-supervised contrastive probing approach, that adjusts the underlying PLMs without using any probing data. While Contrastive-Probe pushes the acc@10 to 28%, the performance gap still remains notable. Our human expert evaluation suggests that the probing performance of our Contrastive-Probe is still under-estimated as UMLS still does not include the full spectrum of factual knowledge. We hope MedLAMA and Contrastive-Probe facilitate further developments of more suited probing techniques for this domain. Our code and dataset are publicly available at https://github.com/cambridgeltl/medlama.

11:00-12:30 (Forum)
#2 When classifying grammatical role, BERT doesn't care about word order... except when it matters
Isabel Papadimitriou, Richard Futrell and Kyle Mahowald
Because meaning can often be inferred from lexical semantics alone, word order is often a redundant cue in natural language. For example, the words chopped, chef, and onion are more likely used to convey "The chef chopped the onion," not "The onion chopped the chef." Recent work has shown large language models to be surprisingly word order invariant, but crucially has largely considered natural prototypical inputs, where compositional meaning mostly matches lexical expectations. To overcome this confound, we probe grammatical role representation in English BERT and GPT-2, on instances where lexical expectations are not sufficient, and word order knowledge is necessary for correct classification. Such non-prototypical instances are naturally occurring English sentences with inanimate subjects or animate objects, or sentences where we systematically swap the arguments to make sentences like "The onion chopped the chef". We find that, while early layer embeddings are largely lexical, word order is in fact crucial in defining the later-layer representations of words in semantically non-prototypical positions. Our experiments isolate the effect of word order on the contextualization process, and highlight how models use context in the uncommon, but critical, instances where it matters.

11:00-12:30 (Forum)
#3 Low-Rank Softmax Can Have Unargmaxable Classes in Theory but Rarely in Practice
Andreas Grivas, Nikolay Bogoychev and Adam Lopez
Classifiers in natural language processing (NLP) often have a large number of output classes. For example, neural language models (LMs) and machine translation (MT) models both predict tokens from a vocabulary of thousands. The Softmax output layer of these models typically receives as input a dense feature representation, which has much lower dimensionality than the output. In theory, the result is some words may be impossible to be predicted via argmax, irrespective of input features, and empirically, there is evidence this happens in small language models (Demeter et al., 2020). In this paper we ask whether it can happen in practical large language models and translation models. To do so, we develop algorithms to detect such unargmaxable tokens in public models. We find that 13 out of 150 models do indeed have such tokens; however, they are very infrequent and unlikely to impact model quality. We release our algorithms and code to the public.

11:00-12:30 (Forum)
#4 Probing for the Usage of Grammatical Number
Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau and Ryan D Cotterell
A central quest of probing is to uncover how pre-trained models encode a linguistic property within their representations. An encoding, however, might be spurious-i.e., the model might not rely on it when making predictions. In this paper, we try to find an encoding that the model actually uses, introducing a usage-based probing setup. We first choose a behavioral task which cannot be solved without using the linguistic property. Then, we attempt to remove the property by intervening on the model's representations. We contend that, if an encoding is used by the model, its removal should harm the performance on the chosen behavioral task. As a case study, we focus on how BERT encodes grammatical number, and on how it uses this encoding to solve the number agreement task. Experimentally, we find that BERT relies on a linear encoding of grammatical number to produce the correct behavioral output. We also find that BERT uses a separate encoding of grammatical number for nouns and verbs. Finally, we identify in which layers information about grammatical number is transferred from a noun to its head verb.

11:00-12:30 (Forum)
#5 How Distributed are Distributed Representations? An Observation on the Locality of Syntactic Information in Verb Agreement Tasks
Bingzhi Li, Guillaume Wisniewski and Benoit Crabbé
This work addresses the question of the localization of syntactic information encoded in the transformers representations. We tackle this question from two perspectives, considering the object-past participle agreement in French, by identifying, first, in which part of the sentence and, second, in which part of the representation the syntactic information is encoded. The results of our experiments, using probing, causal analysis and feature selection method, show that syntactic information is encoded locally in a way consistent with the French grammar.

11:00-12:30 (Forum)
#6 Counterfactual Explanations for Natural Language Interfaces
George Tolkachev, Stephen Mell, Stephan Zdancewic and Osbert Bastani
A key challenge facing natural language interfaces is enabling users to understand the capabilities of the underlying system. We propose a novel approach for generating explanations of a natural language interface based on semantic parsing. We focus on counterfactual explanations, which are post-hoc explanations that describe to the user how they could have minimally modified their utterance to achieve their desired goal. In particular, the user provides an utterance along with a demonstration of their desired goal; then, our algorithm synthesizes a paraphrase of their utterance that is guaranteed to achieve their goal. In two user studies, we demonstrate that our approach substantially improves user performance, and that it generates explanations that more closely match the user's intent compared to two ablations.

11:00-12:30 (Forum)
#7 How does the pre-training objective affect what large language models learn about linguistic properties?
Ahmed Alajrami and Nikolaos Aletras
Several pre-training objectives, such as masked language modeling (MLM), have been proposed to pre-train language models (e.g. BERT) with the aim of learning better language representations. However, to the best of our knowledge, no previous work so far has investigated how different pre-training objectives affect what BERT learns about linguistics properties. We hypothesize that linguistically motivated objectives such as MLM should help BERT to acquire better linguistic knowledge compared to other non-linguistically motivated objectives that are not intuitive or hard for humans to guess the association between the input and the label to be predicted. To this end, we pre-train BERT with two linguistically motivated objectives and three non-linguistically motivated ones. We then probe for linguistic characteristics encoded in the representation of the resulting models. We find strong evidence that there are only small differences in probing performance between the representations learned by the two different types of objectives. These surprising results question the dominant narrative of linguistically informed pre-training.

11:00-12:30 (Forum)
#8 What to Learn, and How: Toward Effective Learning from Rationales
Samuel Carton, Surya Kanoria and Chenhao Tan
Learning from rationales seeks to augment model prediction accuracy using human-annotated rationales (i.e. subsets of input tokens) that justify their chosen labels, often in the form of intermediate or multitask supervision. While intuitive, this idea has proven elusive in practice. We make two observations about human rationales via empirical analyses: 1) maximizing rationale supervision accuracy is not necessarily the optimal objective for improving model accuracy; 2) human rationales vary in whether they provide sufficient information for the model to exploit for prediction. Building on these insights, we propose several novel loss functions and learning strategies, and evaluate their effectiveness on three datasets with human rationales. Our results demonstrate consistent improvements over baselines in both label and rationale accuracy, including a 3

11:00-12:30 (Forum)
#9 Probing Multilingual Cognate Prediction Models
Clémentine Fourrier and Benoît Sagot
Character-based neural machine translation models have become the reference models for cognate prediction, a historical linguistics task. So far, all linguistic interpretations about latent information captured by such models have been based on external analysis (accuracy, raw results, errors). In this paper, we investigate what probing can tell us about both models and previous interpretations, and learn that though our models store linguistic and diachronic information, they do not achieve it in previously assumed ways.

11:00-12:30 (Forum)
#10 Exploring the Capacity of a Large-scale Masked Language Model to Recognize Grammatical Errors
Ryo Nagata, Manabu Kimura and Kazuaki Hanawa
In this paper, we explore the capacity of a language model-based method for grammatical error detection in detail. We first show that 5 to 10% of training data are enough for a BERT-based error detection method to achieve performance equivalent to what a non-language model-based method can achieve with the full training data; recall improves much faster with respect to training data size in the BERT-based method than in the non-language model method. This suggests that (i) the BERT-based method should have a good knowledge of the grammar required to recognize certain types of error and that (ii) it can transform the knowledge into error detection rules by fine-tuning with few training samples, which explains its high generalization ability in grammatical error detection. We further show with pseudo error data that it actually exhibits

such nice properties in learning rules for recognizing various types of error. Finally, based on these findings, we discuss a cost-effective method for detecting grammatical errors with feedback comments explaining relevant grammatical rules to learners.

11:00-12:30 (Forum)
[TACL] **#11 Evaluating Explanations: How Much do Explanations from the Teacher aid Students?**
Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Soares, Michael Collins, Zachary Lipton, Graham Neubig and William Cohen

11:00-12:30 (Forum)
#12 Probing as Quantifying Inductive Bias
Alexander Immer, Lucas Torroba Hennigen, Vincent Fortuin and Ryan D Cotterell
Pre-trained contextual representations have led to dramatic performance improvements on a range of downstream tasks. Such performance improvements have motivated researchers to quantify and understand the linguistic information encoded in these representations. In general, researchers quantify the amount of linguistic information through probing, an endeavor which consists of training a supervised model to predict a linguistic property directly from the contextual representations. Unfortunately, this definition of probing has been subject to extensive criticism in the literature, and has been observed to lead to paradoxical and counter-intuitive results. In the theoretical portion of this paper, we take the position that the goal of probing ought to be measuring the amount of inductive bias that the representations encode on a specific task. We further describe a Bayesian framework that operationalizes this goal and allows us to quantify the representations' inductive bias. In the empirical portion of the paper, we apply our framework to a variety of NLP tasks. Our results suggest that our proposed framework alleviates many previous problems found in probing. Moreover, we are able to offer concrete evidence that—for some tasks—fastText can offer a better inductive bias than BERT.

11:00-12:30 (Forum)
#13 Data Contamination: From Memorization to Exploitation
Inbal Magar and Roy Schwartz
Pretrained language models are typically trained on massive web-based datasets, which are often "contaminated" with downstream test sets. It is not clear to what extent models exploit the contaminated data for downstream tasks. We present a principled method to study this question. We pretrain BERT models on joint corpora of Wikipedia and labeled downstream datasets, and fine-tune them on the relevant task. Comparing performance between samples seen and unseen during pretraining enables us to define and quantify levels of memorization and exploitation. Experiments with two models and three downstream tasks show that exploitation exists in some cases, but in others the models memorize the contaminated data, but do not exploit it. We show that these two measures are affected by different factors such as the number of duplications of the contaminated data and the model size. Our results highlight the importance of analyzing massive web-scale datasets to verify that progress in NLP is obtained by better language understanding and not better data exploitation.

11:00-12:30 (Forum)
#14 Is Attention Explanation? An Introduction to the Debate
Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François and Patrick Watrin
The performance of deep learning models in NLP and other fields of machine learning has led to a rise in their popularity, and so the need for explanations of these models becomes paramount. Attention has been seen as a solution to increase performance, while providing some explanations. However, a debate has started to cast doubt on the explanatory power of attention in neural networks. Although the debate has created a vast literature thanks to contributions from various areas, the lack of communication is becoming more and more tangible. In this paper, we provide a clear overview of the insights on the debate by critically confronting works from these different areas. This holistic vision can be of great interest for future works in all the communities concerned by this debate. We sum up the main challenges spotted in these areas, and we conclude by discussing the most promising future avenues on attention as an explanation.

11:00-12:30 (Forum)
#15 The Paradox of the Compositionality of Natural Language: A Neural Machine Translation Case Study
Verna Dankers, Elia Bruni and Dieuwke Hupkes
Obtaining human-like performance in NLP is often argued to require compositional generalisation. Whether neural networks exhibit this ability is usually studied by training models on highly compositional synthetic data. However, compositionality in natural language is much more complex than the rigid, arithmetic-like version such data adheres to, and artificial compositionality tests thus do not allow us to determine how neural models deal with more realistic forms of compositionality. In this work, we re-instantiate three compositionality tests from the literature and reformulate them for neural machine translation (NMT). Our results highlight that: i) unfavourably, models trained on more data are more compositional; ii) models are sometimes less compositional than expected, but sometimes more, exemplifying that different levels of compositionality are required, and models are not always able to modulate between them correctly; iii) some of the non-compositional behaviours are mistakes, whereas others reflect the natural variation in data. Apart from an empirical study, our work is a call to action: we should rethink the evaluation of compositionality in neural networks and develop benchmarks using real data to evaluate compositionality on natural language, where composing meaning is not as straightforward as doing the math.

11:00-12:30 (Forum)
#16 Finding Structural Knowledge in Multimodal-BERT
Victor Siemen Janusz Milewski, Miryam de Lhoneux and Marie-Francine Moens
In this work, we investigate the knowledge learned in the embeddings of multimodal-BERT models. More specifically, we probe their capabilities of storing the grammatical structure of linguistic data and the structure learned over objects in visual data. To reach that goal, we first make the inherent structure of language and visuals explicit by a dependency parse of the sentences that describe the image and by the dependencies between the object regions in the image, respectively. We call this explicit visual structure the scene tree, that is based on the dependency tree of the language description. Extensive probing experiments show that the multimodal-BERT models do not encode these scene trees.

11:00-12:30 (Forum)
#17 Toward Interpretable Semantic Textual Similarity via Optimal Transport-based Contrastive Sentence Learning
Seonghyeon Lee, Dongha Lee, Seongbo Jang and Hwanjo Yu

Recently, finetuning a pretrained language model to capture the similarity between sentence embeddings has shown the state-of-the-art performance on the semantic textual similarity (STS) task. However, the absence of an interpretation method for the sentence similarity makes it difficult to explain the model output. In this work, we explicitly describe the sentence distance as the weighted sum of contextualized token distances on the basis of a transportation problem, and then present the optimal transport-based distance measure, named RCMD; it identifies and leverages semantically-aligned token pairs. In the end, we propose CLRCMD, a contrastive learning framework that optimizes RCMD of sentence pairs, which enhances the quality of sentence similarity and their interpretation. Extensive experiments demonstrate that our learning framework outperforms other baselines on both STS and interpretable-STS benchmarks, indicating that it computes effective sentence similarity and also provides interpretation consistent with human judgement.

11:00-12:30 (Forum)
#18 Can Explanations Be Useful for Calibrating Black Box Models?
Xi Ye and Greg Durrett
NLP practitioners often want to take existing trained models and apply them to data from new domains. While fine-tuning or few-shot learning can be used to adapt a base model, there is no single recipe for making these techniques work; moreover, one may not have access to the original model weights if it is deployed as a black box. We study how to improve a black box model's performance on a new domain by leveraging explanations of the model's behavior. Our approach first extracts a set of features combining human intuition about the task with model attributions generated by black box interpretation techniques, then uses a simple calibrator, in the form of a classifier, to predict whether the base model was correct or not. We experiment with our method on two tasks, extractive question answering and natural language inference, covering adaptation from several pairs of domains with limited target-domain data. The experimental results across all the domain pairs show that explanations are useful for calibrating these models, boosting accuracy when predictions do not have to be returned on every example. We further show that the calibration model transfers to some extent between tasks.

11:00-12:30 (Forum)
#19 An Empirical Study on Explanations in Out-of-Domain Settings
George Chrysostomou and Nikolaos Aletras
Recent work in Natural Language Processing has focused on developing approaches that extract faithful explanations, either via identifying the most important tokens in the input (i.e. post-hoc explanations) or by designing inherently faithful models that first select the most important tokens and then use them to predict the correct label (i.e. select-then-predict models). Currently, these approaches are largely evaluated on in-domain settings. Yet, little is known about how post-hoc explanations and inherently faithful models perform in out-of-domain settings. In this paper, we conduct an extensive empirical study that examines: (1) the out-of-domain faithfulness of post-hoc explanations, generated by five feature attribution methods; and (2) the out-of-domain performance of two inherently faithful models over six datasets. Contrary to our expectations, results show that in many cases out-of-domain post-hoc explanation faithfulness measured by sufficiency and comprehensiveness is higher compared to in-domain. We find this misleading and suggest using a random baseline as a yardstick for evaluating post-hoc explanation faithfulness. Our findings also show that select-then predict models demonstrate comparable predictive performance in out-of-domain settings to full-text trained models.

11:00-12:30 (Forum)
#20 An Investigation of the (In)effectiveness of Counterfactually Augmented Data
Nitish Joshi and He He
While pretrained language models achieve excellent performance on natural language understanding benchmarks, they tend to rely on spurious correlations and generalize poorly to out-of-distribution (OOD) data. Recent work has explored using counterfactually-augmented data (CAD)—data generated by minimally perturbing examples to flip the ground-truth label—to identify robust features that are invariant under distribution shift. However, empirical results using CAD during training for OOD generalization have been mixed. To explain this discrepancy, through a toy theoretical example and empirical analysis on two crowdsourced CAD datasets, we show that: (a) while features perturbed in CAD are indeed robust features, it may prevent the model from learning unperturbed robust features; and (b) CAD may exacerbate existing spurious correlations in the data. Our results thus show that the lack of perturbation diversity limits CAD's effectiveness on OOD generalization, calling for innovative crowdsourcing procedures to elicit diverse perturbation of examples.

11:00-12:30 (Forum)
#21 ProtoTEx: Explaining Model Decisions with Prototype Tensors
Anubrata Das, Chitrank Gupta, Venelin Kovatchev, Matthew Lease and Junyi Jessy Li
We present ProtoTEx, a novel white-box NLP classification architecture based on prototype networks (Li et al., 2018). ProtoTEx faithfully explains model decisions based on prototype tensors that encode latent clusters of training examples. At inference time, classification decisions are based on the distances between the input text and the prototype tensors, explained via the training examples most similar to the most influential prototypes. We also describe a novel interleaved training algorithm that effectively handles classes characterized by ProtoTEx indicative features. On a propaganda detection task, ProtoTEx accuracy matches BART-large and exceeds BERTlarge with the added benefit of providing faithful explanations. A user study also shows that prototype-based explanations help non-experts to better recognize propaganda in online news.

11:00-12:30 (Forum)
#22 How reparametrization trick broke differentially-private text representation learning
Ivan Habernal
As privacy gains traction in the NLP community, researchers have started adopting various approaches to privacy-preserving methods. One of the favorite privacy frameworks, differential privacy (DP), is perhaps the most compelling thanks to its fundamental theoretical guarantees. Despite the apparent simplicity of the general concept of differential privacy, it seems non-trivial to get it right when applying it to NLP. In this short paper, we formally analyze several recent NLP papers proposing text representation learning using DPText (Beigi et al., 2019a,b; Alnasser et al., 2021; Beigi et al., 2021) and reveal their false claims of being differentially private. Furthermore, we also show a simple yet general empirical sanity check to determine whether a given implementation of a DP mechanism almost certainly violates the privacy loss guarantees. Our main goal is to raise awareness and help the community understand potential pitfalls of applying differential privacy to text representation learning.

11:00-12:30 (Forum)
#23 The Grammar-Learning Trajectories of Neural Language Models

Leshem Choshen, Guy Hacohen, Daphna Weinshall and Omri Abend
The learning trajectories of linguistic phenomena in humans provide insight into linguistic representation, beyond what can be gleaned from inspecting the behavior of an adult speaker. To apply a similar approach to analyze neural language models (NLM), it is first necessary to establish that different models are similar enough in the generalizations they make. In this paper, we show that NLMs with different initialization, architecture, and training data acquire linguistic phenomena in a similar order, despite their different end performance. These findings suggest that there is some mutual inductive bias that underlies these models' learning of linguistic phenomena. Taking inspiration from psycholinguistics, we argue that studying this inductive bias is an opportunity to study the linguistic representation implicit in NLMs. Leveraging these findings, we compare the relative performance on different phenomena at varying learning stages with simpler reference models. Results suggest that NLMs exhibit consistent "developmental" stages. Moreover, we find the learning trajectory to be approximately one-dimensional: given an NLM with a certain overall performance, it is possible to predict what linguistic generalizations it has already acquired. Initial analysis of these stages presents phenomena clusters (notably morphological ones), whose performance progresses in unison, suggesting a potential link between the generalizations behind them.

11:00-12:30 (Forum)
#24 Memorisation versus Generalisation in Pre-trained Language Models
Michael Tänzer, Sebastian Ruder and Marek Rei
State-of-the-art pre-trained language models have been shown to memorise facts and perform well with limited amounts of training data. To gain a better understanding of how these models learn, we study their generalisation and memorisation capabilities in noisy and low-resource scenarios. We find that the training of these models is almost unaffected by label noise and that it is possible to reach near-optimal results even on extremely noisy datasets. However, our experiments also show that they mainly learn from high-frequency patterns and largely fail when tested on low-resource tasks such as few-shot learning and rare entity recognition. To mitigate such limitations, we propose an extension based on prototypical networks that improves performance in low-resource named entity recognition tasks.

11:00-12:30 (Forum)
#25 Can Transformer be Too Compositional? Analysing Idiom Processing in Neural Machine Translation
Verna Dankers, Christopher G. Lucas and Ivan Titov
Unlike literal expressions, idioms' meanings do not directly follow from their parts, posing a challenge for neural machine translation (NMT). NMT models are often unable to translate idioms accurately and over-generate compositional, literal translations. In this work, we investigate whether the non-compositionality of idioms is reflected in the mechanics of the dominant NMT model, Transformer, by analysing the hidden states and attention patterns for models with English as source language and one of seven European languages as target language. When Transformer emits a non-literal translation - i.e. identifies the expression as idiomatic - the encoder processes idioms more strongly as single lexical units compared to literal expressions. This manifests in idioms' parts being grouped through attention and in reduced interaction between idioms and their context. In the decoder's cross-attention, figurative inputs result in reduced attention on source-side tokens. These results suggest that Transformer's tendency to process idioms as compositional expressions contributes to literal translations of idioms.

11:00-12:30 (Forum)
#26 Adaptive Testing and Debugging of NLP Models
Marco Tulio Ribeiro and Scott M Lundberg
Current approaches to testing and debugging NLP models rely on highly variable human creativity and extensive labor, or only work for a very restrictive class of bugs. We present AdaTest, a process which uses large scale language models (LMs) in partnership with human feedback to automatically write unit tests highlighting bugs in a target model. Such bugs are then addressed through an iterative text-fix-retest loop, inspired by traditional software development. In experiments with expert and non-expert users and commercial / research models for 8 different tasks, AdaTest makes users 5-10x more effective at finding bugs than current approaches, and helps users effectively fix bugs without adding new bugs.

11:00-12:30 (Forum)
#27 Factual Consistency of Multilingual Pretrained Language Models
Constanza Fierro and Anders Søgaard
Pretrained language models can be queried for factual knowledge, with potential applications in knowledge base acquisition and tasks that require inference. However, for that, we need to know how reliable this knowledge is, and recent work has shown that monolingual English language models lack consistency when predicting factual knowledge, that is, they fill-in-the-blank differently for paraphrases describing the same fact. In this paper, we extend the analysis of consistency to a multilingual setting. We introduce a resource, mParaRel, and investigate (i) whether multilingual language models such as mBERT and XLM-R are more consistent than their monolingual counterparts; and (ii) if such models are equally consistent across languages. We find that mBERT is as inconsistent as English BERT in English paraphrases, but that both mBERT and XLM-R exhibit a high degree of inconsistency in English and even more so for all the other 45 languages.

11:00-12:30 (Forum)
#28 Extracting Latent Steering Vectors from Pretrained Language Models
Nishant Subramani, Nivedita Suresh and Matthew E Peters
Prior work on controllable text generation has focused on learning how to control language models through trainable decoding, smart-prompt design, or fine-tuning based on a desired objective. We hypothesize that the information needed to steer the model to generate a target sentence is already encoded within the model. Accordingly, we explore a different approach altogether: extracting latent vectors directly from pretrained language model decoders without fine-tuning. Experiments show that there exist steering vectors, which, when added to the hidden states of the language model, generate a target sentence nearly perfectly (> 99 BLEU) for English sentences from a variety of domains. We show that vector arithmetic can be used for unsupervised sentiment transfer on the Yelp sentiment benchmark, with performance comparable to models tailored to this task. We find that distances between steering vectors reflect sentence similarity when evaluated on a textual similarity benchmark (STS-B), outperforming pooled hidden states of models. Finally, we present an analysis of the intrinsic properties of the steering vectors. Taken together, our results suggest that frozen LMs can be effectively controlled through their latent steering space.

11:00-12:30 (Forum)
#29 A Novel Perspective to Look At Attention: Bi-level Attention-based Explainable Topic Modeling for News Classification
Dairui Liu, Derek Greene and Ruihai Dong
Many recent deep learning-based solutions have adopted the attention mechanism in various tasks in the field of NLP. However, the inherent

characteristics of deep learning models and the flexibility of the attention mechanism increase the models' complexity, thus leading to challenges in model explainability. To address this challenge, we propose a novel practical framework by utilizing a two-tier attention architecture to decouple the complexity of explanation and the decision-making process. We apply it in the context of a news article classification task. The experiments on two large-scaled news corpora demonstrate that the proposed model can achieve competitive performance with many state-of-the-art alternatives and illustrate its appropriateness from an explainability perspective. We release the source code here[1].

11:00-12:30 (Forum)
#30 Reframing Instructional Prompts to GPTk's Language
Daniel Khashabi, Chitta Baral, Yejin Choi and Hannaneh Hajishirzi
What kinds of instructional prompts are easier to follow for Language Models (LMs)? We study this question by conducting extensive empirical analysis that shed light on important features of successful instructional prompts. Specifically, we study several classes of reframing techniques for manual reformulation of prompts into more effective ones. Some examples include decomposing a complex task instruction into multiple simpler tasks or itemizing instructions into sequential steps. Our experiments compare the zero-shot and few-shot performance of LMs prompted with reframed instructions on 12 NLP tasks across 6 categories. Compared with original instructions, our reframed instructions lead to significant improvements across LMs with different sizes. For example, the same reframed prompts boost few-shot performance of GPT3-series and GPT2-series by 12.5

11:00-12:30 (Forum)
#31 Does BERT really agree ? Fine-grained Analysis of Lexical Dependence on a Syntactic Task
Karim Lasri, Alessandro Lenci and Thierry Poibeau
Although transformer-based Neural Language Models demonstrate impressive performance on a variety of tasks, their generalization abilities are not well understood. They have been shown to perform strongly on subject-verb number agreement in a wide array of settings, suggesting that they learned to track syntactic dependencies during their training even without explicit supervision. In this paper, we examine the extent to which BERT is able to perform lexically-independent subject-verb number agreement (NA) on targeted syntactic templates. To do so, we disrupt the lexical patterns found in naturally occurring stimuli for each targeted structure in a novel fine-grained analysis of BERT's behavior. Our results on nonce sentences suggest that the model generalizes well for simple templates, but fails to perform lexically-independent syntactic generalization when as little as one attractor is present.

11:00-12:30 (Forum)
#32 Systematicity, Compositionality and Transitivity of Deep NLP Models: a Metamorphic Testing Perspective
Edoardo Manino, Julia Rozanova, Danilo Carvalho, Andre Freitas and Lucas Carvalho Cordeiro
Metamorphic testing has recently been used to check the safety of neural NLP models. Its main advantage is that it does not rely on a ground truth to generate test cases. However, existing studies are mostly concerned with robustness-like metamorphic relations, limiting the scope of linguistic properties they can test. We propose three new classes of metamorphic relations, which address the properties of systematicity, compositionality and transitivity. Unlike robustness, our relations are defined over multiple source inputs, thus increasing the number of test cases that we can produce by a polynomial factor. With them, we test the internal consistency of state-of-the-art NLP models, and show that they do not always behave according to their expected linguistic properties. Lastly, we introduce a novel graphical notation that efficiently summarises the inner structure of metamorphic relations.

11:00-12:30 (Forum)
#33 Local Structure Matters Most: Perturbation Study in NLU
Louis Clouatre, Prasanna Parthasarathi, Amal Zouaq and Sarath Chandar
Recent research analyzing the sensitivity of natural language understanding models to word-order perturbations has shown that neural models are surprisingly insensitive to the order of words. In this paper, we investigate this phenomenon by developing order-altering perturbations on the order of words, subwords, and characters to analyze their effect on neural models' performance on language understanding tasks. We experiment with measuring the impact of perturbations to the local neighborhood of characters and global position of characters in the perturbed texts and observe that perturbation functions found in prior literature only affect the global ordering while the local ordering remains relatively unperturbed. We empirically show that neural models, invariant of their inductive biases, pretraining scheme, or the choice of tokenization, mostly rely on the local structure of text to build understanding and make limited use of the global structure.

11:00-12:30 (Forum)
#34 IsoScore: Measuring the Uniformity of Embedding Space Utilization
William Rudman, Nate Gillman, Taylor Rayne and Carsten Eickhoff
The recent success of distributed word representations has led to an increased interest in analyzing the properties of their spatial distribution. Several studies have suggested that contextualized word embedding models do not isotropically project tokens into vector space. However, current methods designed to measure isotropy, such as average random cosine similarity and the partition score, have not been thoroughly analyzed and are not appropriate for measuring isotropy. We propose IsoScore: a novel tool that quantifies the degree to which a point cloud uniformly utilizes the ambient vector space. Using rigorously designed tests, we demonstrate that IsoScore is the only tool available in the literature that accurately measures how uniformly distributed variance is across dimensions in vector space. Additionally, we use IsoScore to challenge a number of recent conclusions in the NLP literature that have been derived using brittle metrics of isotropy. We caution future studies from using existing tools to measure isotropy in contextualized embedding space as resulting conclusions will be misleading or altogether inaccurate.

11:00-12:30 (Forum)
#35 Detection of Adversarial Examples in Text Classification: Benchmark and Baseline via Robust Density Estimation
KiYoon Yoo, Jangho Kim, Jiho Jang and Nojun Kwak
Word-level adversarial attacks have shown success in NLP models, drastically decreasing the performance of transformer-based models in recent years. As a countermeasure, adversarial defense has been explored, but relatively few efforts have been made to detect adversarial examples. However, detecting adversarial examples may be crucial for automated tasks (e.g. review sentiment analysis) that wish to amass information about a certain population and additionally be a step towards a robust defense system. To this end, we release a dataset for four popular attack methods on four datasets and four models to encourage further research in this field. Along with it, we propose a competitive

[1] https://github.com/Ruixinhua/BATM

baseline based on density estimation that has the highest AUC on 29 out of 30 dataset-attack-model combinations. The source code is released (https://github.com/bangawayoo/adversarial-examples-in-text-classification).

11:00-12:30 (Forum)
#36 On the data requirements of probing
Zining Zhu, Jixuan Wang, Bai Li and Frank Rudzicz
As large and powerful neural language models are developed, researchers have been increasingly interested in developing diagnostic tools to probe them. There are many papers with conclusions of the form "observation X is found in model Y", using their own datasets with varying sizes. Larger probing datasets bring more reliability, but are also expensive to collect. There is yet to be a quantitative method for estimating reasonable probing dataset sizes. We tackle this omission in the context of comparing two probing configurations: after we have collected a small dataset from a pilot study, how many additional data samples are sufficient to distinguish two different configurations? We present a novel method to estimate the required number of data samples in such experiments and, across several case studies, we verify that our estimations have sufficient statistical power. Our framework helps to systematically construct probing datasets to diagnose neural NLP models.

11:00-12:30 (Forum)
#37 Learning Disentangled Representations of Negation and Uncertainty
Jake A Vasilakes, Chrysoula Zerva, Makoto Miwa and Sophia Ananiadou
Negation and uncertainty modeling are long-standing tasks in natural language processing. Linguistic theory postulates that expressions of negation and uncertainty are semantically independent from each other and the content they modify. However, previous works on representation learning do not explicitly model this independence. We therefore attempt to disentangle the representations of negation, uncertainty, and content using a Variational Autoencoder. We find that simply supervising the latent representations results in good disentanglement, but auxiliary objectives based on adversarial learning and mutual information minimization can provide additional disentanglement gains.

11:00-12:30 (Forum)
#38 Training Text-to-Text Transformers with Privacy Guarantees
Natalia Ponomareva, Jasmijn Bastings and Sergei Vassilvitskii
Recent advances in NLP often stem from large transformer-based pre-trained models, which rapidly grow in size and use more and more training data. Such models are often released to the public so that end users can fine-tune them on a task dataset. While it is common to treat pre-training data as public, it may still contain personally identifiable information (PII), such as names, phone numbers, and copyrighted material. Recent findings show that the capacity of these models allows them to memorize parts of the training data, and suggest differentially private (DP) training as a potential mitigation. While there is recent work on DP fine-tuning of NLP models, the effects of DP pre-training are less well understood: it is not clear how downstream performance is affected by DP pre-training, and whether DP pre-training mitigates some of the memorization concerns. We focus on T5 and show that by using recent advances in JAX and XLA we can train models with DP that do not suffer a large drop in pre-training utility, nor in training speed, and can still be fine-tuned to high accuracies on downstream tasks (e.g. GLUE). Moreover, we show that T5's span corruption is a good defense against data memorization.

11:00-12:30 (Forum)
#39 Understanding Gender Bias in Knowledge Base Embeddings
Yupei Du, Qi Zheng, Yuanbin Wu, Man Lan, Yan Yang and Meirong Ma
Knowledge base (KB) embeddings have been shown to contain gender biases. In this paper, we study two questions regarding these biases: how to quantify them, and how to trace their origins in KB? Specifically, first, we develop two novel bias measures respectively for a group of person entities and an individual person entity. Evidence of their validity is observed by comparison with real-world census data. Second, we use the influence function to inspect the contribution of each triple in KB to the overall group bias. To exemplify the potential applications of our study, we also present two strategies (by adding and removing KB triples) to mitigate gender biases in KB embeddings.

11:00-12:30 (Forum)
[TACL] **#40 Word Acquisition in Neural Language Models**
Tyler Chang and Benjamin Bergen

11:00-12:30 (Forum)
#41 Word Order Does Matter and Shuffled Language Models Know It
Mostafa Abdou, Vinit Ravishankar, Artur Kulmizev and Anders Søgaard
Recent studies have shown that language models pretrained and/or fine-tuned on randomly permuted sentences exhibit competitive performance on GLUE, putting into question the importance of word order information. Somewhat counter-intuitively, some of these studies also report that position embeddings appear to be crucial for models' good performance with shuffled text. We probe these language models for word order information and investigate what position embeddings learned from shuffled text encode, showing that these models retain a notion of word order information. We show this is in part due to a subtlety in how shuffling is implemented in previous work – before rather than after subword segmentation. Surprisingly, we find even Language models trained on text shuffled after subword segmentation retain some semblance of information about word order because of the statistical dependencies between sentence length and unigram probabilities. Finally, we show that beyond GLUE, a variety of language understanding tasks do require word order information, often to an extent that cannot be learned through fine-tuning.

11:00-12:30 (Forum)
[DEMO] **DataLab: A Platform for Data Analysis and Intervention**
Yang Xiao, Jinlan Fu, Weizhe Yuan, Vijay Viswanathan, Zhoumianze Liu, Yixin Liu, Graham Neubig and Pengfei Liu
Despite data's crucial role in machine learning, most existing tools and research tend to focus on systems on top of existing data rather than how to interpret and manipulate data. In this paper, we propose DataLab, a unified data-oriented platform that not only allows users to interactively analyze the characteristics of data but also provides a standardized interface so that many data processing operations can be provided within a unified interface. Additionally, in view of the ongoing surge in the proliferation of datasets, DataLab has features for dataset recommendation and global vision analysis that help researchers form a better view of the data ecosystem. So far, DataLab covers 1,300 datasets and 3,583 of its transformed version, where 313 datasets support different types of analysis (e.g., with respect to gender bias) with the help of 119M samples annotated by 318 feature functions. DataLab is under active development and will be supported going forward. We

have released a web platform, web API, Python SDK, and PyPI published package, which hopefully, can meet the diverse needs of researchers.

Poster Session 1: NLP Applications

11:00-12:30 (Forum)

11:00-12:30 (Forum)
#42 What does the sea say to the shore? A BERT based DST style approach for speaker to dialogue attribution in novels
Carolina Cuesta-Lazaro, Animesh Prasad and Trevor Wood
We present a complete pipeline to extract characters in a novel and link them to their direct-speech utterances. Our model is divided into three independent components: extracting direct-speech, compiling a list of characters, and attributing those characters to their utterances. Although we find that existing systems can perform the first two tasks accurately, attributing characters to direct speech is a challenging problem due to the narrator's lack of explicit character mentions, and the frequent use of nominal and pronominal coreference when such explicit mentions are made. We adapt the progress made on Dialogue State Tracking to tackle a new problem: attributing speakers to dialogues. This is the first application of deep learning to speaker attribution, and it shows that is possible to overcome the need for the hand-crafted features and rules used in the past. Our full pipeline improves the performance of state-of-the-art models by a relative 50% in F1-score.

11:00-12:30 (Forum)
#43 On the Robustness of Offensive Language Classifiers
Jonathan Rusert, Zubair Shafiq and Padmini Srinivasan
Social media platforms are deploying machine learning based offensive language classification systems to combat hateful, racist, and other forms of offensive speech at scale. However, despite their real-world deployment, we do not yet comprehensively understand the extent to which offensive language classifiers are robust against adversarial attacks. Prior work in this space is limited to studying robustness of offensive language classifiers against primitive attacks such as misspellings and extraneous spaces. To address this gap, we systematically analyze the robustness of state-of-the-art offensive language classifiers against more crafty adversarial attacks that leverage greedy- and attention-based word selection and context-aware embeddings for word replacement. Our results on multiple datasets show that these crafty adversarial attacks can degrade the accuracy of offensive language classifiers by more than 50

11:00-12:30 (Forum)
#44 Adversarial Authorship Attribution for Deobfuscation
Wanyue Zhai, Jonathan Rusert, Zubair Shafiq and Padmini Srinivasan
Recent advances in natural language processing have enabled powerful privacy-invasive authorship attribution. To counter authorship attribution, researchers have proposed a variety of rule-based and learning-based text obfuscation approaches. However, existing authorship obfuscation approaches do not consider the adversarial threat model. Specifically, they are not evaluated against adversarially trained authorship attributors that are aware of potential obfuscation. To fill this gap, we investigate the problem of adversarial authorship attribution for deobfuscation. We show that adversarially trained authorship attributors are able to degrade the effectiveness of existing obfuscators from 20-30

11:00-12:30 (Forum)
#45 Sibylvariant Transformations for Robust Text Classification
Fabrice Y Harel-Canada, Muhammad Ali Gulzar, Nanyun Peng and Miryung Kim
The vast majority of text transformation techniques in NLP are inherently limited in their ability to expand input space coverage due to an implicit constraint to preserve the original class label. In this work, we propose the notion of sibylvariance (SIB) to describe the broader set of transforms that relax the label-preserving constraint, knowably vary the expected class, and lead to significantly more diverse input distributions. We offer a unified framework to organize all data transformations, including two types of SIB: (1) Transmutations convert one discrete kind into another, (2) Mixture Mutations blend two or more classes together. To explore the role of sibylvariance within NLP, we implemented 41 text transformations, including several novel techniques like Concept2Sentence and SentMix. Sibylvariance also enables a unique form of adaptive training that generates new input mixtures for the most confused class pairs, challenging the learner to differentiate with greater nuance. Our experiments on six benchmark datasets strongly support the efficacy of sibylvariance for generalization performance, defect detection, and adversarial robustness.

11:00-12:30 (Forum)
#46 Question Generation for Reading Comprehension Assessment by Modeling How and What to Ask
Bilal Ghanem, Lauren Lutz Coleman, Julia Rivard Dexter, Spencer McIntosh von der Ohe and Alona Fyshe
Reading is integral to everyday life, and yet learning to read is a struggle for many young learners. During lessons, teachers can use comprehension questions to increase engagement, test reading skills, and improve retention. Historically such questions were written by skilled teachers, but recently language models have been used to generate comprehension questions. However, many existing Question Generation (QG) systems focus on generating extractive questions from the text, and have no way to control the type of the generated question. In this paper, we study QG for reading comprehension where inferential questions are critical and extractive techniques cannot be used. We propose a two-step model (HTA-WTA) that takes advantage of previous datasets, and can generate questions for a specific targeted comprehension skill. We propose a new reading comprehension dataset that contains questions annotated with story-based reading comprehension skills (SBRCS), allowing for a more complete reader assessment. Across several experiments, our results show that HTA-WTA outperforms multiple strong baselines on this new dataset. We show that the HTA-WTA model tests for strong SCRS by asking deep inferential questions.

11:00-12:30 (Forum)
[TACL] **#47 A Neighbourhood Framework for Resource-Lean Content Flagging**
Momchil Hardalov, Sheikh Sarwar, Dimitrina Zlatkova, Yoan Dinkov, Isabelle Augenstein and Preslav Nakov

11:00-12:30 (Forum)
#48 TableFormer: Robust Transformer Modeling for Table-Text Encoding
Jingfeng Yang, Aditya Gupta, Shyam Upadhyay, Luheng He, Rahul Goel and Shachi Paul
Understanding tables is an important aspect of natural language understanding. Existing models for table understanding require linearization

of the table structure, where row or column order is encoded as an unwanted bias. Such spurious biases make the model vulnerable to row and column order perturbations. Additionally, prior work has not thoroughly modeled the table structures or table-text alignments, hindering the table-text understanding ability. In this work, we propose a robust and structurally aware table-text encoding architecture TableFormer, where tabular structural biases are incorporated completely through learnable attention biases. TableFormer is (1) strictly invariant to row and column orders, and, (2) could understand tables better due to its tabular inductive biases. Our evaluations showed that TableFormer outperforms strong baselines in all settings on SQA, WTQ and TabFact table reasoning datasets, and achieves state-of-the-art performance on SQA, especially when facing answer-invariant row and column order perturbations (6

11:00-12:30 (Forum)
#49 Differentiable Multi-Agent Actor-Critic for Multi-Step Radiology Report Summarization
Sanjeev Kumar Karn, Ning Liu, Hinrich Schuetze and Oladimeji Farri
The IMPRESSIONS section of a radiology report about an imaging study is a summary of the radiologist's reasoning and conclusions, and it also aids the referring physician in confirming or excluding certain diagnoses. A cascade of tasks are required to automatically generate an abstractive summary of the typical information-rich radiology report. These tasks include acquisition of salient content from the report and generation of a concise, easily consumable IMPRESSIONS section. Prior research on radiology report summarization has focused on single-step end-to-end models – which subsume the task of salient content acquisition. To fully explore the cascade structure and explainability of radiology report summarization, we introduce two innovations. First, we design a two-step approach: extractive summarization followed by abstractive summarization. Second, we additionally break down the extractive part into two independent tasks: extraction of salient (1) sentences and (2) keywords. Experiments on a publicly available radiology report dataset show our novel approach leads to a more precise summary compared to single-step and to two-step-with-single-extractive-process baselines with an overall improvement in F1 score of 3-4

11:00-12:30 (Forum)
#50 From the Detection of Toxic Spans in Online Discussions to the Analysis of Toxic-to-Civil Transfer
John Pavlopoulos, Leo Laugier, Alexandros Xenos, Jeffrey S. Sorensen and Ion Androutsopoulos
We study the task of toxic spans detection, which concerns the detection of the spans that make a text toxic, when detecting such spans is possible. We introduce a dataset for this task, ToxicSpans, which we release publicly. By experimenting with several methods, we show that sequence labeling models perform best, but methods that add generic rationale extraction mechanisms on top of classifiers trained to predict if a post is toxic or not are also surprisingly promising. Finally, we use ToxicSpans and systems trained on it, to provide further analysis of state-of-the-art toxic to non-toxic transfer systems, as well as of human performance on that latter task. Our work highlights challenges in finer toxicity detection and mitigation.

11:00-12:30 (Forum)
#51 Clickbait Spoiling via Question Answering and Passage Retrieval
Matthias Hagen, Maik Fröbe, Artur Jurk and Martin Potthast
We introduce and study the task of clickbait spoiling: generating a short text that satisfies the curiosity induced by a clickbait post. Clickbait links to a web page and advertises its contents by arousing curiosity instead of providing an informative summary. Our contributions are approaches to classify the type of spoiler needed (i.e., a phrase or a passage), and to generate appropriate spoilers. A large-scale evaluation and error analysis on a new corpus of 5,000 manually spoiled clickbait posts—the Webis Clickbait Spoiling Corpus 2022—shows that our spoiler type classifier achieves an accuracy of 80%, while the question answering model DeBERTa-large outperforms all others in generating spoilers for both types.

11:00-12:30 (Forum)
#52 Interpretability for Language Learners Using Example-Based Grammatical Error Correction
Masahiro Kaneko, Sho Takase, Ayana Niwa and Naoaki Okazaki
Grammatical Error Correction (GEC) should not focus only on high accuracy of corrections but also on interpretability for language learning. However, existing neural-based GEC models mainly aim at improving accuracy, and their interpretability has not been explored. A promising approach for improving interpretability is an example-based method, which uses similar retrieved examples to generate corrections. In addition, examples are beneficial in language learning, helping learners understand the basis of grammatically incorrect/correct texts and improve their confidence in writing. Therefore, we hypothesize that incorporating an example-based method into GEC can improve interpretability as well as support language learners. In this study, we introduce an Example-Based GEC (EB-GEC) that presents examples to language learners as a basis for a correction result. The examples consist of pairs of correct and incorrect sentences similar to a given input and its predicted correction. Experiments demonstrate that the examples presented by EB-GEC help language learners decide to accept or refuse suggestions from the GEC output. Furthermore, the experiments also show that retrieved examples improve the accuracy of corrections.

11:00-12:30 (Forum)
#53 Your Answer is Incorrect... Would you like to know why? Introducing a Bilingual Short Answer Feedback Dataset
Anna Filighera, Siddharth Singh Parihar, Tim Steuer, Tobias Meuser and Sebastian Ochs
Handing in a paper or exercise and merely receiving "bad" or "incorrect" as feedback is not very helpful when the goal is to improve. Unfortunately, this is currently the kind of feedback given by Automatic Short Answer Grading (ASAG) systems. One of the reasons for this is a lack of content-focused elaborated feedback datasets. To encourage research on explainable and understandable feedback systems, we present the Short Answer Feedback dataset (SAF). Similar to other ASAG datasets, SAF contains learner responses and reference answers to German and English questions. However, instead of only assigning a label or score to the learners' answers, SAF also contains elaborated feedback explaining the given score. Thus, SAF enables supervised training of models that grade answers and explain where and why mistakes were made. This paper discusses the need for enhanced feedback models in real-world pedagogical scenarios, describes the dataset annotation process, gives a comprehensive analysis of SAF, and provides T5-based baselines for future comparison.

11:00-12:30 (Forum)
#54 Ensembling and Knowledge Distilling of Large Sequence Taggers for Grammatical Error Correction
Maksym Tarnavskyi, Artem Chernodub and Kostiantyn Omelianchuk
In this paper, we investigate improvements to the GEC sequence tagging architecture with a focus on ensembling of recent cutting-edge Transformer-based encoders in Large configurations. We encourage ensembling models by majority votes on span-level edits because this approach is tolerant to the model architecture and vocabulary size. Our best ensemble achieves a new SOTA result with an $F_{0.5}$ score of 76.05 on BEA-2019 (test), even without pre-training on synthetic datasets. In addition, we perform knowledge distillation with a trained ensemble to generate new synthetic training datasets, "Troy-Blogs" and "Troy-1BW". Our best single sequence tagging model that is pretrained on the generated Troy- datasets in combination with the publicly available synthetic PIE dataset achieves a near-SOTA result with an $F_{0.5}$ score of

73.21 on BEA-2019 (test). The code, datasets, and trained models are publicly available.

11:00-12:30 (Forum)
#55 Few-Shot Tabular Data Enrichment Using Fine-Tuned Transformer Architectures
Asaf Harari and Gilad Katz
The enrichment of tabular datasets using external sources has gained significant attention in recent years. Existing solutions, however, either ignore external unstructured data completely or devise dataset-specific solutions. In this study we proposed Few-Shot Transformer based Enrichment (FeSTE), a generic and robust framework for the enrichment of tabular datasets using unstructured data. By training over multiple datasets, our approach is able to develop generic models that can be applied to additional datasets with minimal training (i.e., few-shot). Our approach is based on an adaptation of BERT, for which we present a novel fine-tuning approach that reformulates the tuples of the datasets as sentences. Our evaluation, conducted on 17 datasets, shows that FeSTE is able to generate high quality features and significantly outperform existing fine-tuning solutions.

11:00-12:30 (Forum)
#56 Improving Generalizability in Implicitly Abusive Language Detection with Concept Activation Vectors
Isar Nejadgholi, Kathleen C. Fraser and Svetlana Kiritchenko
Robustness of machine learning models on ever-changing real-world data is critical, especially for applications affecting human well-being such as content moderation. New kinds of abusive language continually emerge in online discussions in response to current events (e.g., COVID-19), and the deployed abuse detection systems should be updated regularly to remain accurate. In this paper, we show that general abusive language classifiers tend to be fairly reliable in detecting out-of-domain explicitly abusive utterances but fail to detect new types of more subtle, implicit abuse. Next, we propose an interpretability technique, based on the Testing Concept Activation Vector (TCAV) method from computer vision, to quantify the sensitivity of a trained model to the human-defined concepts of explicit and implicit abusive language, and use that to explain the generalizability of the model on new data, in this case, COVID-related anti-Asian hate speech. Extending this technique, we introduce a novel metric, Degree of Explicitness, for a single instance and show that the new metric is beneficial in suggesting out-of-domain unlabeled examples to effectively enrich the training data with informative, implicitly abusive texts.

11:00-12:30 (Forum)
#57 Legal Judgment Prediction via Event Extraction with Constraints
Yi Feng, Chuanyi Li and Vincent Ng
While significant progress has been made on the task of Legal Judgment Prediction (LJP) in recent years, the incorrect predictions made by SOTA LJP models can be attributed in part to their failure to (1) locate the key event information that determines the judgment, and (2) exploit the cross-task consistency constraints that exist among the subtasks of LJP. To address these weaknesses, we propose EPM, an Event-based Prediction Model with constraints, which surpasses existing SOTA models in performance on a standard LJP dataset.

11:00-12:30 (Forum)
#58 RNSum: A Large-Scale Dataset for Automatic Release Note Generation via Commit Logs Summarization
Hisashi Kamezawa, Noriki Nishida, Nobuyuki Shimizu, Takashi Miyazaki and Hideki Nakayama
A release note is a technical document that describes the latest changes to a software product and is crucial in open source software development. However, it still remains challenging to generate release notes automatically. In this paper, we present a new dataset called RNSum, which contains approximately 82,000 English release notes and the associated commit messages derived from the online repositories in GitHub. Then, we propose classwise extractive-then-abstractive/abstractive summarization approaches to this task, which can employ a modern transformer-based seq2seq network like BART and can be applied to various repositories without specific constraints. The experimental results on the RNSum dataset show that the proposed methods can generate less noisy release notes at higher coverage than the baselines. We also observe that there is a significant gap in the coverage of essential information when compared to human references. Our dataset and the code are publicly available.

11:00-12:30 (Forum)
#59 Letters From the Past: Modeling Historical Sound Change Through Diachronic Character Embeddings
Sidsel Boldsen and Patrizia Paggio
While a great deal of work has been done on NLP approaches to lexical semantic change detection, other aspects of language change have received less attention from the NLP community. In this paper, we address the detection of sound change through historical spelling. We propose that a sound change can be captured by comparing the relative distance through time between the distributions of the characters involved before and after the change has taken place. We model these distributions using PPMI character embeddings. We verify this hypothesis in synthetic data and then test the method's ability to trace the well-known historical change of lenition of plosives in Danish historical sources. We show that the models are able to identify several of the changes under consideration and to uncover meaningful contexts in which they appeared. The methodology has the potential to contribute to the study of open questions such as the relative chronology of sound shifts and their geographical distribution.

11:00-12:30 (Forum)
#60 Learning and Evaluating Character Representations in Novels
Naoya Inoue, Charuta Pethe, Allen Kim and Steven Skiena
We address the problem of learning fixed-length vector representations of characters in novels. Recent advances in word embeddings have proven successful in learning entity representations from short texts, but fall short on longer documents because they do not capture full book-level information. To overcome the weakness of such text-based embeddings, we propose two novel methods for representing characters: (i) graph neural network-based embeddings from a full corpus-based character network; and (ii) low-dimensional embeddings constructed from the occurrence pattern of characters in each novel. We test the quality of these character embeddings using a new benchmark suite to evaluate character representations, encompassing 12 different tasks. We show that our representation techniques combined with text-based embeddings lead to the best character representations, outperforming text-based embeddings in four tasks. Our dataset and evaluation script will be made publicly available to stimulate additional work in this area.

11:00-12:30 (Forum)
[TACL] **#61 A Survey on Automated Fact-Checking**

Zhijiang Guo, Michael Schlichtkrull and Andreas Vlachos

11:00-12:30 (Forum)
#62 Learning to Describe Solutions for Bug Reports Based on Developer Discussions
Sheena Panthaplackel, Junyi Jessy Li, Milos Gligoric and Ray Mooney
When a software bug is reported, developers engage in a discussion to collaboratively resolve it. While the solution is likely formulated within the discussion, it is often buried in a large amount of text, making it difficult to comprehend and delaying its implementation. To expedite bug resolution, we propose generating a concise natural language description of the solution by synthesizing relevant content within the discussion, which encompasses both natural language and source code. We build a corpus for this task using a novel technique for obtaining noisy supervision from repository changes linked to bug reports, with which we establish benchmarks. We also design two systems for generating a description during an ongoing discussion by classifying when sufficient context for performing the task emerges in real-time. With automated and human evaluation, we find this task to form an ideal testbed for complex reasoning in long, bimodal dialogue context.

11:00-12:30 (Forum)
#63 LexGLUE: A Benchmark Dataset for Legal Language Understanding in English
Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz and Nikolaos Aletras
Laws and their interpretations, legal arguments and agreements are typically expressed in writing, leading to the production of vast corpora of legal text. Their analysis, which is at the center of legal practice, becomes increasingly elaborate as these collections grow in size. Natural language understanding (NLU) technologies can be a valuable tool to support legal practitioners in these endeavors. Their usefulness, however, largely depends on whether current state-of-the-art models can generalize across various tasks in the legal domain. To answer this currently open question, we introduce the Legal General Language Understanding Evaluation (LexGLUE) benchmark, a collection of datasets for evaluating model performance across a diverse set of legal NLU tasks in a standardized way. We also provide an evaluation and analysis of several generic and legal-oriented models demonstrating that the latter consistently offer performance improvements across multiple tasks.

11:00-12:30 (Forum)
#64 It is AI's Turn to Ask Humans a Question: Question-Answer Pair Generation for Children's Story Books
Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Mo Yu and Ying Xu
Existing question answering (QA) techniques are created mainly to answer questions asked by humans. But in educational applications, teachers often need to decide what questions they should ask, in order to help students to improve their narrative understanding capabilities. We design an automated question-answer generation (QAG) system for this education scenario: given a story book at the kindergarten to eighth-grade level as input, our system can automatically generate QA pairs that are capable of testing a variety of dimensions of a student's comprehension skills. Our proposed QAG model architecture is demonstrated using a new expert-annotated FairytaleQA dataset, which has 278 child-friendly storybooks with 10,580 QA pairs. Automatic and human evaluations show that our model outperforms state-of-the-art QAG baseline systems. On top of our QAG system, we also start to build an interactive story-telling application for the future real-world deployment in this educational scenario.

11:00-12:30 (Forum)
#65 Learning to Reason Deductively: Math Word Problem Solving as Complex Relation Extraction
Zhanming Jie, Jierui Li and Wei Lu
Solving math word problems requires deductive reasoning over the quantities in the text. Various recent research efforts mostly relied on sequence-to-sequence or sequence-to-tree models to generate mathematical expressions without explicitly performing relational reasoning between quantities in the given context. While empirically effective, such approaches typically do not provide explanations for the generated expressions. In this work, we view the task as a complex relation extraction problem, proposing a novel approach that presents explainable deductive reasoning steps to iteratively construct target expressions, where each step involves a primitive operation over two quantities defining their relation. Through extensive experiments on four benchmark datasets, we show that the proposed model significantly outperforms existing strong baselines. We further demonstrate that the deductive procedure not only presents more explainable steps but also enables us to make more accurate predictions on questions that require more complex reasoning.

11:00-12:30 (Forum)
#66 The AI Doctor Is In: A Survey of Task-Oriented Dialogue Systems for Healthcare Applications
Mina Valizadeh and Natalie Parde
Task-oriented dialogue systems are increasingly prevalent in healthcare settings, and have been characterized by a diverse range of architectures and objectives. Although these systems have been surveyed in the medical community from a non-technical perspective, a systematic review from a rigorous computational perspective has to date remained noticeably absent. As a result, many important implementation details of healthcare-oriented dialogue systems remain limited or underspecified, slowing the pace of innovation in this area. To fill this gap, we investigated an initial pool of 4070 papers from well-known computer science, natural language processing, and artificial intelligence venues, identifying 70 papers discussing the system-level implementation of task-oriented dialogue systems for healthcare applications. We conducted a comprehensive technical review of these papers, and present our key findings including identified gaps and corresponding recommendations.

11:00-12:30 (Forum)
[DEMO] **QiuNiu: A Chinese Lyrics Generation System with Passage-Level Input**
Le Zhang, Rongsheng Zhang, Xiaoxi Mao and Yongzhu Chang
Lyrics generation has been a very popular application of natural language generation. Previous works mainly focused on generating lyrics based on a couple of attributes or keywords, rendering very limited control over the content of the lyrics. In this paper, we demonstrate the QiuNiu, a Chinese lyrics generation system which is conditioned on passage-level text rather than a few attributes or keywords. By using the passage-level text as input, the content of generated lyrics is expected to reflect the nuances of users' needs. The QiuNiu system supports various forms of passage-level input, such as short stories, essays, poetry. The training of it is conducted under the framework of unsupervised machine translation, due to the lack of aligned passage-level text-to-lyrics corpus. We initialize the parameters of QiuNiu with a custom pretrained Chinese GPT-2 model and adopt a two-step process to finetune the model for better alignment between passage-level text and lyrics. Additionally, a postprocess module is used to filter and rerank the generated lyrics to select the ones of highest quality. The demo video of the system is available at https://youtu.be/OCQNzahqWgM.

11:00-12:30 (Forum)
[DEMO] **PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts**
Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang and Alexander Rush
PromptSource is a system for creating, sharing, and using natural language prompts. Prompts are functions that map an example from a dataset to a natural language input and target output. Using prompts to train and query language models is an emerging area in NLP that requires new tools that let users develop and refine these prompts collaboratively. PromptSource addresses the emergent challenges in this new setting with (1) a templating language for defining data-linked prompts, (2) an interface that lets users quickly iterate on prompt development by observing outputs of their prompts on many examples, and (3) a community-driven set of guidelines for contributing new prompts to a common pool. Over 2,000 prompts for roughly 170 datasets are already available in PromptSource. PromptSource is available at https://github.com/bigscience-workshop/promptsource.

11:00-12:30 (Forum)
[DEMO] **COVID-19 Claim Radar: A Structured Claim Extraction and Tracking System**
Manling Li, Revanth Gangi Reddy, Ziqi Wang, Yi-shyuan Chiang, Tuan Lai, Pengfei Yu, Zixuan Zhang and Heng Ji
To tackle the challenge of accurate and timely communication regarding the COVID-19 pandemic, we present a COVID-19 Claim Radar to automatically extract supporting and refuting claims on a daily basis. We provide a comprehensive structured view of claims, including rich claim attributes (such as claimers and claimer affiliations) and associated knowledge elements as claim semantics (such as events, relations and entities), enabling users to explore equivalent, refuting, or supporting claims with structural evidence, such as shared claimers, similar centroid events and arguments. In order to consolidate claim structures at the corpus-level, we leverage Wikidata as the hub to merge coreferential knowledge elements. The system automatically provides users a comprehensive exposure to COVID-19 related claims, their importance, and their interconnections. The system is publicly available at GitHub and DockerHub, with complete documentation.

Poster Session 1: Student Research Workshop

11:00-12:30 (Forum)

11:00-12:30 (Forum)
#67 Darkness can not drive out darkness: Investigating Bias in Hate SpeechDetection Models
Fatma Elsafoury
It has become crucial to develop tools for automated hate speech and abuse detection. These tools would help to stop the bullies and the haters and provide a safer environment for individuals especially from marginalized groups to freely express themselves. However, recent research shows that machine learning models are biased and they might make the right decisions for the wrong reasons. In this thesis, I set out to understand the performance of hate speech and abuse detection models and the different biases that could influence them. I show that hate speech and abuse detection models are not only subject to social bias but also to other types of bias that have not been explored before. Finally, I investigate the causal effect of the social and intersectional bias on the performance and unfairness of hate speech detection models.

11:00-12:30 (Forum)
#68 What Do You Mean by Relation Extraction? A Survey on Datasets and Study on Scientific Relation Classification
Elisa Bassignana and Barbara Plank
Over the last five years, research on Relation Extraction (RE) witnessed extensive progress with many new dataset releases. At the same time, setup clarity has decreased, contributing to increased difficulty of reliable empirical evaluation (Taillé et al., 2020). In this paper, we provide a comprehensive survey of RE datasets, and revisit the task definition and its adoption by the community. We find that cross-dataset and cross-domain setups are particularly lacking. We present an empirical study on scientific Relation Classification across two datasets. Despite large data overlap, our analysis reveals substantial discrepancies in annotation. Annotation discrepancies strongly impact Relation Classification performance, explaining large drops in cross-dataset evaluations. Variation within further sub-domains exists but impacts Relation Classification only to limited degrees. Overall, our study calls for more rigour in reporting setups in RE and evaluation across multiple test sets.

11:00-12:30 (Forum)
#69 GNNer: Reducing Overlapping in Span-based NER Using Graph Neural Networks
Urchade Zaratiana, Nadi Tomeh, Pierre Holat and Thierry Charnois
There are two main paradigms for Named Entity Recognition (NER): sequence labelling and span classification. Sequence labelling aims to assign a label to each word in an input text using, for example, BIO (Begin, Inside and Outside) tagging, while span classification involves enumerating all possible spans in a text and classifying them into their labels. In contrast to sequence labelling, unconstrained span-based methods tend to assign entity labels to overlapping spans, which is generally undesirable, especially for NER tasks without nested entities. Accordingly, we propose GNNer, a framework that uses Graph Neural Networks to enrich the span representation to reduce the number of overlapping spans during prediction. Our approach reduces the number of overlapping spans compared to strong baseline while maintaining competitive metric performance. Code is available at `https://github.com/urchade/GNNer`.

11:00-12:30 (Forum)
#70 Towards Unification of Discourse Annotation Frameworks
Yingxue Fu
Discourse information is difficult to represent and annotate. Among the major frameworks for annotating discourse information, RST, PDTB and SDRT are widely discussed and used, each having its own theoretical foundation and focus. Corpora annotated under different frameworks vary considerably. To make better use of the existing discourse corpora and achieve the possible synergy of different frameworks, it is worthwhile to investigate the systematic relations between different frameworks and devise methods of unifying the frameworks. Although the issue of framework unification has been a topic of discussion for a long time, there is currently no comprehensive approach which considers unifying both discourse structure and discourse relations and evaluates the unified framework intrinsically and extrinsically. We plan to use automatic means for the unification task and evaluate the result with structural complexity and downstream tasks. We will also explore

the application of the unified framework in multi-task learning and graphical models.

11:00-12:30 (Forum)
#71 What do Models Learn From Training on More Than Text? Measuring Visual Commonsense Knowledge
Lovisa Hagström and Richard Johansson
There are limitations in learning language from text alone. Therefore, recent focus has been on developing multimodal models. However, few benchmarks exist that can measure what language models learn about language from multimodal training. We hypothesize that training on a visual modality should improve on the visual commonsense knowledge in language models. Therefore, we introduce two evaluation tasks for measuring visual commonsense knowledge in language models (code publicly available at: github.com/lovhag/measure-visual-commonsense-knowledge) and use them to evaluate different multimodal models and unimodal baselines. Primarily, we find that the visual commonsense knowledge is not significantly different between the multimodal models and unimodal baseline models trained on visual text data.

11:00-12:30 (Forum)
#72 TeluguNER: Leveraging Multi-Domain Named Entity Recognition with Deep Transformers
Suma Reddy Duggenpudi, Subba Reddy Oota, Mounika Marreddy and Radhika Mamidi
Named Entity Recognition (NER) is a successful and well-researched problem in English due to the availability of resources. The transformer models, specifically the masked-language models (MLM), have shown remarkable performance in NER during recent times. With growing data in different online platforms, there is a need for NER in other languages too. NER remains to be underexplored in Indian languages due to the lack of resources and tools. Our contributions in this paper include (i) Two annotated NER datasets for the Telugu language in multiple domains: Newswire Dataset (ND) and Medical Dataset (MD), and we combined ND and MD to form Combined Dataset (CD) (ii) Comparison of the finetuned Telugu pretrained transformer models (BERT-Te, RoBERTa-Te, and ELECTRA-Te) with other baseline models (CRF, LSTM-CRF, and BiLSTM-CRF) (iii) Further investigation of the performance of Telugu pretrained transformer models against the multilingual models mBERT, XLM-R, and IndicBERT. We find that pretrained Telugu language models (BERT-Te and RoBERTa) outperform the existing pretrained multilingual and baseline models in NER. On a large dataset (CD) of 38,363 sentences, the BERT-Te achieves a high F1-score of 0.80 (entity-level) and 0.75 (token-level). Further, these pretrained Telugu models have shown state-of-the-art performance on various existing Telugu NER datasets. We open-source our dataset, pretrained models, and code.

11:00-12:30 (Forum)
#73 Mining Logical Event Schemas From Pre-Trained Language Models
Lane Lawley and Lenhart Schubert
We present NESL (the Neuro-Episodic Schema Learner), an event schema learning system that combines large language models, FrameNet parsing, a powerful logical representation of language, and a set of simple behavioral schemas meant to bootstrap the learning process. In lieu of a pre-made corpus of stories, our dataset is a continuous feed of "situation samples" from a pre-trained language model, which are then parsed into FrameNet frames, mapped into simple behavioral schemas, and combined and generalized into complex, hierarchical schemas for a variety of everyday scenarios. We show that careful sampling from the language model can help emphasize stereotypical properties of situations and de-emphasize irrelevant details, and that the resulting schemas specify situations more comprehensively than those learned by other systems.

11:00-12:30 (Forum)
#74 Discourse on ASR Measurement: Introducing the ARPOCA Assessment Tool
Megan Merz and Olga Scrivner
Automatic speech recognition (ASR) has evolved from a pipeline architecture with pronunciation dictionaries, phonetic features and language models to the end-to-end systems performing a direct translation from a raw waveform into a word sequence. With the increase in accuracy and the availability of pre-trained models, the ASR systems are now omnipresent in our daily applications. On the other hand, the models' interpretability and their computational cost have become more challenging, particularly when dealing with less-common languages or identifying regional variations of speakers. This research proposal will follow a four-stage process: 1) Proving an overview of acoustic features and feature extraction algorithms; 2) Exploring current ASR models, tools, and performance assessment techniques; 3) Aligning features with interpretable phonetic transcripts; and 4) Designing a prototype ARPOCA to increase awareness of regional language variation and improve models feedback by developing a semi-automatic acoustic features extraction using PRAAT in conjunction with phonetic transcription.

11:00-12:30 (Forum)
#75 Pretrained Knowledge Base Embeddings for improved Sentential Relation Extraction
andrea papaluca, Daniel Krefl, Hanna Suominen and Artem Lenskiy
In this work we put forward to combine pretrained knowledge base graph embeddings with transformer based language models to improve performance on the sentential Relation Extraction task in natural language processing. Our proposed model is based on a simple variation of existing models to incorporate off-task pretrained graph embeddings with an on-task finetuned BERT encoder. We perform a detailed statistical evaluation of the model on standard datasets. We provide evidence that the added graph embeddings improve the performance, making such a simple approach competitive with the state-of-the-art models that perform explicit on-task training of the graph embeddings. Furthermore, we ob- serve for the underlying BERT model an interesting power-law scaling behavior between the variance of the F1 score obtained for a relation class and its support in terms of training examples.

11:00-12:30 (Forum)
#76 On the Locality of Attention in Direct Speech Translation
Belen Alastruey, Javier Ferrando, Gerard I. Gállego and Marta R. Costa-jussà
Transformers have achieved state-of-the-art results across multiple NLP tasks. However, the self-attention mechanism complexity scales quadratically with the sequence length, creating an obstacle for tasks involving long sequences, like in the speech domain. In this paper, we discuss the usefulness of self-attention for Direct Speech Translation. First, we analyze the layer-wise token contributions in the self-attention of the encoder, unveiling local diagonal patterns. To prove that some attention weights are avoidable, we propose to substitute the standard self-attention with a local efficient one, setting the amount of context used based on the results of the analysis. With this approach, our model matches the baseline performance, and improves the efficiency by skipping the computation of those weights that standard attention discards.

11:00-12:30 (Forum)
#77 Extraction of Diagnostic Reasoning Relations for Clinical Knowledge Graphs
Vimig Socrates
Clinical knowledge graphs lack meaningful diagnostic relations (e.g. comorbidities, sign/symptoms), limiting their ability to represent real-world diagnostic processes. Previous methods in biomedical relation extraction have focused on concept relations, such as gene-disease and disease-drug, and largely ignored clinical processes. In this thesis, we leverage a clinical reasoning ontology and propose methods to extract such relations from a physician-facing point-of-care reference wiki and consumer health resource texts. Given the lack of data labeled with diagnostic relations, we also propose new methods of evaluating the correctness of extracted triples in the zero-shot setting. We describe a process for the intrinsic evaluation of new facts by triple confidence filtering and clinician manual review, as well extrinsic evaluation in the form of a differential diagnosis prediction task.

11:00-12:30 (Forum)
#78 A Checkpoint on Multilingual Misogyny Identification
Arianna Muti and Alberto Barrón-Cedeño
We address the problem of identifying misogyny in tweets in mono and multilingual settings in three languages: English, Italian, and Spanish. We explore model variations considering single and multiple languages both in the pre-training of the transformer and in the training of the downstream task to explore the feasibility of detecting misogyny through a transfer learning approach across multiple languages. That is, we train monolingual transformers with monolingual data, and multilingual transformers with both monolingual and multilingual data. Our models reach state-of-the-art performance on all three languages. The single-language BERT models perform the best, closely followed by different configurations of multilingual BERT models. The performance drops in zero-shot classification across languages. Our error analysis shows that multilingual and monolingual models tend to make the same mistakes.

11:00-12:30 (Forum)
#79 EDDIE: End-to-End Humor Generation Model with Keyword Control
Irene Lee, Anish Thite, Mohan Dodda, Xu Zeng and Diyi Yang
Humor is a common experience that has great societal value. However, humor generation is a challenging problem for the language technologies community as it involves a significant understanding of context. To address this, previous works on humor generation have employed linguistic expertise to produce jokes with a set structure. However this approach is not scalable, since there is a wide range of possible joke structures across various types of jokes (e.g. question-answer formats, one-liners, etc.). In addition, studies on topic-controlled joke generation has been limited, which can hinder the usability of such models in real life. Topic control is important since jokes are often used in a specific context or setting, and choosing the right topic for the setting is crucial for a good use of a joke. This work introduces EDDIE, a keyword-based humor generation model that allows the user to have a fine-grained topic control on the jokes produced. Our system primarily focuses on producing Question and Answer (Q&A) style jokes with occasional one-liners. We have developed a humor generation pipeline that incorporates humor classifier, humor generator, and toxicity filter in order to produce responsible and humorous content. We evaluate EDDIE in a quantitative and qualitative manner. For the former evaluation, we assess the ability of various language models to learn and utilize mechanisms of humor given no linguistic knowledge. For the latter, we have designed three evaluation criteria : funniness, understandability, and complexity. The study has been conducted with both expert comedians and general volunteers from the general public in order to gain insights into machine humor generation and its application in real life.

Lunch Break
12:30-14:00 - **Auditorium** (Lunch is not served)

Session 2 - 14:00-15:00

Information Retrieval and Text Mining
14:00-14:55 (Liffey Hall 1)

14:00-14:15 (Liffey Hall 1)
SDR: Efficient Neural Re-ranking using Succinct Document Representation
Nachshon Cohen, Amit Portnoy, Besnik Fetahu and Amir Ingber
BERT based ranking models have achieved superior performance on various information retrieval tasks. However, the large number of parameters and complex self-attention operations come at a significant latency overhead. To remedy this, recent works propose late-interaction architectures, which allow pre-computation of intermediate document representations, thus reducing latency. Nonetheless, having solved the immediate latency issue, these methods now introduce storage costs and network fetching latency, which limit their adoption in real-life production systems.
In this work, we propose the Succinct Document Representation (SDR) scheme that computes *highly compressed* intermediate document representations, mitigating the storage/network issue. Our approach first reduces the dimension of token representations by encoding them using a novel autoencoder architecture that uses the document's textual content in both the encoding and decoding phases. After this token encoding step, we further reduce the size of the document representations using modern quantization techniques.
Evaluation on MSMARCO's passage re-reranking task show that compared to existing approaches using compressed document representations, our method is highly efficient, achieving 4x–11.6x higher compression rates for the same ranking quality. Similarly, on the TREC CAR dataset, we achieve 7.7x higher compression rate for the same ranking quality.

14:15-14:30 (Liffey Hall 1)
Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval
Luyu Gao and Jamie Callan
Recent research demonstrates the effectiveness of using fine-tuned language models (LM) for dense retrieval. However, dense retrievers are hard to train, typically requiring heavily engineered fine-tuning pipelines to realize their full potential. In this paper, we identify and address two underlying problems of dense retrievers: i) fragility to training data noise and ii) requiring large batches to robustly learn the embedding space. We use the recently proposed Condenser pre-training architecture, which learns to condense information into the dense vector through LM pre-training. On top of it, we propose coCondenser, which adds an unsupervised corpus-level contrastive loss to warm up the passage

embedding space. Experiments on MS-MARCO, Natural Question, and Trivia QA datasets show that coCondenser removes the need for heavy data engineering such as augmentation, synthesis, or filtering, and the need for large batch training. It shows comparable performance to RocketQA, a state-of-the-art, heavily engineered system, using simple small batch fine-tuning.

14:30-14:45 (Liffey Hall 1)
[TACL] **ABNIRML: Analyzing the Behavior of Neural IR Models**
Sean MacAvaney, Sergey Feldman, Nazli Goharian, Doug Downey and Arman Cohan

14:45-14:55 (Liffey Hall 1)
Augmenting Document Representations for Dense Retrieval with Interpolation and Perturbation
Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang and Jong C. Park
Dense retrieval models, which aim at retrieving the most relevant document for an input query on a dense representation space, have gained considerable attention for their remarkable success. Yet, dense models require a vast amount of labeled training data for notable performance, whereas it is often challenging to acquire query-document pairs annotated by humans. To tackle this problem, we propose a simple but effective Document Augmentation for dense Retrieval (DAR) framework, which augments the representations of documents with their interpolation and perturbation. We validate the performance of DAR on retrieval tasks with two benchmark datasets, showing that the proposed DAR significantly outperforms relevant baselines on the dense retrieval of both the labeled and unlabeled documents.

Phonology, Morphology and Word Segmentation

14:00-15:00 (Wicklow Hall 1)

14:00-14:15 (Wicklow Hall 1)
CaMEL: Case Marker Extraction without Labels
Leonie Weissweiler, Valentin Hofmann, Masoud Jalili Sabet and Hinrich Schuetze
We introduce **CaMEL** (**Ca**se **M**arker **E**xtraction without **L**abels), a novel and challenging task in computational morphology that is especially relevant for low-resource languages. We propose a first model for CaMEL that uses a massively multilingual corpus to extract case markers in 83 languages based only on a noun phrase chunker and an alignment system. To evaluate CaMEL, we automatically construct a silver standard from UniMorph. The case markers extracted by our model can be used to detect and visualise similarities and differences between the case systems of different languages as well as to annotate fine-grained deep cases in languages in which they are not overtly marked.

14:15-14:25 (Wicklow Hall 1)
(Un)solving Morphological Inflection: Lemma Overlap Artificially Inflates Models' Performance
Omer Goldman, David Guriel and Reut Tsarfaty
In the domain of Morphology, Inflection is a fundamental and important task that gained a lot of traction in recent years, mostly via SIGMORPHON's shared-tasks. With average accuracy above 0.9 over the scores of all languages, the task is considered mostly solved using relatively generic neural seq2seq models, even with little data provided. In this work, we propose to re-evaluate morphological inflection models by employing harder train-test splits that will challenge the generalization capacity of the models. In particular, as opposed to the naïve split-by-form, we propose a split-by-lemma method to challenge the performance on existing benchmarks. Our experiments with the three top-ranked systems on the SIGMORPHON's 2020 shared-task show that the lemma-split presents an average drop of 30 percentage points in macro-average for the 90 languages included. The effect is most significant for low-resourced languages with a drop as high as 95 points, but even high-resourced languages lose about 10 points on average. Our results clearly show that generalizing inflection to unseen lemmas is far from being solved, presenting a simple yet effective means to promote more sophisticated models.

14:25-14:35 (Wicklow Hall 1)
Morphological Reinflection with Multiple Arguments: An Extended Annotation schema and a Georgian Case Study
David Guriel, Omer Goldman and Reut Tsarfaty
In recent years, a flurry of morphological datasets had emerged, most notably UniMorph, aa multi-lingual repository of inflection tables. However, the flat structure of the current morphological annotation makes the treatment of some languages quirky, if not impossible, specifically in cases of polypersonal agreement. In this paper we propose a general solution for such cases and expand the UniMorph annotation schema to naturally address this phenomenon, in which verbs agree with multiple arguments using true affixes. We apply this extended schema to one such language, Georgian, and provide a human-verified, accurate and balanced morphological dataset for Georgian verbs. The dataset has 4 times more tables and 6 times more verb forms compared to the existing UniMorph dataset, covering all possible variants of argument marking, demonstrating the adequacy of our proposed scheme. Experiments on a reinflection task show that generalization is easy when the data is split at the form level, but extremely hard when splitting along lemma lines. Expanding the other languages in UniMorph according to this schema is expected to improve both the coverage, consistency and interpretability of this benchmark.

14:35-14:45 (Wicklow Hall 1)
WLASL-LEX: a Dataset for Recognising Phonological Properties in American Sign Language
Federico Tavella, Viktor Schlegel, Marta Romeo, Aphrodite Galata and Angelo Cangelosi
Signed Language Processing (SLP) concerns the automated processing of signed languages, the main means of communication of Deaf and hearing impaired individuals. SLP features many different tasks, ranging from sign recognition to translation and production of signed speech, but has been overlooked by the NLP community thus far. In this paper, we bring to attention the task of modelling the phonology of sign languages. We leverage existing resources to construct a large-scale dataset of American Sign Language signs annotated with six different phonological properties. We then conduct an extensive empirical study to investigate whether data-driven end-to-end and feature-based approaches can be optimised to automatically recognise these properties. We find that, despite the inherent challenges of the task, graph-based neural networks that operate over skeleton features extracted from raw videos are able to succeed at the task to a varying degree. Most importantly, we show that this performance pertains even on signs unobserved during training.

14:45-14:55 (Wicklow Hall 1)
An Embarrassingly Simple Method to Mitigate Undesirable Properties of Pretrained Language Model Tokenizers
Valentin Hofmann, Hinrich Schuetze and Janet B. Pierrehumbert
We introduce FLOTA (Few Longest Token Approximation), a simple yet effective method to improve the tokenization of pretrained language models (PLMs). FLOTA uses the vocabulary of a standard tokenizer but tries to preserve the morphological structure of words during tokenization. We evaluate FLOTA on morphological gold segmentations as well as a text classification task, using BERT, GPT-2, and XLNet as example PLMs. FLOTA leads to performance gains, makes inference more efficient, and enhances the robustness of PLMs with respect to whitespace noise.

Dialogue and Interactive Systems 1

14:00-15:00 (The Liffey B)

14:00-14:15 (The Liffey B)
DEAM: Dialogue Coherence Evaluation using AMR-based Semantic Manipulations
Sarik Ghazarian, Nuan Wen, Aram Galstyan and Nanyun Peng
Automatic evaluation metrics are essential for the rapid development of open-domain dialogue systems as they facilitate hyper-parameter tuning and comparison between models. Although recently proposed trainable conversation-level metrics have shown encouraging results, the quality of the metrics is strongly dependent on the quality of training data. Prior works mainly resort to heuristic text-level manipulations (e.g. utterances shuffling) to bootstrap incoherent conversations (negative examples) from coherent dialogues (positive examples). Such approaches are insufficient to appropriately reflect the incoherence that occurs in interactions between advanced dialogue models and humans. To tackle this problem, we propose DEAM, a Dialogue coherence Evaluation metric that relies on Abstract Meaning Representation (AMR) to apply semantic-level Manipulations for incoherent (negative) data generation. AMRs naturally facilitate the injection of various types of incoherence sources, such as coreference inconsistency, irrelevancy, contradictions, and decrease engagement, at the semantic level, thus resulting in more natural incoherent samples. Our experiments show that DEAM achieves higher correlations with human judgments compared to baseline methods on several dialog datasets by significant margins. We also show that DEAM can distinguish between coherent and incoherent dialogues generated by baseline manipulations, whereas those baseline models cannot detect incoherent examples generated by DEAM. Our results demonstrate the potential of AMR-based semantic manipulations for natural negative example generation.

14:15-14:30 (The Liffey B)
Achieving Reliable Human Assessment of Open-Domain Dialogue Systems
Tianbo Ji, Yvette Graham, Gareth J. F. Jones, Chenyang Lyu and Qun Liu
Evaluation of open-domain dialogue systems is highly challenging and development of better techniques is highlighted time and again as desperately needed. Despite substantial efforts to carry out reliable live evaluation of systems in recent competitions, annotations have been abandoned and reported as too unreliable to yield sensible results. This is a serious problem since automatic metrics are not known to provide a good indication of what may or may not be a high-quality conversation. Answering the distress call of competitions that have emphasized the urgent need for better evaluation techniques in dialogue, we present the successful development of human evaluation that is highly reliable while still remaining feasible and low cost. Self-replication experiments reveal almost perfectly repeatable results with a correlation of $r = 0.969$. Furthermore, due to the lack of appropriate methods of statistical significance testing, the likelihood of potential improvements to systems occurring due to chance is rarely taken into account in dialogue evaluation, and the evaluation we propose facilitates application of standard tests. Since we have developed a highly reliable evaluation method, new insights into system performance can be revealed. We therefore include a comparison of state-of-the-art models (i) with and without personas, to measure the contribution of personas to conversation quality, as well as (ii) prescribed versus freely chosen topics. Interestingly with respect to personas, results indicate that personas do not positively contribute to conversation quality as expected.

14:30-14:45 (The Liffey B)
Internet-Augmented Dialogue Generation
Mojtaba Komeili, Kurt Shuster and Jason E Weston
The largest store of continually updating knowledge on our planet can be accessed via internet search. In this work we study giving access to this information to conversational agents. Large language models, even though they store an impressive amount of knowledge within their weights, are known to hallucinate facts when generating dialogue (Shuster et al., 2021); moreover, those facts are frozen in time at the point of model training. In contrast, we propose an approach that learns to generate an internet search query based on the context, and then conditions on the search results to finally generate a response, a method that can employ up-to-the-minute relevant information. We train and evaluate such models on a newly collected dataset of human-human conversations whereby one of the speakers is given access to internet search during knowledgedriven discussions in order to ground their responses. We find that search-query based access of the internet in conversation provides superior performance compared to existing approaches that either use no augmentation or FAISS-based retrieval (Lewis et al., 2020b).

14:45-15:00 (The Liffey B)
Online Semantic Parsing for Latency Reduction in Task-Oriented Dialogue
Jiawei Zhou, Jason Eisner, Michael Newman, Emmanouil Antonios Platanios and Sam Thomson
Standard conversational semantic parsing maps a complete user utterance into an executable program, after which the program is executed to respond to the user. This could be slow when the program contains expensive function calls. We investigate the opportunity to reduce latency by predicting and executing function calls while the user is still speaking. We introduce the task of online semantic parsing for this purpose, with a formal latency reduction metric inspired by simultaneous machine translation. We propose a general framework with first a learned prefix-to-program prediction module, and then a simple yet effective thresholding heuristic for subprogram selection for early execution. Experiments on the SMCalFlow and TreeDST datasets show our approach achieves large latency reduction with good parsing quality, with a 30

Ethics in NLP

14:00-15:00 (The Liffey A)

14:00-14:15 (The Liffey A)
An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models
Nicholas Meade, Elinor Poole-Dayan and Siva Reddy
Recent work has shown pre-trained language models capture social biases from the large amounts of text they are trained on. This has attracted attention to developing techniques that mitigate such biases. In this work, we perform an empirical survey of five recently proposed bias mitigation techniques: Counterfactual Data Augmentation (CDA), Dropout, Iterative Nullspace Projection, Self-Debias, and SentenceDebias. We quantify the effectiveness of each technique using three intrinsic bias benchmarks while also measuring the impact of these techniques on a model's language modeling ability, as well as its performance on downstream NLU tasks. We experimentally find that: (1) Self-Debias is the strongest debiasing technique, obtaining improved scores on all bias benchmarks; (2) Current debiasing techniques perform less consistently when mitigating non-gender biases; And (3) improvements on bias benchmarks such as StereoSet and CrowS-Pairs by using debiasing strategies are often accompanied by a decrease in language modeling ability, making it difficult to determine whether the bias mitigation was effective.

14:15-14:30 (The Liffey A)
Ethics Sheets for AI Tasks
Saif M. Mohammad
Several high-profile events, such as the mass testing of emotion recognition systems on vulnerable sub-populations and using question answering systems to make moral judgments, have highlighted how technology will often lead to more adverse outcomes for those that are already marginalized. At issue here are not just individual systems and datasets, but also the AI tasks themselves. In this position paper, I make a case for thinking about ethical considerations not just at the level of individual models and datasets, but also at the level of AI tasks. I will present a new form of such an effort, Ethics Sheets for AI Tasks, dedicated to fleshing out the assumptions and ethical considerations hidden in how a task is commonly framed and in the choices we make regarding the data, method, and evaluation. I will also present a template for ethics sheets with 50 ethical considerations, using the task of emotion recognition as a running example. Ethics sheets are a mechanism to engage with and document ethical considerations before building datasets and systems. Similar to survey articles, a small number of carefully created ethics sheets can serve numerous researchers and developers.

14:30-14:45 (The Liffey A)
The Dangers of Underclaiming: Reasons for Caution When Reporting How NLP Systems Fail
Samuel R. Bowman
Researchers in NLP often frame and discuss research results in ways that serve to deemphasize the field's successes, often in response to the field's widespread hype. Though well-meaning, this has yielded many misleading or false claims about the limits of our best technology. This is a problem, and it may be more serious than it looks: It harms our credibility in ways that can make it harder to mitigate present-day harms, like those involving biased systems for content moderation or resume screening. It also limits our ability to prepare for the potentially enormous impacts of more distant future advances. This paper urges researchers to be careful about these claims and suggests some research directions and communication strategies that will make it easier to avoid or rebut them.

14:45-15:00 (The Liffey A)
French CrowS-Pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English
Aurélie Névéol, Yoann Dupont, Julien Bezançon and Karën Fort
Warning: This paper contains explicit statements of offensive stereotypes which may be upsetting. Much work on biases in natural language processing has addressed biases linked to the social and cultural experience of English speaking individuals in the United States. We seek to widen the scope of bias studies by creating material to measure social bias in language models (LMs) against specific demographic groups in France. We build on the US-centered CrowS-pairs dataset to create a multilingual stereotypes dataset that allows for comparability across languages while also characterizing biases that are specific to each country and language. We introduce 1,679 sentence pairs in French that cover stereotypes in ten types of bias like gender and age. 1,467 sentence pairs are translated from CrowS-pairs and 212 are newly crowdsourced. The sentence pairs contrast stereotypes concerning underadvantaged groups with the same sentence concerning advantaged groups. We find that four widely used language models (three French, one multilingual) favor sentences that express stereotypes in most bias categories. We report on the translation process from English into French, which led to a characterization of stereotypes in CrowS-pairs including the identification of US-centric cultural traits. We offer guidelines to further extend the dataset to other languages and cultural environments.

Special Theme 1

14:00-15:00 (Liffey Hall 2)

14:00-14:15 (Liffey Hall 2)
Cree Corpus: A Collection of nêhiyawêwin Resources
Daniela Teodorescu, Josie Matalski, Delaney Alexa Lothian, Denilson Barbosa and Carrie Demmans Epp
Plains Cree (nêhiyawêwin) is an Indigenous language that is spoken in Canada and the USA. It is the most widely spoken dialect of Cree and a morphologically complex language that is polysynthetic, highly inflective, and agglutinative. It is an extremely low resource language, with no existing corpus that is both available and prepared for supporting the development of language technologies. To support nêhiyawêwin revitalization and preservation, we developed a corpus covering diverse genres, time periods, and texts for a variety of intended audiences. The data has been verified and cleaned; it is ready for use in developing language technologies for nêhiyawêwin. The corpus includes the corresponding English phrases or audio files where available. We demonstrate the utility of the corpus through its community use and its use to build language technologies that can provide the types of support that community members have expressed are desirable. The corpus is available for public use.

14:15-14:30 (Liffey Hall 2)
Phone-ing it in: Towards Flexible Multi-Modal Language Model Training by Phonetic Representations of Data

Colin Leong and Daniel Lee Whitenack
Multi-modal techniques offer significant untapped potential to unlock improved NLP technology for local languages. However, many advances in language model pre-training are focused on text, a fact that only increases systematic inequalities in the performance of NLP tasks across the world's languages. In this work, we propose a multi-modal approach to train language models using whatever text and/or audio data might be available in a language. Initial experiments using Swahili and Kinyarwanda data suggest the viability of the approach for downstream Named Entity Recognition (NER) tasks, with models pre-trained on phone data showing an improvement of up to 6% F1-score above models that are trained from scratch. Preprocessing and training code will be uploaded to https://github.com/sil-ai/phone-it-in.

14:30-14:45 (Liffey Hall 2)
From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology
Mark Dingemanse and Andreas Liesenfeld
Informal social interaction is the primordial home of human language. Linguistically diverse conversational corpora are an important and largely untapped resource for computational linguistics and language technology. Through the efforts of a worldwide language documentation movement, such corpora are increasingly becoming available. We show how interactional data from 63 languages (26 families) harbours insights about turn-taking, timing, sequential structure and social action, with implications for language technology, natural language understanding, and the design of conversational interfaces. Harnessing linguistically diverse conversational corpora will provide the empirical foundations for flexible, localizable, humane language technologies of the future.

14:45-15:00 (Liffey Hall 2)
How can NLP Help Revitalize Endangered Languages? A Case Study and Roadmap for the Cherokee Language
Shiyue Zhang, Ben Frey and Mohit Bansal
More than 43

Discourse and Pragmatics

14:00-15:00 (Wicklow Hall 2a)

14:00-14:15 (Wicklow Hall 2a)
How Do We Answer Complex Questions: Discourse Structure of Long-form Answers
Fangyuan Xu, Junyi Jessy Li and Eunsol Choi
Long-form answers, consisting of multiple sentences, can provide nuanced and comprehensive answers to a broader set of questions. To better understand this complex and understudied task, we study the functional structure of long-form answers collected from three datasets, ELI5, WebGPT and Natural Questions. Our main goal is to understand how humans organize information to craft complex answers. We develop an ontology of six sentence-level functional roles for long-form answers, and annotate 3.9k sentences in 640 answer paragraphs. Different answer collection methods manifest in different discourse structures. We further analyze model-generated answers – finding that annotators agree less with each other when annotating model-generated answers compared to annotating human-written answers. Our annotated data enables training a strong classifier that can be used for automatic analysis. We hope our work can inspire future research on discourse-level modeling and evaluation of long-form QA systems.

14:15-14:30 (Wicklow Hall 2a)
Entity-based Neural Local Coherence Modeling
Sungho Jeon and Michael Strube
In this paper, we propose an entity-based neural local coherence model which is linguistically more sound than previously proposed neural coherence models. Recent neural coherence models encode the input document using large-scale pretrained language models. Hence their basis for computing local coherence are words and even sub-words. The analysis of their output shows that these models frequently compute coherence on the basis of connections between (sub-)words which, from a linguistic perspective, should not play a role. Still, these models achieve state-of-the-art performance in several end applications. In contrast to these models, we compute coherence on the basis of entities by constraining the input to noun phrases and proper names. This provides us with an explicit representation of the most important items in sentences leading to the notion of focus. This brings our model linguistically in line with pre-neural models of computing coherence. It also gives us better insight into the behaviour of the model thus leading to better explainability. Our approach is also in accord with a recent study (O'Connor and Andreas, 2021), which shows that most usable information is captured by nouns and verbs in transformer-based language models. We evaluate our model on three downstream tasks showing that it is not only linguistically more sound than previous models but also that it outperforms them in end applications.

14:30-14:45 (Wicklow Hall 2a)
Rethinking Self-Supervision Objectives for Generalizable Coherence Modeling
Prathyusha Jwalapuram, Shafiq Joty and Xiang Lin
Given the claims of improved text generation quality across various pre-trained neural models, we consider the coherence evaluation of machine generated text to be one of the principal applications of coherence models that needs to be investigated. Prior work in neural coherence modeling has primarily focused on devising new architectures for solving the permuted document task. We instead use a basic model architecture and show significant improvements over state of the art within the same training regime. We then design a harder self-supervision objective by increasing the ratio of negative samples within a contrastive learning setup, and enhance the model further through automatic hard negative mining coupled with a large global negative queue encoded by a momentum encoder. We show empirically that increasing the density of negative samples improves the basic model, and using a global negative queue further improves and stabilizes the model while training with hard negative samples. We evaluate the coherence model on task-independent test sets that resemble real-world applications and show significant improvements in coherence evaluations of downstream tasks.

14:45-15:00 (Wicklow Hall 2a)
[TACL] **Out-of-Domain Discourse Dependency Parsing via Bootstrapping: An Empirical Analysis on Its Effectiveness and Limitation**
Noriki Nishida and Yuji Matsumoto

Sentiment Analysis, Stylistic Analysis, and Argument Mining 1

14:00-15:00 (Wicklow Hall 2b)

14:00-14:15 (Wicklow Hall 2b)
The Moral Debater: A Study on the Computational Generation of Morally Framed Arguments
Milad Alshomary, Roxanne El Baff, Timon Gurcke and Henning Wachsmuth
An audience's prior beliefs and morals are strong indicators of how likely they will be affected by a given argument. Utilizing such knowledge can help focus on shared values to bring disagreeing parties towards agreement. In argumentation technology, however, this is barely exploited so far. This paper studies the feasibility of automatically generating morally framed arguments as well as their effect on different audiences. Following the moral foundation theory, we propose a system that effectively generates arguments focusing on different morals. In an in-depth user study, we ask liberals and conservatives to evaluate the impact of these arguments. Our results suggest that, particularly when prior beliefs are challenged, an audience becomes more affected by morally framed arguments.

14:15-14:30 (Wicklow Hall 2b)
Identifying the Human Values behind Arguments
Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth and Benno Stein
This paper studies the (often implicit) human values behind natural language arguments, such as to have freedom of thought or to be broad-minded. Values are commonly accepted answers to why some option is desirable in the ethical sense and are thus essential both in real-world argumentation and theoretical argumentation frameworks. However, their large variety has been a major obstacle to modeling them in argument mining. To overcome this obstacle, we contribute an operationalization of human values, namely a multi-level taxonomy with 54 values that is in line with psychological research. Moreover, we provide a dataset of 5270 arguments from four geographical cultures, manually annotated for human values. First experiments with the automatic classification of human values are promising, with F_1-scores up to 0.81 and 0.25 on average.

14:30-14:45 (Wicklow Hall 2b)
Can Unsupervised Knowledge Transfer from Social Discussions Help Argument Mining?
Subhabrata Dutta, Jeevesh Juneja, Dipankar Das and Tanmoy Chakraborty
Identifying argument components from unstructured texts and predicting the relationships expressed among them are two primary steps of argument mining. The intrinsic complexity of these tasks demands powerful learning models. While pretrained Transformer-based Language Models (LM) have been shown to provide state-of-the-art results over different NLP tasks, the scarcity of manually annotated data and the highly domain-dependent nature of argumentation restrict the capabilities of such models. In this work, we propose a novel transfer learning strategy to overcome these challenges. We utilize argumentation-rich social discussions from the *ChangeMyView* subreddit as a source of unsupervised, argumentative discourse-aware knowledge by finetuning pretrained LMs on a selectively masked language modeling task. Furthermore, we introduce a novel prompt-based strategy for inter-component relation prediction that compliments our proposed finetuning method while leveraging on the discourse context. Exhaustive experiments show the generalization capability of our method on these two tasks over within-domain as well as out-of-domain datasets, outperforming several existing and employed strong baselines.

14:45-15:00 (Wicklow Hall 2b)
Fair and Argumentative Language Modeling for Computational Argumentation
Carolin Holtermann, Anne Lauscher and Simone Paolo Ponzetto
Although much work in NLP has focused on measuring and mitigating stereotypical bias in semantic spaces, research addressing bias in computational argumentation is still in its infancy. In this paper, we address this research gap and conduct a thorough investigation of bias in argumentative language models. To this end, we introduce ABBA, a novel resource for bias measurement specifically tailored to argumentation. We employ our resource to assess the effect of argumentative fine-tuning and debiasing on the intrinsic bias found in transformer-based language models using a lightweight adapter-based approach that is more sustainable and parameter-efficient than full fine-tuning. Finally, we analyze the potential impact of language model debiasing on the performance in argument quality prediction, a downstream task of computational argumentation. Our results show that we are able to successfully and sustainably remove bias in general and argumentative language models while preserving (and sometimes improving) model performance in downstream tasks. We make all experimental code and data available at https://github.com/umanlp/FairArgumentativeLM.

Poster Session 2: Language Grounding, Speech and Multimodality

14:00-15:00 (Forum)

14:00-15:00 (Forum)
#1 CARETS: A Consistency And Robustness Evaluative Test Suite for VQA
Carlos E Jimenez, Olga Russakovsky and Karthik R Narasimhan
We introduce CARETS, a systematic test suite to measure consistency and robustness of modern VQA models through a series of six fine-grained capability tests. In contrast to existing VQA test sets, CARETS features balanced question generation to create pairs of instances to test models, with each pair focusing on a specific capability such as rephrasing, logical symmetry or image obfuscation. We evaluate six modern VQA systems on CARETS and identify several actionable weaknesses in model comprehension, especially with concepts such as negation, disjunction, or hypernym invariance. Interestingly, even the most sophisticated models are sensitive to aspects such as swapping the order of terms in a conjunction or varying the number of answer choices mentioned in the question. We release CARETS to be used as an extensible tool for evaluating multi-modal model robustness.

14:00-15:00 (Forum)
#2 Understanding Multimodal Procedural Knowledge by Sequencing Multimodal Instructional Manuals
Te-Lin Wu, Alex Spangher, Pegah Alipoormolabashi, Marjorie Freedman, Ralph M. Weischedel and Nanyun Peng
The ability to sequence unordered events is evidence of comprehension and reasoning about real world tasks/procedures. It is essential for applications such as task planning and multi-source instruction summarization. It often requires thorough understanding of temporal common sense and multimodal information, since these procedures are often conveyed by a combination of texts and images. While humans are capable

of reasoning about and sequencing unordered procedural instructions, the extent to which the current machine learning methods possess such capability is still an open question. In this work, we benchmark models' capability of reasoning over and sequencing unordered multimodal instructions by curating datasets from online instructional manuals and collecting comprehensive human annotations. We find current state-of-the-art models not only perform significantly worse than humans but also seem incapable of efficiently utilizing multimodal information. To improve machines' performance on multimodal event sequencing, we propose sequence-aware pretraining techniques exploiting the sequential alignment properties of both texts and images, resulting in > 5

14:00-15:00 (Forum)
#3 Skill Induction and Planning with Latent Language
Pratyusha Sharma, Antonio Torralba and Jacob Andreas
We present a framework for learning hierarchical policies from demonstrations, using sparse natural language annotations to guide the discovery of reusable skills for autonomous decision-making. We formulate a generative model of action sequences in which goals generate sequences of high-level subtask descriptions, and these descriptions generate sequences of low-level actions. We describe how to train this model using primarily unannotated demonstrations by parsing demonstrations into sequences of named high-level sub-tasks, using only a small number of seed annotations to ground language in action. In trained models, natural language commands index a combinatorial library of skills; agents can use these skills to plan by generating high-level instruction sequences tailored to novel goals. We evaluate this approach in the ALFRED household simulation environment, providing natural language annotations for only 10

14:00-15:00 (Forum)
#4 Multimodal fusion via cortical network inspired losses
Shiv Shankar
Information integration from different modalities is an active area of research. Human beings and, in general, biological neural systems are quite adept at using a multitude of signals from different sensory perceptive fields to interact with the environment and each other. Recent work in deep fusion models via neural networks has led to substantial improvements over unimodal approaches in areas like speech recognition, emotion recognition and analysis, captioning and image description. However, such research has mostly focused on architectural changes allowing for fusion of different modalities while keeping the model complexity manageable. Inspired by neuroscientific ideas about multisensory integration and processing, we investigate the effect of introducing neural dependencies in the loss functions. Experiments on multimodal sentiment analysis tasks with different models show that our approach provides a consistent performance boost.

14:00-15:00 (Forum)
#5 Leveraging Visual Knowledge in Language Tasks: An Empirical Study on Intermediate Pre-training for Cross-Modal Knowledge Transfer
Woojeong Jin, Dong-Ho Lee, Chenguang Zhu, Jay Pujara and Xiang Ren
Pre-trained language models are still far from human performance in tasks that need understanding of properties (e.g. appearance, measurable quantity) and affordances of everyday objects in the real world since the text lacks such information due to reporting bias. In this work, we study whether integrating visual knowledge into a language model can fill the gap. We investigate two types of knowledge transfer: (1) *text knowledge transfer using image captions that may contain enriched visual knowledge and (2) cross-modal knowledge transfer using both images and captions with vision-language training objectives. On 5 downstream tasks that may need visual knowledge to solve the problem, we perform extensive empirical comparisons over the presented objectives. Our experiments show that visual knowledge transfer can improve performance in both low-resource and fully supervised settings.*

14:00-15:00 (Forum)
[TACL] **#6 Retrieve Fast, Rerank Smart: Cooperative and Joint Approaches for Improved Cross-Modal Retrieval**
Gregor Geigle, Jonas Pfeiffer, Nils Reimers, Ivan Vulić and Iryna Gurevych

14:00-15:00 (Forum)
[TACL] **#7 Word Representation Learning in Multimodal Pre-Trained Transformers: An Intrinsic Evaluation**
Sandro Pezzelle, Ece Takmaz and Raquel Fernández

14:00-15:00 (Forum)
#8 Analyzing Generalization of Vision and Language Navigation to Unseen Outdoor Areas
Raphael Schumann and Stefan Riezler
Vision and language navigation (VLN) is a challenging visually-grounded language understanding task. Given a natural language navigation instruction, a visual agent interacts with a graph-based environment equipped with panorama images and tries to follow the described route. Most prior work has been conducted in indoor scenarios where best results were obtained for navigation on routes that are similar to the training routes, with sharp drops in performance when testing on unseen environments. We focus on VLN in outdoor scenarios and find that in contrast to indoor VLN, most of the gain in outdoor VLN on unseen data is due to features like junction type embedding or heading delta that are specific to the respective environment graph, while image information plays a very minor role in generalizing VLN to unseen outdoor areas. These findings show a bias to specifics of graph representations of urban environments, demanding that VLN tasks grow in scale and diversity of geographical environments.

14:00-15:00 (Forum)
#9 Understanding Game-Playing Agents with Natural Language Annotations
Nicholas Tomlin, Andre Wang He and Dan Klein
We present a new dataset containing 10K human-annotated games of Go and show how these natural language annotations can be used as a tool for model interpretability. Given a board state and its associated comment, our approach uses linear probing to predict mentions of domain-specific terms (e.g., ko, atari) from the intermediate state representations of game-playing agents like AlphaGo Zero. We find these game concepts are nontrivially encoded in two distinct policy networks, one trained via imitation learning and another trained via reinforcement learning. Furthermore, mentions of domain-specific terms are most easily predicted from the later layers of both models, suggesting that these policy networks encode high-level abstractions similar to those used in the natural language annotations.

14:00-15:00 (Forum)
#10 VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto and Albert Gatt
We propose VALSE (Vision And Language Structured Evaluation), a novel benchmark designed for testing general-purpose pretrained vision and language (V&L) models for their visio-linguistic grounding capabilities on specific linguistic phenomena. VALSE offers a suite of six tests covering various linguistic constructs. Solving these requires models to ground linguistic phenomena in the visual modality, allowing more fine-grained evaluations than hitherto possible. We build VALSE using methods that support the construction of valid foils, and report results from evaluating five widely-used V&L models. Our experiments suggest that current models have considerable difficulty addressing most phenomena. Hence, we expect VALSE to serve as an important benchmark to measure future progress of pretrained V&L models from a linguistic perspective, complementing the canonical task-centred V&L evaluations.

14:00-15:00 (Forum)
#11 Voxel-informed Language Grounding
Rodolfo Corona, Shizhan Zhu, Dan Klein and Trevor Darrell
Natural language applied to natural 2D images describes a fundamentally 3D world. We present the Voxel-informed Language Grounder (VLG), a language grounding model that leverages 3D geometric information in the form of voxel maps derived from the visual input using a volumetric reconstruction model. We show that VLG significantly improves grounding accuracy on SNARE, an object reference game task. At the time of writing, VLG holds the top place on the SNARE leaderboard, achieving SOTA results with a 2.0

14:00-15:00 (Forum)
#12 Image Retrieval from Contextual Descriptions
Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti and Siva Reddy
The ability to integrate context, including perceptual and temporal cues, plays a pivotal role in grounding the meaning of a linguistic utterance. In order to measure to what extent current vision-and-language models master this ability, we devise a new multimodal challenge, Image Retrieval from Contextual Descriptions (ImageCoDe). In particular, models are tasked with retrieving the correct image from a set of 10 minimally contrastive candidates based on a contextual description. As such, each description contains only the details that help distinguish between images. Because of this, descriptions tend to be complex in terms of syntax and discourse and require drawing pragmatic inferences. Images are sourced from both static pictures and video frames. We benchmark several state-of-the-art models, including both cross-encoders such as ViLBERT and bi-encoders such as CLIP, on ImageCoDe. Our results reveal that these models dramatically lag behind human performance: the best variant achieves an accuracy of 20.9 on video frames and 59.4 on static pictures, compared with 90.8 in humans. Furthermore, we experiment with new model variants that are better equipped to incorporate visual and temporal context into their representations, which achieve modest gains. Our hope is that ImageCoDE will foster progress in grounded language understanding by encouraging models to focus on fine-grained visual differences.

14:00-15:00 (Forum)
#13 XDBERT: Distilling Visual Information to BERT from Cross-Modal Systems to Improve Language Understanding
Chan-Jan Hsu, Hung-yi Lee and Yu Tsao
Transformer-based models are widely used in natural language understanding (NLU) tasks, and multimodal transformers have been effective in visual-language tasks. This study explores distilling visual information from pretrained multimodal transformers to pretrained language encoders. Our framework is inspired by cross-modal encoders' success in visual-language tasks while we alter the learning objective to cater to the language-heavy characteristics of NLU. After training with a small number of extra adapting steps and finetuned, the proposed XDBERT (cross-modal distilled BERT) outperforms pretrained-BERT in general language understanding evaluation (GLUE), situations with adversarial generations (SWAG) benchmarks, and readability benchmarks. We analyze the performance of XDBERT on GLUE to show that the improvement is likely visually grounded.

14:00-15:00 (Forum)
#14 ReCLIP: A Strong Zero-Shot Baseline for Referring Expression Comprehension
Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh and Anna Rohrbach
Training a referring expression comprehension (ReC) model for a new visual domain requires collecting referring expressions, and potentially corresponding bounding boxes, for images in the domain. While large-scale pre-trained models are useful for image classification across domains, it remains unclear if they can be applied in a zero-shot manner to more complex tasks like ReC. We present ReCLIP, a simple but strong *zero-shot* baseline that repurposes CLIP, a state-of-the-art large-scale model, for ReC. Motivated by the close connection between ReC and CLIP's contrastive pre-training objective, the first component of ReCLIP is a region-scoring method that isolates object proposals via cropping and blurring, and passes them to CLIP. However, through controlled experiments on a synthetic dataset, we find that CLIP is largely incapable of performing spatial reasoning off-the-shelf. We reduce the gap between zero-shot baselines from prior work and supervised models by as much as 29% on RefCOCOg, and on RefGTA (video game imagery), ReCLIP's relative improvement over supervised ReC models trained on real images is 8%.

14:00-15:00 (Forum)
#15 A Good Prompt Is Worth Millions of Parameters: Low-resource Prompt-based Learning for Vision-Language Models
Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen and Xiang Ren
Large pre-trained vision-language (VL) models can learn a new task with a handful of examples and generalize to a new task without fine-tuning. However, these VL models are hard to deploy for real-world applications due to their impractically huge sizes and slow inference speed. To solve this limitation, we study prompt-based low-resource learning of VL tasks with our proposed method, FewVLM, relatively smaller than recent few-shot learners. For FewVLM, we pre-train a sequence-to-sequence transformer model with prefix language modeling (PrefixLM) and masked language modeling (MaskedLM). Furthermore, we analyze the effect of diverse prompts for few-shot tasks. Experimental results on VQA show that FewVLM with prompt-based learning outperforms Frozen which is 31x larger than FewVLM by 18.2In our analysis, we observe that (1) prompts significantly affect zero-shot performance but marginally affect few-shot performance, (2) models with noisy prompts learn as quickly as hand-crafted prompts given larger training data, and (3) MaskedLM helps VQA tasks while PrefixLM boosts captioning performance. Our code is publicly available at `https://github.com/woojeongjin/FewVLM`

14:00-15:00 (Forum)
#16 FIBER: Fill-in-the-Blanks as a Challenging Video Understanding Evaluation Framework
Santiago Castro, Ruoyao Wang, Pingxuan Huang, Ian Stewart, Oana Ignat, Nan Liu, Jonathan C. Stroud and Rada Mihalcea
We propose fill-in-the-blanks as a video understanding evaluation framework and introduce FIBER – a novel dataset consisting of 28,000 videos and descriptions in support of this evaluation framework. The fill-in-the-blanks setting tests a model's understanding of a video by

requiring it to predict a masked noun phrase in the caption of the video, given the video and the surrounding text. The FIBER benchmark does not share the weaknesses of the current state-of-the-art language-informed video understanding tasks, namely: (1) video question answering using multiple-choice questions, where models perform relatively well because they exploit linguistic biases in the task formulation, thus making our framework challenging for the current state-of-the-art systems to solve; and (2) video captioning, which relies on an open-ended evaluation framework that is often inaccurate because system answers may be perceived as incorrect if they differ in form from the ground truth. The FIBER dataset and our code are available at https://lit.eecs.umich.edu/fiber/.

14:00-15:00 (Forum)
#17 Inferring Rewards from Language in Context
Jessy Lin, Daniel Fried, Dan Klein and Anca Dragan
In classic instruction following, language like "I'd like the JetBlue flight" maps to actions (e.g., selecting that flight). However, language also conveys information about a user's underlying reward function (e.g., a general preference for JetBlue), which can allow a model to carry out desirable actions in new contexts. We present a model that infers rewards from language pragmatically: reasoning about how speakers choose utterances not only to elicit desired actions, but also to reveal information about their preferences. On a new interactive flight–booking task with natural language, our model more accurately infers rewards and predicts optimal actions in unseen environments, in comparison to past work that first maps language to actions (instruction following) and then maps actions to rewards (inverse reinforcement learning).

14:00-15:00 (Forum)
#18 Language-Agnostic Meta-Learning for Low-Resource Text-to-Speech with Articulatory Features
Florian Lux and Thang Vu
While neural text-to-speech systems perform remarkably well in high-resource scenarios, they cannot be applied to the majority of the over 6,000 spoken languages in the world due to a lack of appropriate training data. In this work, we use embeddings derived from articulatory vectors rather than embeddings derived from phoneme identities to learn phoneme representations that hold across languages. In conjunction with language agnostic meta learning, this enables us to fine-tune a high-quality text-to-speech model on just 30 minutes of data in a previously unseen language spoken by a previously unseen speaker.

14:00-15:00 (Forum)
#19 Cross-Modal Discrete Representation Learning
Alexander H. Liu, SouYoung Jin, Cheng-I Lai, Andrew Rouditchenko, Aude Oliva and James R. Glass
In contrast to recent advances focusing on high-level representation learning across modalities, in this work we present a self-supervised learning framework that is able to learn a representation that captures finer levels of granularity across different modalities such as concepts or events represented by visual objects or spoken words. Our framework relies on a discretized embedding space created via vector quantization that is shared across different modalities. Beyond the shared embedding space, we propose a Cross-Modal Code Matching objective that forces the representations from different views (modalities) to have a similar distribution over the discrete embedding space such that cross-modal objects/actions localization can be performed without direct supervision. We show that the proposed discretized multi-modal fine-grained representation (e.g., pixel/word/frame) can complement high-level summary representations (e.g., video/sentence/waveform) for improved performance on cross-modal retrieval tasks. We also observe that the discretized representation uses individual clusters to represent the same semantic concept across modalities.

14:00-15:00 (Forum)
#20 Sample, Translate, Recombine: Leveraging Audio Alignments for Data Augmentation in End-to-end Speech Translation
Tsz Kin Lam, Shigehiko Schamoni and Stefan Riezler
End-to-end speech translation relies on data that pair source-language speech inputs with corresponding translations into a target language. Such data are notoriously scarce, making synthetic data augmentation by back-translation or knowledge distillation a necessary ingredient of end-to-end training. In this paper, we present a novel approach to data augmentation that leverages audio alignments, linguistic properties, and translation. First, we augment a transcription by sampling from a suffix memory that stores text and audio data. Second, we translate the augmented transcript. Finally, we recombine concatenated audio segments and the generated translation. Our method delivers consistent improvements of up to 0.9 and 1.1 BLEU points on top of augmentation with knowledge distillation on five language pairs on CoVoST 2 and on two language pairs on Europarl-ST, respectively.

14:00-15:00 (Forum)
#21 SUPERB-SG: Enhanced Speech processing Universal PERformance Benchmark for Semantic and Generative Capabilities
Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-wen Yang, Shuyan Dong, Andy T. Liu, Cheng-I Lai, Jiatong Shi, Xuankai Chang, Phil Hall, Hsuan-Jui Chen, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed and Hung-yi Lee
Transfer learning has proven to be crucial in advancing the state of speech and natural language processing research in recent years. In speech, a model pre-trained by self-supervised learning transfers remarkably well on multiple tasks. However, the lack of a consistent evaluation methodology is limiting towards a holistic understanding of the efficacy of such models. SUPERB was a step towards introducing a common benchmark to evaluate pre-trained models across various speech tasks. In this paper, we introduce SUPERB-SG, a new benchmark focusing on evaluating the semantic and generative capabilities of pre-trained models by increasing task diversity and difficulty over SUPERB. We use a lightweight methodology to test the robustness of representations learned by pre-trained models under shifts in data domain and quality across different types of tasks. It entails freezing pre-trained model parameters, only using simple task-specific trainable heads. The goal is to be inclusive of all researchers, and encourage efficient use of computational resources. We also show that the task diversity of SUPERB-SG coupled with limited task supervision is an effective recipe for evaluating the generalizability of model representation.

14:00-15:00 (Forum)
#22 Modeling Intensification for Sign Language Generation: A Computational Approach
Mert Inan, Yang Zhong, Sabit Hassan, Lorna Quandt and Malihe Alikhani
End-to-end sign language generation models do not accurately represent the prosody in sign language. A lack of temporal and spatial variations leads to poor-quality generated presentations that confuse human interpreters. In this paper, we aim to improve the prosody in generated sign languages by modeling intensification in a data-driven manner. We present different strategies grounded in linguistics of sign language that inform how intensity modifiers can be represented in gloss annotations. To employ our strategies, we first annotate a subset of the benchmark PHOENIX-14T, a German Sign Language dataset, with different levels of intensification. We then use a supervised intensity tagger to extend the annotated dataset and obtain labels for the remaining portion of it. This enhanced dataset is then used to train state-of-the-art

transformer models for sign language generation. We find that our efforts in intensification modeling yield better results when evaluated with automatic metrics. Human evaluation also indicates a higher preference of the videos generated using our model.

14:00-15:00 (Forum)
#23 Comprehensive Multi-Modal Interactions for Referring Image Segmentation
Kanishk Jain and Vineet Gandhi
We investigate Referring Image Segmentation (RIS), which outputs a segmentation map corresponding to the natural language description. Addressing RIS efficiently requires considering the interactions happening across visual and linguistic modalities and the interactions within each modality. Existing methods are limited because they either compute different forms of interactions sequentially (leading to error propagation) or ignore intra-modal interactions. We address this limitation by performing all three interactions simultaneously through a Synchronous Multi-Modal Fusion Module (SFM). Moreover, to produce refined segmentation masks, we propose a novel Hierarchical Cross-Modal Aggregation Module (HCAM), where linguistic features facilitate the exchange of contextual information across the visual hierarchy. We present thorough ablation studies and validate our approach's performance on four benchmark datasets, showing considerable performance gains over the existing state-of-the-art (SOTA) methods.

14:00-15:00 (Forum)
#24 Attention as Grounding: Exploring Textual and Cross-Modal Attention on Entities and Relations in Language-and-Vision Transformer
Nikolai Ilinykh and Simon Dobnik
We explore how a multi-modal transformer trained for generation of longer image descriptions learns syntactic and semantic representations about entities and relations grounded in objects at the level of masked self-attention (text generation) and cross-modal attention (information fusion). We observe that cross-attention learns the visual grounding of noun phrases into objects and high-level semantic information about spatial relations, while text-to-text attention captures low-level syntactic knowledge between words. This concludes that language models in a multi-modal task learn different semantic information about objects and relations cross-modally and uni-modally (text-only). Our code is available here: https://github.com/GU-CLASP/attention-as-grounding.

14:00-15:00 (Forum)
#25 Interpreting Character Embeddings With Perceptual Representations: The Case of Shape, Sound, and Color
Sidsel Boldsen, Manex Agirrezabal and Nora Hollenstein
Character-level information is included in many NLP models, but evaluating the information encoded in character representations is an open issue. We leverage perceptual representations in the form of shape, sound, and color embeddings and perform a representational similarity analysis to evaluate their correlation with textual representations in five languages. This cross-lingual analysis shows that textual character representations correlate strongly with sound representations for languages using an alphabetic script, while shape correlates with featural scripts. We further develop a set of probing classifiers to intrinsically evaluate what phonological information is encoded in character embeddings. Our results suggest that information on features such as voicing are embedded in both LSTM and transformer-based representations.

14:00-15:00 (Forum)
#26 Zero-shot Learning for Grapheme to Phoneme Conversion with Language Ensemble
Xinjian Li, Florian Metze, David R Mortensen, Shinji Watanabe and Alan Black
Grapheme-to-Phoneme (G2P) has many applications in NLP and speech fields. Most existing work focuses heavily on languages with abundant training datasets, which limits the scope of target languages to less than 100 languages. This work attempts to apply zero-shot learning to approximate G2P models for all low-resource and endangered languages in Glottolog (about 8k languages). For any unseen target language, we first build the phylogenetic tree (i.e. language family tree) to identify top-k nearest languages for which we have training sets. Then we run models of those languages to obtain a hypothesis set, which we combine into a confusion network to propose a most likely hypothesis as an approximation to the target language. We test our approach on over 600 unseen languages and demonstrate it significantly outperforms baselines.

14:00-15:00 (Forum)
[DEMO] **ViLMedic: a framework for research at the intersection of vision and language in medical AI**
Jean-benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari and Curtis Langlotz
There is a growing need to model interactions between data modalities (e.g., vision, language) — both to improve AI predictions on existing tasks and to enable new applications. In the recent field of multimodal medical AI, integrating multiple modalities has gained widespread popularity as multimodal models have proven to improve performance, robustness, require less training samples and add complementary information. To improve technical reproducibility and transparency for multimodal medical tasks as well as speed up progress across medical AI, we present ViLMedic, a Vision-and-Language medical library. As of 2022, the library contains a dozen reference implementations replicating the state-of-the-art results for problems that range from medical visual question answering and radiology report generation to multimodal representation learning on widely adopted medical datasets. In addition, ViLMedic hosts a model-zoo with more than twenty pretrained models for the above tasks designed to be extensible by researchers but also simple for practitioners. Ultimately, we hope our reproducible pipelines can enable clinical translation and create real impact. The library is available at https://github.com/jbdel/vilmedic.

14:00-15:00 (Forum)
[DEMO] **MMEKG: Multi-modal Event Knowledge Graph towards Universal Representation across Modalities**
Yubo Ma, Zehao Wang, Mukai Li, Yixin Cao, Meiqi Chen, Xinze Li, Wenqi Sun, Kunquan Deng, Kun Wang, Aixin Sun and Jing Shao
Events are fundamental building blocks of real-world happenings. In this paper, we present a large-scale, multi-modal event knowledge graph named MMEKG. MMEKG unifies different modalities of knowledge via events, which complement and disambiguate each other. Specifically, MMEKG incorporates (i) over 990 thousand concept events with 644 relation types to cover most types of happenings, and (ii) over 863 million instance events connected through 934 million relations, which provide rich contextual information in texts and/or images. To collect billion-scale instance events and relations among them, we additionally develop an efficient yet effective pipeline for textual/visual knowledge extraction system. We also develop an induction strategy to create million-scale concept events and a schema organizing all events and relations in MMEKG. To this end, we also provide a pipeline enabling our system to seamlessly parse texts/images to event graphs and to retrieve multi-modal knowledge at both concept- and instance-levels.

Poster Session 2: Machine Learning for NLP

14:00-15:00 (Forum)

14:00-15:00 (Forum)
#27 ConTinTin: Continual Learning from Task Instructions
Wenpeng Yin, Jia Li and Caiming Xiong
The mainstream machine learning paradigms for NLP often work with two underlying presumptions. First, the target task is predefined and static; a system merely needs to learn to solve it exclusively. Second, the supervision of a task mainly comes from a set of labeled examples. A question arises: how to build a system that can keep learning new tasks from their instructions? This work defines a new learning paradigm ConTinTin (Continual Learning from Task Instructions), in which a system should learn a sequence of new tasks one by one, each task is explained by a piece of textual instruction. The system is required to (i) generate the expected outputs of a new task by learning from its instruction, (ii) transfer the knowledge acquired from upstream tasks to help solve downstream tasks (i.e., forward-transfer), and (iii) retain or even improve the performance on earlier tasks after learning new tasks (i.e., backward-transfer). This new problem is studied on a stream of more than 60 tasks, each equipped with an instruction. Technically, our method InstructionSpeak contains two strategies that make full use of task instructions to improve forward-transfer and backward-transfer: one is to learn from negative outputs, the other is to re-visit instructions of previous tasks. To our knowledge, this is the first time to study ConTinTin in NLP. In addition to the problem formulation and our promising approach, this work also contributes to providing rich analyses for the community to better understand this novel learning problem.

14:00-15:00 (Forum)
#28 Learning Disentangled Textual Representations via Statistical Measures of Similarity
Pierre Colombo, Guillaume Staerman, Nathan Noiry and Pablo Piantanida
When working with textual data, a natural application of disentangled representations is the fair classification where the goal is to make predictions without being biased (or influenced) by sensible attributes that may be present in the data (e.g., age, gender or race). Dominant approaches to disentangle a sensitive attribute from textual representations rely on learning simultaneously a penalization term that involves either an adversary loss (e.g., a discriminator) or an information measure (e.g., mutual information). However, these methods require the training of a deep neural network with several parameter updates for each update of the representation model. As a matter of fact, the resulting nested optimization loop is both times consuming, adding complexity to the optimization dynamic, and requires a fine hyperparameter selection (e.g., learning rates, architecture). In this work, we introduce a family of regularizers for learning disentangled representations that do not require training. These regularizers are based on statistical measures of similarity between the conditional probability distributions with respect to the sensible attributes. Our novel regularizers do not require additional training, are faster and do not involve additional tuning while achieving better results both when combined with pretrained and randomly initialized text encoders.

14:00-15:00 (Forum)
#29 Coherence boosting: When your pretrained language model is not paying enough attention
Nikolay Malkin, Zhen Wang and Nebojsa Jojic
Long-range semantic coherence remains a challenge in automatic language generation and understanding. We demonstrate that large language models have insufficiently learned the effect of distant words on next-token prediction. We present coherence boosting, an inference procedure that increases a LM's focus on a long context. We show the benefits of coherence boosting with pretrained models by distributional analyses of generated ordinary text and dialog responses. It is also found that coherence boosting with state-of-the-art models for various zero-shot NLP tasks yields performance gains with no additional training.

14:00-15:00 (Forum)
#30 Sparsifying Transformer Models with Trainable Representation Pooling
Michał Pietruszka, Łukasz Borchmann and Łukasz Garncarek
We propose a novel method to sparsify attention in the Transformer model by learning to select the most-informative token representations during the training process, thus focusing on the task-specific parts of an input. A reduction of quadratic time and memory complexity to sublinear was achieved due to a robust trainable top-k operator. Our experiments on a challenging long document summarization task show that even our simple baseline performs comparably to the current SOTA, and with trainable pooling we can retain its top quality, while being $1.8\times$ faster during training, $4.5\times$ faster during inference, and up to $13\times$ more computationally efficient in the decoder.

14:00-15:00 (Forum)
#31 Imputing Out-of-Vocabulary Embeddings with LOVE Makes LanguageModels Robust with Little Cost
Lihu Chen, Gael Varoquaux and Fabian M. Suchanek
State-of-the-art NLP systems represent inputs with word embeddings, but these are brittle when faced with Out-of-Vocabulary (OOV) words. To address this issue, we follow the principle of mimick-like models to generate vectors for unseen words, by learning the behavior of pre-trained embeddings using only the surface form of words. We present a simple contrastive learning framework, LOVE, which extends the word representation of an existing pre-trained language model (such as BERT) and makes it robust to OOV with few additional parameters. Extensive evaluations demonstrate that our lightweight model achieves similar or even better performances than prior competitors, both on original datasets and on corrupted variants. Moreover, it can be used in a plug-and-play fashion with FastText and BERT, where it significantly improves their robustness.

14:00-15:00 (Forum)
#32 BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models
Elad Ben Zaken, Yoav Goldberg and Shauli Ravfogel
We introduce BitFit, a sparse-finetuning method where only the bias-terms of the model (or a subset of them) are being modified. We show that with small-to-medium training data, applying BitFit on pre-trained BERT models is competitive with (and sometimes better than) fine-tuning the entire model. For larger data, the method is competitive with other sparse fine-tuning methods. Besides their practical utility, these findings are relevant for the question of understanding the commonly-used process of finetuning: they support the hypothesis that finetuning is mainly about exposing knowledge induced by language-modeling training, rather than learning new task-specific linguistic knowledge.

14:00-15:00 (Forum)
#33 An Information-theoretic Approach to Prompt Engineering Without Ground Truth Labels

Taylor Sorensen, Joshua Robinson, Christopher Michael Rytting, Alexander Glenn Shaw, Kyle Jeffrey Rogers, Alexia Pauline Delorey, Mahmoud Khalil, Nancy Fulda and David Wingate
Pre-trained language models derive substantial linguistic and factual knowledge from the massive corpora on which they are trained, and prompt engineering seeks to align these models to specific tasks. Unfortunately, existing prompt engineering methods require significant amounts of labeled data, access to model parameters, or both. We introduce a new method for selecting prompt templates *without labeled examples* and *without direct access to the model*. Specifically, over a set of candidate templates, we choose the template that maximizes the mutual information between the input and the corresponding model output. Across 8 datasets representing 7 distinct NLP tasks, we show that when a template has high mutual information, it also has high accuracy on the task. On the largest model, selecting prompts with our method gets 90% of the way from the average prompt accuracy to the best prompt accuracy and requires no ground truth labels.

14:00-15:00 (Forum)
#34 Kronecker Decomposition for GPT Compression
Ali Edalati, Marzieh S. Tahaei, Ahmad Rashid, Vahid Partovi Nia, James J. Clark and Mehdi Rezagholizadeh
GPT is an auto-regressive Transformer-based pre-trained language model which has attracted a lot of attention in the natural language processing (NLP) domain. The success of GPT is mostly attributed to its pre-training on huge amount of data and its large number of parameters. Despite the superior performance of GPT, this overparameterized nature of GPT can be very prohibitive for deploying this model on devices with limited computational power or memory. This problem can be mitigated using model compression techniques; however, compressing GPT models has not been investigated much in the literature. In this work, we use Kronecker decomposition to compress the linear mappings of the GPT-2 model. Our Kronecker GPT-2 model (KnGPT2) is initialized based on the Kronecker decomposed version of the GPT-2 model and then is undergone a very light pre- training on only a small portion of the training data with intermediate layer knowledge distillation (ILKD). Finally, our KnGPT2 is fine-tuned on downstream tasks using ILKD as well. We evaluate our model on both language modeling and General Language Understanding Evaluation benchmark tasks and show that with more efficient pre-training and similar number of parameters, our KnGPT2 outperforms the existing DistilGPT2 model significantly.

14:00-15:00 (Forum)
#35 Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models
Robert L. Logan IV, Ivana Balazevic, Eric Wallace, Fabio Petroni, Sameer Singh and Sebastian Riedel
Prompting language models (LMs) with training examples and task descriptions has been seen as critical to recent successes in few-shot learning. In this work, we show that finetuning LMs in the few-shot setting can considerably reduce the need for prompt engineering. In fact, one can use null prompts, prompts that contain neither task-specific templates nor training examples, and achieve competitive accuracy to manually-tuned prompts across a wide range of tasks. While finetuning LMs does introduce new parameters for each downstream task, we show that this memory overhead can be substantially reduced: finetuning only the bias terms can achieve comparable or better accuracy than standard finetuning while only updating 0.1

14:00-15:00 (Forum)
#36 Open Vocabulary Extreme Classification Using Generative Models
Daniel Simig, Fabio Petroni, Pouya Yanki, Kashyap Popat, Christina Du, Sebastian Riedel and Majid Yazdani
The extreme multi-label classification (XMC) task aims at tagging content with a subset of labels from an extremely large label set. The label vocabulary is typically defined in advance by domain experts and assumed to capture all necessary tags. However in real world scenarios this label set, although large, is often incomplete and experts frequently need to refine it. To develop systems that simplify this process, we introduce the task of open vocabulary XMC (OXMC): given a piece of content, predict a set of labels, some of which may be outside of the known tag set. Hence, in addition to not having training data for some labels–as is the case in zero-shot classification–models need to invent some labels on-thefly. We propose GROOV, a fine-tuned seq2seq model for OXMC that generates the set of labels as a flat sequence and is trained using a novel loss independent of predicted label order. We show the efficacy of the approach, experimenting with popular XMC datasets for which GROOV is able to predict meaningful labels outside the given vocabulary while performing on par with state-of-the-art solutions for known labels.

14:00-15:00 (Forum)
#37 CUE Vectors: Modular Training of Language Models Conditioned on Diverse Contextual Signals
Scott Novotney, Sreeparna Mukherjee, Zeeshan Ahmed and Andreas Stolcke
We propose a framework to modularize the training of neural language models that use diverse forms of context by eliminating the need to jointly train context and within-sentence encoders. Our approach, contextual universal embeddings (CUE), trains LMs on one type of contextual data and adapts to novel context types. The model consists of a pretrained neural sentence LM, a BERT-based contextual encoder, and a masked transfomer decoder that estimates LM probabilities using sentence-internal and contextual evidence. When contextually annotated data is unavailable, our model learns to combine contextual and sentence-internal information using noisy oracle unigram embeddings as a proxy. Real context data can be introduced later and used to adapt a small number of parameters that map contextual data into the decoder's embedding space. We validate the CUE framework on a NYTimes text corpus with multiple metadata types, for which the LM perplexity can be lowered from 36.6 to 27.4 by conditioning on context. Bootstrapping a contextual LM with only a subset of the metadata during training retains 85

14:00-15:00 (Forum)
#38 Aligned Weight Regularizers for Pruning Pretrained Neural Networks
James O' Neill, Sourav Dutta and Haytham Assem
Pruning aims to reduce the number of parameters while maintaining performance close to the original network. This work proposes a novel *self-distillation* based pruning strategy, whereby the representational similarity between the pruned and unpruned versions of the same network is maximized. Unlike previous approaches that treat distillation and pruning separately, we use distillation to inform the pruning criteria, without requiring a separate student network as in knowledge distillation. We show that the proposed *cross-correlation objective for self-distilled pruning* implicitly encourages sparse solutions, naturally complementing magnitude-based pruning criteria. Experiments on the GLUE and XGLUE benchmarks show that self-distilled pruning increases mono- and cross-lingual language model performance. Self-distilled pruned models also outperform smaller Transformers with an equal number of parameters and are competitive against (6 times) larger distilled networks. We also observe that self-distillation (1) maximizes class separability, (2) increases the signal-to-noise ratio, and (3) converges faster after pruning steps, providing further insights into why self-distilled pruning improves generalization.

14:00-15:00 (Forum)
[TACL] **#39 PADA: Example-based Prompt Learning for on-the-fly Adaptation to Unseen Domains**

Roi Reichart, Eyal Ben-David and Nadav Oved

14:00-15:00 (Forum)
[TACL] **#40 Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP**
Timo Schick, Sahana Udupa and Hinrich Schütze

14:00-15:00 (Forum)
[TACL] **#41 Compressing Large-Scale Transformer-Based Models: A Case Study on BERT**
Prakhar Ganesh, Yao Chen, Xin Lou, Mohammad Khan, Yin Yang, Hassan Sajjad, Preslav Nakov, Deming Chen and Marianne Winslett

14:00-15:00 (Forum)
[TACL] **#42 Towards General Natural Language Understanding with Probabilistic Worldbuilding**
Abulhair Saparov and Tom Mitchell

14:00-15:00 (Forum)
#43 When to Use Multi-Task Learning vs Intermediate Fine-Tuning for Pre-Trained Encoder Transfer Learning
Orion Weller, Kevin Seppi and Matt Gardner
Transfer learning (TL) in natural language processing (NLP) has seen a surge of interest in recent years, as pre-trained models have shown an impressive ability to transfer to novel tasks. Three main strategies have emerged for making use of multiple supervised datasets during fine-tuning: training on an intermediate task before training on the target task (STILTs), using multi-task learning (MTL) to train jointly on a supplementary task and the target task (pairwise MTL), or simply using MTL to train jointly on all available datasets (MTL-ALL). In this work, we compare all three TL methods in a comprehensive analysis on the GLUE dataset suite. We find that there is a simple heuristic for when to use one of these techniques over the other: pairwise MTL is better than STILTs when the target task has fewer instances than the supporting task and vice versa. We show that this holds true in more than 92

14:00-15:00 (Forum)
#44 Word2Box: Capturing Set-Theoretic Semantics of Words using Box Embeddings
Shib Sankar Dasgupta, Michael Boratko, Siddhartha Mishra, Shriya Atmakuri, Dhruvesh Patel, Xiang Lorraine Li and Andrew McCallum
Learning representations of words in a continuous space is perhaps the most fundamental task in NLP, however words interact in ways much richer than vector dot product similarity can provide. Many relationships between words can be expressed set-theoretically, for example, adjective-noun compounds (eg. "red cars"$\subseteq$"cars") and homographs (eg. "tongue"$\cap$"body" should be similar to "mouth", while "tongue"$\cap$"language" should be similar to "dialect") have natural set-theoretic interpretations. Box embeddings are a novel region-based representation which provide the capability to perform these set-theoretic operations. In this work, we provide a fuzzy-set interpretation of box embeddings, and learn box representations of words using a set-theoretic training objective. We demonstrate improved performance on various word similarity tasks, particularly on less common words, and perform a quantitative and qualitative analysis exploring the additional unique expressivity provided by Word2Box.

14:00-15:00 (Forum)
#45 E-LANG: Energy-Based Joint Inferencing of Super and Swift Language Models
Mohammad Akbari, Amin Banitalebi-Dehkordi and Yong Zhang
Building huge and highly capable language models has been a trend in the past years. Despite their great performance, they incur high computational cost. A common solution is to apply model compression or choose light-weight architectures, which often need a separate fixed-size model for each desirable computational budget, and may lose performance in case of heavy compression. This paper proposes an effective dynamic inference approach, called E-LANG, which distributes the inference between large accurate Super-models and light-weight Swift models. To this end, a decision making module routes the inputs to Super or Swift models based on the energy characteristics of the representations in the latent space. This method is easily adoptable and architecture agnostic. As such, it can be applied to black-box pre-trained models without a need for architectural manipulations, reassembling of modules, or re-training. Unlike existing methods that are only applicable to encoder-only backbones and classification tasks, our method also works for encoder-decoder structures and sequence-to-sequence tasks such as translation. The E-LANG performance is verified through a set of experiments with T5 and BERT backbones on GLUE, SuperGLUE, and WMT. In particular, we outperform T5-11B with an average computations speed-up of 3.3X on GLUE and 2.9X on SuperGLUE. We also achieve BERT-based SOTA on GLUE with 3.2X less computations. Code and demo are available in supplementary materials.

14:00-15:00 (Forum)
#46 SHIELD: Defending Textual Neural Networks against Multiple Black-Box Adversarial Attacks with Stochastic Multi-Expert Patcher
Thai Le, Noseong Park and Dongwon Lee
Even though several methods have proposed to defend textual neural network (NN) models against black-box adversarial attacks, they often defend against a specific text perturbation strategy and/or require re-training the models from scratch. This leads to a lack of generalization in practice and redundant computation. In particular, the state-of-the-art transformer models (e.g., BERT, RoBERTa) require great time and computation resources. By borrowing an idea from software engineering, in order to address these limitations, we propose a novel algorithm, SHIELD, which modifies and re-trains only the last layer of a textual NN, and thus it "patches" and "transforms" the NN into a stochastic weighted ensemble of multi-expert prediction heads. Considering that most of current black-box attacks rely on iterative search mechanisms to optimize their adversarial perturbations, SHIELD confuses the attackers by automatically utilizing different weighted ensembles of predictors depending on the input. In other words, SHIELD breaks a fundamental assumption of the attack, which is a victim NN model remains constant during an attack. By conducting comprehensive experiments, we demonstrate that all of CNN, RNN, BERT, and RoBERTa-based textual NNs, once patched by SHIELD, exhibit a relative enhancement of 15

14:00-15:00 (Forum)
#47 Unsupervised multiple-choice question generation for out-of-domain Q&A fine-tuning
Guillaume Le Berre, Christophe Cerisara, Philippe Langlais and Guy Lapalme
Pre-trained models have shown very good performances on a number of question answering benchmarks especially when fine-tuned on multiple question answering datasets at once. In this work, we propose an approach for generating a fine-tuning dataset thanks to a rule-based

algorithm that generates questions and answers from unannotated sentences. We show that the state-of-the-art model UnifiedQA can greatly benefit from such a system on a multiple-choice benchmark about physics, biology and chemistry it has never been trained on. We further show that improved performances may be obtained by selecting the most challenging distractors (wrong answers), with a dedicated ranker based on a pretrained RoBERTa model.

14:00-15:00 (Forum)
#48 "That Is a Suspicious Reaction!": Interpreting Logits Variation to Detect NLP Adversarial Attacks
Edoardo Mosca, Shreyash Agarwal, Javier Rando Ramírez and Georg Groh
Adversarial attacks are a major challenge faced by current machine learning research. These purposely crafted inputs fool even the most advanced models, precluding their deployment in safety-critical applications. Extensive research in computer vision has been carried to develop reliable defense strategies. However, the same issue remains less explored in natural language processing. Our work presents a model-agnostic detector of adversarial text examples. The approach identifies patterns in the logits of the target classifier when perturbing the input text. The proposed detector improves the current state-of-the-art performance in recognizing adversarial inputs and exhibits strong generalization capabilities across different NLP models, datasets, and word-level attacks.

14:00-15:00 (Forum)
#49 Label Semantic Aware Pre-training for Few-shot Text Classification
Aaron Mueller, Jason Krone, Salvatore Romeo, Saab Mansour, Elman Mansimov, Yi Zhang and Dan Roth
In text classification tasks, useful information is encoded in the label names. Label semantic aware systems have leveraged this information for improved text classification performance during fine-tuning and prediction. However, use of label-semantics during pre-training has not been extensively explored. We therefore propose Label Semantic Aware Pre-training (LSAP) to improve the generalization and data efficiency of text classification systems. LSAP incorporates label semantics into pre-trained generative models (T5 in our case) by performing secondary pre-training on labeled sentences from a variety of domains. As domain-general pre-training requires large amounts of data, we develop a filtering and labeling pipeline to automatically create sentence-label pairs from unlabeled text. We perform experiments on intent (ATIS, Snips, TOPv2) and topic classification (AG News, Yahoo! Answers). LSAP obtains significant accuracy improvements over state-of-the-art models for few-shot text classification while maintaining performance comparable to state of the art in high-resource settings.

14:00-15:00 (Forum)
#50 Deduplicating Training Data Makes Language Models Better
Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch and Nicholas Carlini
We find that existing language modeling datasets contain many near-duplicate examples and long repetitive substrings. As a result, over 1We develop two tools that allow us to deduplicate training datasets—for example removing from C4 a single 61 word English sentence that is repeated over 60,000 times. Deduplication allows us to train models that emit memorized text ten times less frequently and require fewer training steps to achieve the same or better accuracy. We can also reduce train-test overlap, which affects over 4Code for deduplication is released at https://github.com/google-research/deduplicate-text-datasets.

14:00-15:00 (Forum)
#51 ∞-former: Infinite Memory Transformer
Pedro Henrique Martins, Zita Marinho and Andre Martins
Transformers are unable to model long-term memories effectively, since the amount of computation they need to perform grows with the context length. While variations of efficient transformers have been proposed, they all have a finite memory capacity and are forced to drop old information. In this paper, we propose the ∞-former, which extends the vanilla transformer with an unbounded long-term memory. By making use of a continuous-space attention mechanism to attend over the long-term memory, the ∞-former's attention complexity becomes independent of the context length, trading off memory length with precision. In order to control where precision is more important, ∞-former maintains "sticky memories," being able to model arbitrarily long contexts while keeping the computation budget fixed. Experiments on a synthetic sorting task, language modeling, and document grounded dialogue generation demonstrate the ∞-former's ability to retain information from long sequences.

14:00-15:00 (Forum)
#52 Adapting Coreference Resolution Models through Active Learning
Michelle Yuan, Patrick Xia, Chandler May, Benjamin Van Durme and Jordan Lee Boyd-Graber
Neural coreference resolution models trained on one dataset may not transfer to new, low-resource domains. Active learning mitigates this problem by sampling a small subset of data for annotators to label. While active learning is well-defined for classification tasks, its application to coreference resolution is neither well-defined nor fully understood. This paper explores how to actively label coreference, examining sources of model uncertainty and document reading costs. We compare uncertainty sampling strategies and their advantages through thorough error analysis. In both synthetic and human experiments, labeling spans within the same document is more effective than annotating spans across documents. The findings contribute to a more realistic development of coreference resolution models.

14:00-15:00 (Forum)
#53 mLUKE: The Power of Entity Representations in Multilingual Pretrained Language Models
Ryokan Ri, Ikuya Yamada and Yoshimasa Tsuruoka
Recent studies have shown that multilingual pretrained language models can be effectively improved with cross-lingual alignment information from Wikipedia entities. However, existing methods only exploit entity information in pretraining and do not explicitly use entities in downstream tasks. In this study, we explore the effectiveness of leveraging entity representations for downstream cross-lingual tasks. We train a multilingual language model with 24 languages with entity representations and show the model consistently outperforms word-based pretrained models in various cross-lingual transfer tasks. We also analyze the model and the key insight is that incorporating entity representations into the input allows us to extract more language-agnostic features. We also evaluate the model with a multilingual cloze prompt task with the mLAMA dataset. We show that entity-based prompt elicits correct factual knowledge more likely than using only word representations.

14:00-15:00 (Forum)
#54 Continual Sequence Generation with Adaptive Compositional Modules
Yanzhe Zhang, Xuezhi Wang and Diyi Yang
Continual learning is essential for real-world deployment when there is a need to quickly adapt the model to new tasks without forgetting knowledge of old tasks. Existing work on continual sequence generation either always reuses existing parameters to learn new tasks, which is vulnerable to catastrophic forgetting on dissimilar tasks, or blindly adds new parameters for every new task, which could prevent knowledge sharing between similar tasks. To get the best of both worlds, in this work, we propose continual sequence generation with adaptive compositional modules to adaptively add modules in transformer architectures and compose both old and new modules for new tasks. We also incorporate pseudo experience replay to facilitate knowledge transfer in those shared modules. Experiment results on various sequences of generation tasks show that our framework can adaptively add modules or reuse modules based on task similarity, outperforming state-of-the-art baselines in terms of both performance and parameter efficiency. We make our code public at https://github.com/GT-SALT/Adaptive-Compositional-Modules.

14:00-15:00 (Forum)
#55 Sharpness-Aware Minimization Improves Language Model Generalization
Dara Bahri, Hossein Mobahi and Yi Tay
The allure of superhuman-level capabilities has led to considerable interest in language models like GPT-3 and T5, wherein the research has, by and large, revolved around new model architectures, training tasks, and loss objectives, along with substantial engineering efforts to scale up model capacity and dataset size. Comparatively little work has been done to improve the generalization of these models through better optimization. In this work, we show that Sharpness-Aware Minimization (SAM), a recently proposed optimization procedure that encourages convergence to flatter minima, can substantially improve the generalization of language models without much computational overhead. We show that SAM is able to boost performance on SuperGLUE, GLUE, Web Questions, Natural Questions, Trivia QA, and TyDiQA, with particularly large gains when training data for these tasks is limited.

14:00-15:00 (Forum)
#56 Cluster & Tune: Boost Cold Start Performance in Text Classification
Eyal Shnarch, Ariel Gera, Alon Halfon, Lena Dankin, Leshem Choshen, Ranit Aharonov and Noam Slonim
In real-world scenarios, a text classification task often begins with a cold start, when labeled data is scarce. In such cases, the common practice of fine-tuning pre-trained models, such as BERT, for a target classification task, is prone to produce poor performance. We suggest a method to boost the performance of such models by adding an intermediate unsupervised classification task, between the pre-training and fine-tuning phases. As such an intermediate task, we perform clustering and train the pre-trained model on predicting the cluster labels. We test this hypothesis on various data sets, and show that this additional classification phase can significantly improve performance, mainly for topical classification tasks, when the number of labeled instances available for fine-tuning is only a couple of dozen to a few hundred.

14:00-15:00 (Forum)
#57 Uncertainty Determines the Adequacy of the Mode and the Tractability of Decoding in Sequence-to-Sequence Models
Felix Stahlberg, Ilia Kulikov and Shankar Kumar
In many natural language processing (NLP) tasks the same input (e.g. source sentence) can have multiple possible outputs (e.g. translations). To analyze how this ambiguity (also known as intrinsic uncertainty) shapes the distribution learned by neural sequence models we measure sentence-level uncertainty by computing the degree of overlap between references in multi-reference test sets from two different NLP tasks: machine translation (MT) and grammatical error correction (GEC). At both the sentence- and the task-level, intrinsic uncertainty has major implications for various aspects of search such as the inductive biases in beam search and the complexity of exact search. In particular, we show that well-known pathologies such as a high number of beam search errors, the inadequacy of the mode, and the drop in system performance with large beam sizes apply to tasks with high level of ambiguity such as MT but not to less uncertain tasks such as GEC. Furthermore, we propose a novel exact n-best search algorithm for neural sequence models, and show that intrinsic uncertainty affects model uncertainty as the model tends to overly spread out the probability mass for uncertain tasks and sentences.

14:00-15:00 (Forum)
#58 BERT Learns to Teach: Knowledge Distillation with Meta Learning
Wangchunshu Zhou, Canwen Xu and Julian McAuley
We present Knowledge Distillation with Meta Learning (MetaDistil), a simple yet effective alternative to traditional knowledge distillation (KD) methods where the teacher model is fixed during training. We show the teacher network can learn to better transfer knowledge to the student network (i.e., *learning to teach*) with the feedback from the performance of the distilled student network in a meta learning framework. Moreover, we introduce a pilot update mechanism to improve the alignment between the inner-learner and meta-learner in meta learning algorithms that focus on an improved inner-learner. Experiments on various benchmarks show that MetaDistil can yield significant improvements compared with traditional KD algorithms and is less sensitive to the choice of different student capacity and hyperparameters, facilitating the use of KD on different tasks and models.

14:00-15:00 (Forum)
#59 Early Stopping Based on Unlabeled Samples in Text Classification
HongSeok Choi, Dongha Choi and Hyunju Lee
Early stopping, which is widely used to prevent overfitting, is generally based on a separate validation set. However, in low resource settings, validation-based stopping can be risky because a small validation set may not be sufficiently representative, and the reduction in the number of samples by validation split may result in insufficient samples for training. In this study, we propose an early stopping method that uses unlabeled samples. The proposed method is based on confidence and class distribution similarities. To further improve the performance, we present a calibration method to better estimate the class distribution of the unlabeled samples. The proposed method is advantageous because it does not require a separate validation set and provides a better stopping point by using a large unlabeled set. Extensive experiments are conducted on five text classification datasets and several stop-methods are compared. Our results show that the proposed model even performs better than using an additional validation set as well as the existing stop-methods, in both balanced and imbalanced data settings. Our code is available at https://github.com/DMCB-GIST/BUS-stop.

14:00-15:00 (Forum)
#60 Domain Knowledge Transferring for Pre-trained Language Model via Calibrated Activation Boundary Distillation

Dongha Choi, HongSeok Choi and Hyunju Lee
Since the development and wide use of pretrained language models (PLMs), several approaches have been applied to boost their performance on downstream tasks in specific domains, such as biomedical or scientific domains. Additional pre-training with in-domain texts is the most common approach for providing domain-specific knowledge to PLMs. However, these pre-training methods require considerable in-domain data and training resources and a longer training time. Moreover, the training must be re-performed whenever a new PLM emerges. In this study, we propose a domain knowledge transferring (DoKTra) framework for PLMs without additional in-domain pretraining. Specifically, we extract the domain knowledge from an existing in-domain pretrained language model and transfer it to other PLMs by applying knowledge distillation. In particular, we employ activation boundary distillation, which focuses on the activation of hidden neurons. We also apply an entropy regularization term in both teacher training and distillation to encourage the model to generate reliable output probabilities, and thus aid the distillation. By applying the proposed DoKTra framework to downstream tasks in the biomedical, clinical, and financial domains, our student models can retain a high percentage of teacher performance and even outperform the teachers in certain tasks. Our code is available at https://github.com/DMCB-GIST/DoKTra.

14:00-15:00 (Forum)
#61 The Power of Prompt Tuning for Low-Resource Semantic Parsing
Nathan Schucher, Siva Reddy and Harm de Vries
Prompt tuning has recently emerged as an effective method for adapting pre-trained language models to a number of language understanding and generation tasks. In this paper, we investigate prompt tuning for semantic parsing—the task of mapping natural language utterances onto formal meaning representations. On the low-resource splits of Overnight and TOPv2, we find that a prompt tuned T5-xl significantly outperforms its fine-tuned counterpart, as well as strong GPT-3 and BART baselines. We also conduct ablation studies across different model scales and target representations, finding that, with increasing model scale, prompt tuned T5 models improve at generating target representations that are far from the pre-training distribution.

14:00-15:00 (Forum)
#62 Structured Pruning Learns Compact and Accurate Models
Mengzhou Xia, Zexuan Zhong and Danqi Chen
The growing size of neural language models has led to increased attention in model compression. The two predominant approaches are pruning, which gradually removes weights from a pre-trained model, and distillation, which trains a smaller compact model to match a larger one. Pruning methods can significantly reduce the model size but hardly achieve large speedups as distillation. However, distillation methods require large amounts of unlabeled data and are expensive to train. In this work, we propose a task-specific structured pruning method CoFi (Coarse- and Fine-grained Pruning), which delivers highly parallelizable subnetworks and matches the distillation methods in both accuracy and latency, without resorting to any unlabeled data. Our key insight is to jointly prune coarse-grained (e.g., layers) and fine-grained (e.g., heads and hidden units) modules, which controls the pruning decision of each parameter with masks of different granularity. We also devise a layerwise distillation strategy to transfer knowledge from unpruned to pruned models during optimization. Our experiments on GLUE and SQuAD datasets show that CoFi yields models with over 10X speedups with a small accuracy drop, showing its effectiveness and efficiency compared to previous pruning and distillation approaches.

14:00-15:00 (Forum)
#63 SPoT: Better Frozen Model Adaptation through Soft Prompt Transfer
Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou' and Daniel Cer
There has been growing interest in parameter-efficient methods to apply pre-trained language models to downstream tasks. Building on the Prompt Tuning approach of Lester et al. (2021), which learns task-specific soft prompts to condition a frozen pre-trained model to perform different tasks, we propose a novel prompt-based transfer learning approach called SPoT: Soft Prompt Transfer. SPoT first learns a prompt on one or more source tasks and then uses it to initialize the prompt for a target task. We show that SPoT significantly boosts the performance of Prompt Tuning across many tasks. More remarkably, across all model sizes, SPoT matches or outperforms standard Model Tuning (which fine-tunes all model parameters) on the SuperGLUE benchmark, while using up to 27,000× fewer task-specific parameters. To understand where SPoT is most effective, we conduct a large-scale study on task transferability with 26 NLP tasks in 160 combinations, and demonstrate that many tasks can benefit each other via prompt transfer. Finally, we propose an efficient retrieval approach that interprets task prompts as task embeddings to identify similar tasks and predict the most transferable source tasks for a novel target task.

14:00-15:00 (Forum)
#64 Revisiting the Compositional Generalization Abilities of Neural Sequence Models
Arkil Patel, Satwik Bhattamishra, Phil Blunsom and Navin Goyal
Compositional generalization is a fundamental trait in humans, allowing us to effortlessly combine known phrases to form novel sentences. Recent works have claimed that standard seq-to-seq models severely lack the ability to compositionally generalize. In this paper, we focus on one-shot primitive generalization as introduced by the popular SCAN benchmark. We demonstrate that modifying the training distribution in simple and intuitive ways enables standard seq-to-seq models to achieve near-perfect generalization performance, thereby showing that their compositional generalization abilities were previously underestimated. We perform detailed empirical analysis of this phenomenon. Our results indicate that the generalization performance of models is highly sensitive to the characteristics of the training data which should be carefully considered while designing such benchmarks in future.

14:00-15:00 (Forum)
#65 Efficient, Uncertainty-based Moderation of Neural Networks Text Classifiers
Jakob Smedegaard Andersen and Walid Maalej
To maximize the accuracy and increase the overall acceptance of text classifiers, we propose a framework for the efficient, in-operation moderation of classifiers' output. Our framework focuses on use cases in which F1-scores of modern Neural Networks classifiers (ca. 90

14:00-15:00 (Forum)
#66 Why Exposure Bias Matters: An Imitation Learning Perspective of Error Accumulation in Language Generation
Kushal Arora, Layla El Asri, Hareesh Bahuleyan and Jackie CK Cheung
Current language generation models suffer from issues such as repetition, incoherence, and hallucinations. An often-repeated hypothesis for this brittleness of generation models is that it is caused by the training and the generation procedure mismatch, also referred to as exposure bias. In this paper, we verify this hypothesis by analyzing exposure bias from an imitation learning perspective. We show that exposure bias leads to an accumulation of errors during generation, analyze why perplexity fails to capture this accumulation of errors, and empirically show

that this accumulation results in poor generation quality. [2]

14:00-15:00 (Forum)
#67 Metadata Shaping: A Simple Approach for Knowledge-Enhanced Language Models
Simran Arora, Sen Wu, Enci Liu and Christopher Re
Popular language models (LMs) struggle to capture knowledge about rare tail facts and entities. Since widely used systems such as search and personal-assistants must support the long tail of entities that users ask about, there has been significant effort towards enhancing these base LMs with factual knowledge. We observe proposed methods typically start with a base LM and data that has been annotated with entity metadata, then change the model, by modifying the architecture or introducing auxiliary loss terms to better capture entity knowledge. In this work, we question this typical process and ask to what extent can we match the quality of model modifications, with a simple alternative: using a base LM and only changing the data. We propose metadata shaping, a method which inserts substrings corresponding to the readily available entity metadata, e.g. types and descriptions, into examples at train and inference time based on mutual information. Despite its simplicity, metadata shaping is quite effective. On standard evaluation benchmarks for knowledge-enhanced LMs, the method exceeds the base-LM baseline by an average of 4.3 F1 points and achieves state-of-the-art results. We further show the gains are on average 4.4x larger for the slice of examples containing tail vs. popular entities.

14:00-15:00 (Forum)
#68 Composing Structure-Aware Batches for Pairwise Sentence Classification
Andreas Waldis, Tilman Beck and Iryna Gurevych
Identifying the relation between two sentences requires datasets with pairwise annotations. In many cases, these datasets contain instances that are annotated multiple times as part of different pairs. They constitute a structure that contains additional helpful information about the inter-relatedness of the text instances based on the annotations. This paper investigates how this kind of structural dataset information can be exploited during training. We propose three batch composition strategies to incorporate such information and measure their performance over 14 heterogeneous pairwise sentence classification tasks. Our results show statistically significant improvements (up to 3.9

14:00-15:00 (Forum)
#69 Learning Adaptive Axis Attentions in Fine-tuning: Beyond Fixed Sparse Attention Patterns
Zihan Wang, Jiuxiang Gu, Jason Kuen, Handong Zhao, Vlad I Morariu, Ruiyi Zhang, Ani Nenkova, Tong Sun and Jingbo Shang
We present a comprehensive study of sparse attention patterns in Transformer models. We first question the need for pre-training with sparse attention and present experiments showing that an efficient fine-tuning only approach yields a slightly worse but still competitive model. Then we compare the widely used local attention pattern and the less-well-studied global attention pattern, demonstrating that global patterns have several unique advantages. We also demonstrate that a flexible approach to attention, with different patterns across different layers of the model, is beneficial for some tasks. Drawing on this insight, we propose a novel Adaptive Axis Attention method, which learns—during fine-tuning—different attention patterns for each Transformer layer depending on the downstream task. Rather than choosing a fixed attention pattern, the adaptive axis attention method identifies important tokens—for each task and model layer—and focuses attention on those. It does not require pre-training to accommodate the sparse patterns and demonstrates competitive and sometimes better performance against fixed sparse attention patterns that require resource-intensive pre-training.

14:00-15:00 (Forum)
#70 Learning to Robustly Aggregate Labeling Functions for Semi-supervised Data Programming
Ayush Maheshwari, Krishnateja Killamsetty, Ganesh Ramakrishnan, Rishabh K Iyer, Marina Danilevsky and Lucian Popa
A critical bottleneck in supervised machine learning is the need for large amounts of labeled data which is expensive and time-consuming to obtain. Although a small amount of labeled data cannot be used to train a model, it can be used effectively for the generation of human-interpretable labeling functions (LFs). These LFs, in turn, have been used to generate a large amount of additional noisy labeled data in a paradigm that is now commonly referred to as data programming. Previous methods of generating LFs do not attempt to use the given labeled data further to train a model, thus missing opportunities for improving performance. Additionally, since the LFs are generated automatically, they are likely to be noisy, and naively aggregating these LFs can lead to suboptimal results. In this work, we propose an LF-based bi-level optimization framework WISDOM to solve these two critical limitations. WISDOM learns a joint model on the (same) labeled dataset used for LF induction along with any unlabeled data in a semi-supervised manner, and more critically, reweighs each LF according to its goodness, influencing its contribution to the semi-supervised loss using a robust bi-level optimization algorithm. We show that WISDOM significantly outperforms prior approaches on several text classification datasets.

14:00-15:00 (Forum)
#71 Perturbations in the Wild: Leveraging Human-Written Text Perturbations for Realistic Adversarial Attack and Defense
Thai Le, Jooyoung Lee, Kevin Yen, Yifan Hu and Dongwon Lee
We proposes a novel algorithm, ANTHRO, that inductively extracts over 600K human-written text perturbations in the wild and leverages them for realistic adversarial attack. Unlike existing character-based attacks which often deductively hypothesize a set of manipulation strategies, our work is grounded on actual observations from real-world texts. We find that adversarial texts generated by ANTHRO achieve the best trade-off between (1) attack success rate, (2) semantic preservation of the original text, and (3) stealthiness–i.e. indistinguishable from human writings hence harder to be flagged as suspicious. Specifically, our attacks accomplished around 83

14:00-15:00 (Forum)
#72 Revisiting Uncertainty-based Query Strategies for Active Learning with Transformers
Christopher Schröder, Andreas Niekler and Martin Potthast
Active learning is the iterative construction of a classification model through targeted labeling, enabling significant labeling cost savings. As most research on active learning has been carried out before transformer-based language models ("transformers") became popular, despite its practical importance, comparably few papers have investigated how transformers can be combined with active learning to date. This can be attributed to the fact that using state-of-the-art query strategies for transformers induces a prohibitive runtime overhead, which effectively nullifies, or even outweighs the desired cost savings. For this reason, we revisit uncertainty-based query strategies, which had been largely outperformed before, but are particularly suited in the context of fine-tuning transformers. In an extensive evaluation, we connect transformers to experiments from previous research, assessing their performance on five widely used text classification benchmarks. For active learning

[2] Source code to reproduce these experiments is available at https://github.com/kushalarora/quantifying_exposure_bias.

with transformers, several other uncertainty-based approaches outperform the well-known prediction entropy query strategy, thereby challenging its status as most popular uncertainty baseline in active learning for text classification.

14:00-15:00 (Forum)
[TACL] **#73 CANINE: Pre-training an Efficient Tokenization-Free Encoder for Language Representation**
Jonathan Clark, Dan Garrette, Iulia Turc and John Wieting

14:00-15:00 (Forum)
#74 Improved Multi-label Classification under Temporal Concept Drift: Rethinking Group-Robust Algorithms in a Label-Wise Setting
Ilias Chalkidis and Anders Søgaard
In document classification for, e.g., legal and biomedical text, we often deal with hundreds of classes, including very infrequent ones, as well as temporal concept drift caused by the influence of real world events, e.g., policy changes, conflicts, or pandemics. Class imbalance and drift can sometimes be mitigated by resampling the training data to simulate (or compensate for) a known target distribution, but what if the target distribution is determined by unknown future events? Instead of simply resampling uniformly to hedge our bets, we focus on the underlying optimization algorithms used to train such document classifiers and evaluate several group-robust optimization algorithms, initially proposed to mitigate group-level disparities. Reframing group-robust algorithms as adaptation algorithms under concept drift, we find that Invariant Risk Minimization and Spectral Decoupling outperform sampling-based approaches to class imbalance and concept drift, and lead to much better performance on minority classes. The effect is more pronounced the larger the label set.

14:00-15:00 (Forum)
#75 The Trade-offs of Domain Adaptation for Neural Language Models
David Grangier and Dan Iter
This work connects language model adaptation with concepts of machine learning theory. We consider a training setup with a large out-of-domain set and a small in-domain set. We derive how the benefit of training a model on either set depends on the size of the sets and the distance between their underlying distributions. We analyze how out-of-domain pre-training before in-domain fine-tuning achieves better generalization than either solution independently. Finally, we present how adaptation techniques based on data selection, such as importance sampling, intelligent data selection and influence functions, can be presented in a common framework which highlights their similarity and also their subtle differences.

14:00-15:00 (Forum)
#76 AdapLeR: Speeding up Inference by Adaptive Length Reduction
Ali Modarressi, Hosein Mohebbi and Mohammad Taher Pilehvar
Pre-trained language models have shown stellar performance in various downstream tasks. But, this usually comes at the cost of high latency and computation, hindering their usage in resource-limited settings. In this work, we propose a novel approach for reducing the computational cost of BERT with minimal loss in downstream performance. Our method dynamically eliminates less contributing tokens through layers, resulting in shorter lengths and consequently lower computational cost. To determine the importance of each token representation, we train a Contribution Predictor for each layer using a gradient-based saliency method. Our experiments on several diverse classification tasks show speedups up to 22x during inference time without much sacrifice in performance. We also validate the quality of the selected tokens in our method using human annotations in the ERASER benchmark. In comparison to other widely used strategies for selecting important tokens, such as saliency and attention, our proposed method has a significantly lower false positive rate in generating rationales. Our code is freely available at https://github.com/amodaresi/AdapLeR.

14:00-15:00 (Forum)
#77 Disentangled Sequence to Sequence Learning for Compositional Generalization
Hao Zheng and Mirella Lapata
There is mounting evidence that existing neural network models, in particular the very popular sequence-to-sequence architecture, struggle to systematically generalize to unseen compositions of seen components. We demonstrate that one of the reasons hindering compositional generalization relates to representations being entangled. We propose an extension to sequence-to-sequence models which encourage disentanglement by adaptively re-encoding (at each time step) the source input. Specifically, we condition the source representations on the newly decoded target context which makes it easier for the encoder to exploit specialized information for each prediction rather than capturing it all in a single forward pass. Experimental results on semantic parsing and machine translation empirically show that our proposal delivers more disentangled representations and better generalization.

14:00-15:00 (Forum)
#78 Distributionally Robust Finetuning BERT for Covariate Drift in Spoken Language Understanding
Samuel Broscheit, Quynh Do and Judith Gaspers
In this study, we investigate robustness against covariate drift in spoken language understanding (SLU). Covariate drift can occur in SLU when there is a drift between training and testing regarding what users request or how they request it. To study this we propose a method that exploits natural variations in data to create a covariate drift in SLU datasets. Experiments show that a state-of-the-art BERT-based model suffers performance loss under this drift. To mitigate the performance loss, we investigate distributionally robust optimization (DRO) for finetuning BERT-based models. We discuss some recent DRO methods, propose two new variants and empirically show that DRO improves robustness under drift.

14:00-15:00 (Forum)
#79 DMix: Adaptive Distance-aware Interpolative Mixup
Ramit Sawhney, Megh Thakkar, Shrey Pandit, Ritesh Singh Soun, Di Jin, Diyi Yang and Lucie Flek
Interpolation-based regularisation methods such as Mixup, which generate virtual training samples, have proven to be effective for various tasks and modalities. We extend Mixup and propose DMix, an adaptive distance-aware interpolative Mixup that selects samples based on their diversity in the embedding space. DMix leverages the hyperbolic space as a similarity measure among input samples for a richer encoded representation. DMix achieves state-of-the-art results on sentence classification over existing data augmentation methods on 8 benchmark datasets across English, Arabic, Turkish, and Hindi languages while achieving benchmark F1 scores in 3 times less number of iterations. We

probe the effectiveness of DMix in conjunction with various similarity measures and qualitatively analyze the different components. DMix being generalizable, can be applied to various tasks, models and modalities.

14:00-15:00 (Forum)
#80 Noisy Channel Language Model Prompting for Few-Shot Text Classification
Sewon Min, Mike Lewis, Hannaneh Hajishirzi and Luke Zettlemoyer
We introduce a noisy channel approach for language model prompting in few-shot text classification. Instead of computing the likelihood of the label given the input (referred as direct models), channel models compute the conditional probability of the input given the label, and are thereby required to explain every word in the input. We use channel models for recently proposed few-shot learning methods with no or very limited updates to the language model parameters, via either in-context demonstration or prompt tuning. Our experiments show that, for both methods, channel models significantly outperform their direct counterparts, which we attribute to their stability, i.e., lower variance and higher worst-case accuracy. We also present extensive ablations that provide recommendations for when to use channel prompt tuning instead of other competitive models (e.g., direct head tuning): channel prompt tuning is preferred when the number of training examples is small, labels in the training data are imbalanced, or generalization to unseen labels is required.

14:00-15:00 (Forum)
#81 On the Importance of Effectively Adapting Pretrained Language Models for Active Learning
Katerina Margatina, Loic Barrault and Nikolaos Aletras
Recent active learning (AL) approaches in Natural Language Processing (NLP) proposed using off-the-shelf pretrained language models (LMs). In this paper, we argue that these LMs are not adapted effectively to the downstream task during AL and we explore ways to address this issue. We suggest to first adapt the pretrained LM to the target task by continuing training with all the available unlabeled data and then use it for AL. We also propose a simple yet effective fine-tuning method to ensure that the adapted LM is properly trained in both low and high resource scenarios during AL. Our experiments demonstrate that our approach provides substantial data efficiency improvements compared to the standard fine-tuning approach, suggesting that a poor training strategy can be catastrophic for AL.

14:00-15:00 (Forum)
#82 *Generalized but not Robust?* Comparing the Effects of Data Modification Methods on Out-of-Domain Generalization and Adversarial Robustness
Tejas Gokhale, Swaroop Mishra, Man Luo, Bhavdeep Singh Sachdeva and Chitta Baral
Data modification, either via additional training datasets, data augmentation, debiasing, and dataset filtering, has been proposed as an effective solution for generalizing to out-of-domain (OOD) inputs, in both natural language processing and computer vision literature. However, the effect of data modification on adversarial robustness remains unclear. In this work, we conduct a comprehensive study of common data modification strategies and evaluate not only their in-domain and OOD performance, but also their adversarial robustness (AR). We also present results on a two-dimensional synthetic dataset to visualize the effect of each method on the training distribution. This work serves as an empirical study towards understanding the relationship between generalizing to unseen domains and defending against adversarial perturbations. Our findings suggest that more data (either via additional datasets or data augmentation) benefits both OOD accuracy and AR. However, data filtering (previously shown to improve OOD accuracy on natural language inference) hurts OOD accuracy on other tasks such as question answering and image classification. We provide insights from our experiments to inform future work in this direction.

14:00-15:00 (Forum)
[DEMO] **TimeLMs: Diachronic Language Models from Twitter**
Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke and Jose Camacho-collados
Despite its importance, the time variable has been largely neglected in the NLP and language model literature. In this paper, we present TimeLMs, a set of language models specialized on diachronic Twitter data. We show that a continual learning strategy contributes to enhancing Twitter-based language models' capacity to deal with future and out-of-distribution tweets, while making them competitive with standardized and more monolithic benchmarks. We also perform a number of qualitative analyses showing how they cope with trends and peaks in activity involving specific named entities or concept drift. TimeLMs is available at github.com/cardiffnlp/timelms.

14:00-15:00 (Forum)
[DEMO] **Adaptor: Objective-Centric Adaptation Framework for Language Models**
Michal Štefánik, Vít Novotný, Nikola Groverová and Petr Sojka
This paper introduces Adaptor library, which transposes traditional model-centric approach composed of pre-training + fine-tuning steps to objective-centric approach, composing the training process by applications of selected objectives. We survey research directions that can benefit from enhanced objective-centric experimentation in multitask training, custom objectives development, dynamic training curricula, or domain adaptation. Adaptor aims to ease reproducibility of these research directions in practice. Finally, we demonstrate the practical applicability of Adaptor in selected unsupervised domain adaptation scenarios.

Mini Break
15:00-15:15 - **Auditorium** (Forum)

Spotlight Talks by Young Rising Stars (STIRS)
15:15-16:30 - **Auditorium** (Auditorium)

Coffee Break
16:30-17:00 - **Auditorium** (Forum)

Session 3 - 17:00-18:00

Machine Learning for NLP 2

17:00-18:00 (The Liffey B)

17:00-17:15 (The Liffey B)
SHIELD: Defending Textual Neural Networks against Multiple Black-Box Adversarial Attacks with Stochastic Multi-Expert Patcher
Thai Le, Noseong Park and Dongwon Lee
Even though several methods have proposed to defend textual neural network (NN) models against black-box adversarial attacks, they often defend against a specific text perturbation strategy and/or require re-training the models from scratch. This leads to a lack of generalization in practice and redundant computation. In particular, the state-of-the-art transformer models (e.g., BERT, RoBERTa) require great time and computation resources. By borrowing an idea from software engineering, in order to address these limitations, we propose a novel algorithm, SHIELD, which modifies and re-trains only the last layer of a textual NN, and thus it "patches" and "transforms" the NN into a stochastic weighted ensemble of multi-expert prediction heads. Considering that most of current black-box attacks rely on iterative search mechanisms to optimize their adversarial perturbations, SHIELD confuses the attackers by automatically utilizing different weighted ensembles of predictors depending on the input. In other words, SHIELD breaks a fundamental assumption of the attack, which is a victim NN model remains constant during an attack. By conducting comprehensive experiments, we demonstrate that all of CNN, RNN, BERT, and RoBERTa-based textual NNs, once patched by SHIELD, exhibit a relative enhancement of 15

17:15-17:30 (The Liffey B)
"That Is a Suspicious Reaction!": Interpreting Logits Variation to Detect NLP Adversarial Attacks
Edoardo Mosca, Shreyash Agarwal, Javier Rando Ramírez and Georg Groh
Adversarial attacks are a major challenge faced by current machine learning research. These purposely crafted inputs fool even the most advanced models, precluding their deployment in safety-critical applications. Extensive research in computer vision has been carried to develop reliable defense strategies. However, the same issue remains less explored in natural language processing. Our work presents a model-agnostic detector of adversarial text examples. The approach identifies patterns in the logits of the target classifier when perturbing the input text. The proposed detector improves the current state-of-the-art performance in recognizing adversarial inputs and exhibits strong generalization capabilities across different NLP models, datasets, and word-level attacks.

17:30-17:45 (The Liffey B)
[TACL] **Compressing Large-Scale Transformer-Based Models: A Case Study on BERT**
Prakhar Ganesh, Yao Chen, Xin Lou, Mohammad Khan, Yin Yang, Hassan Sajjad, Preslav Nakov, Deming Chen and Marianne Winslett

17:45-18:00 (The Liffey B)
ABC: Attention with Bounded-memory Control
Hao Peng, Jungo Kasai, Nikolaos Pappas, Dani Yogatama, Zhaofeng Wu, Lingpeng Kong, Roy Schwartz and Noah Smith
Transformer architectures have achieved state- of-the-art results on a variety of natural language processing (NLP) tasks. However, their attention mechanism comes with a quadratic complexity in sequence lengths, making the computational overhead prohibitive, especially for long sequences. Attention context can be seen as a random-access memory with each token taking a slot. Under this perspective, the memory size grows linearly with the sequence length, and so does the overhead of reading from it. One way to improve the efficiency is to bound the memory size. We show that disparate approaches can be subsumed into one abstraction, attention with bounded-memory control (ABC), and they vary in their organization of the memory. ABC reveals new, unexplored possibilities. First, it connects several efficient attention variants that would otherwise seem apart. Second, this abstraction gives new insights—an established approach (Wang et al., 2020b) previously thought to not be applicable in causal attention, actually is. Last, we present a new instance of ABC, which draws inspiration from existing ABC approaches, but replaces their heuristic memory-organizing functions with a learned, contextualized one. Our experiments on language modeling, machine translation, and masked language model finetuning show that our approach outperforms previous efficient attention models; compared to the strong transformer baselines, it significantly improves the inference time and space efficiency with no or negligible accuracy loss.

Machine Translation and Multilinguality 2

17:00-18:00 (The Liffey A)

17:00-17:15 (The Liffey A)
Bilingual alignment transfers to multilingual alignment for unsupervised parallel text mining
Chih-chan Tien and Shane Steinert-Threlkeld
This work presents methods for learning cross-lingual sentence representations using paired or unpaired bilingual texts. We hypothesize that the cross-lingual alignment strategy is transferable, and therefore a model trained to align only two languages can encode multilingually more aligned representations. We thus introduce dual-pivot transfer: training on one language pair and evaluating on other pairs. To study this theory, we design unsupervised models trained on unpaired sentences and single-pair supervised models trained on bitexts, both based on the unsupervised language model XLM-R with its parameters frozen. The experiments evaluate the models as universal sentence encoders on the task of unsupervised bitext mining on two datasets, where the unsupervised model reaches the state of the art of unsupervised retrieval, and the alternative single-pair supervised model approaches the performance of multilingually supervised models. The results suggest that bilingual training techniques as proposed can be applied to get sentence representations with multilingual alignment.

17:15-17:30 (The Liffey A)
Accurate Online Posterior Alignments for Principled Lexically-Constrained Decoding
Soumya Chatterjee, Sunita Sarawagi and Preethi Jyothi
Online alignment in machine translation refers to the task of aligning a target word to a source word when the target sequence has only been partially decoded. Good online alignments facilitate important applications such as lexically constrained translation where user-defined dictionaries are used to inject lexical constraints into the translation model. We propose a novel posterior alignment technique that is truly online in its execution and superior in terms of alignment error rates compared to existing methods. Our proposed inference technique jointly considers alignment and token probabilities in a principled manner and can be seamlessly integrated within existing constrained beam-search decoding algorithms. On five language pairs, including two distant language pairs, we achieve consistent drop in alignment error rates. When

deployed on seven lexically constrained translation tasks, we achieve significant improvements in BLEU specifically around the constrained positions.

17:30-17:45 (The Liffey A)
Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder Translation Models
Lorenzo Lupo, Marco Dinarelli and Laurent Besacier
Multi-encoder models are a broad family of context-aware neural machine translation systems that aim to improve translation quality by encoding document-level contextual information alongside the current sentence. The context encoding is undertaken by contextual parameters, trained on document-level data. In this work, we discuss the difficulty of training these parameters effectively, due to the sparsity of the words in need of context (i.e., the training signal), and their relevant context. We propose to pre-train the contextual parameters over split sentence pairs, which makes an efficient use of the available data for two reasons. Firstly, it increases the contextual training signal by breaking intra-sentential syntactic relations, and thus pushing the model to search the context for disambiguating clues more frequently. Secondly, it eases the retrieval of relevant context, since context segments become shorter. We propose four different splitting methods, and evaluate our approach with BLEU and contrastive test sets. Results show that it consistently improves learning of contextual parameters, both in low and high resource settings.

17:45-18:00 (The Liffey A)
Prediction Difference Regularization against Perturbation for Neural Machine Translation
Dengji Guo, Zhengrui Ma, Min Zhang and Yang Feng
Regularization methods applying input perturbation have drawn considerable attention and have been frequently explored for NMT tasks in recent years. Despite their simplicity and effectiveness, we argue that these methods are limited by the under-fitting of training data. In this paper, we utilize prediction difference for ground-truth tokens to analyze the fitting of token-level samples and find that under-fitting is almost as common as over-fitting. We introduce prediction difference regularization (PD-R), a simple and effective method that can reduce over-fitting and under-fitting at the same time. For all token-level samples, PD-R minimizes the prediction difference between the original pass and the input-perturbed pass, making the model less sensitive to small input changes, thus more robust to both perturbations and under-fitted training data. Experiments on three widely used WMT translation tasks show that our approach can significantly improve over existing perturbation regularization methods. On WMT16 En-De task, our model achieves 1.80 SacreBLEU improvement over vanilla transformer.

NLP Applications 1

17:00-18:00 (Wicklow Hall 2a)

17:00-17:15 (Wicklow Hall 2a)
Interpretability for Language Learners Using Example-Based Grammatical Error Correction
Masahiro Kaneko, Sho Takase, Ayana Niwa and Naoaki Okazaki
Grammatical Error Correction (GEC) should not focus only on high accuracy of corrections but also on interpretability for language learning. However, existing neural-based GEC models mainly aim at improving accuracy, and their interpretability has not been explored. A promising approach for improving interpretability is an example-based method, which uses similar retrieved examples to generate corrections. In addition, examples are beneficial in language learning, helping learners understand the basis of grammatically incorrect/correct texts and improve their confidence in writing. Therefore, we hypothesize that incorporating an example-based method into GEC can improve interpretability as well as support language learners. In this study, we introduce an Example-Based GEC (EB-GEC) that presents examples to language learners as a basis for a correction result. The examples consist of pairs of correct and incorrect sentences similar to a given input and its predicted correction. Experiments demonstrate that the examples presented by EB-GEC help language learners decide to accept or refuse suggestions from the GEC output. Furthermore, the experiments also show that retrieved examples improve the accuracy of corrections.

17:15-17:30 (Wicklow Hall 2a)
Ensembling and Knowledge Distilling of Large Sequence Taggers for Grammatical Error Correction
Maksym Tarnavskyi, Artem Chernodub and Kostiantyn Omelianchuk
In this paper, we investigate improvements to the GEC sequence tagging architecture with a focus on ensembling of recent cutting-edge Transformer-based encoders in Large configurations. We encourage ensembling models by majority votes on span-level edits because this approach is tolerant to the model architecture and vocabulary size. Our best ensemble achieves a new SOTA result with an $F_{0.5}$ score of 76.05 on BEA-2019 (test), even without pre-training on synthetic datasets. In addition, we perform knowledge distillation with a trained ensemble to generate new synthetic training datasets, "Troy-Blogs" and "Troy-1BW". Our best single sequence tagging model that is pretrained on the generated Troy- datasets in combination with the publicly available synthetic PIE dataset achieves a near-SOTA result with an $F_{0.5}$ score of 73.21 on BEA-2019 (test). The code, datasets, and trained models are publicly available.

17:30-17:45 (Wicklow Hall 2a)
Few-Shot Tabular Data Enrichment Using Fine-Tuned Transformer Architectures
Asaf Harari and Gilad Katz
The enrichment of tabular datasets using external sources has gained significant attention in recent years. Existing solutions, however, either ignore external unstructured data completely or devise dataset-specific solutions. In this study we proposed Few-Shot Transformer based Enrichment (FeSTE), a generic and robust framework for the enrichment of tabular datasets using unstructured data. By training over multiple datasets, our approach is able to develop generic models that can be applied to additional datasets with minimal training (i.e., few-shot). Our approach is based on an adaptation of BERT, for which we present a novel fine-tuning approach that reformulates the tuples of the datasets as sentences. Our evaluation, conducted on 17 datasets, shows that FeSTE is able to generate high quality features and significantly outperform existing fine-tuning solutions.

17:45-18:00 (Wicklow Hall 2a)
[CL] **Linguistic Parameters of Spontaneous Speech for identifying Mild Cognitive Impairment and Alzheimer's Disease**
Veronika Vincze, Martina Katalin Szabó, Ildikó Hoffmann, László Tóth, Magdolna Pákáski, János Kálmán and Gábor Gosztolya

Syntax: Tagging, Chunking and Parsing

17:00-18:00 (Wicklow Hall 2b)

17:00-17:15 (Wicklow Hall 2b)
Compositional Generalization in Dependency Parsing
Emily Goodwin, Siva Reddy, Timothy J. O'Donnell and Dzmitry Bahdanau
Compositionality— the ability to combine familiar units like words into novel phrases and sentences— has been the focus of intense interest in artificial intelligence in recent years. To test compositional generalization in semantic parsing, Keysers et al. (2020) introduced Compositional Freebase Queries (CFQ). This dataset maximizes the similarity between the test and train distributions over primitive units, like words, while maximizing the compound divergence: the dissimilarity between test and train distributions over larger structures, like phrases. Dependency parsing, however, lacks a compositional generalization benchmark. In this work, we introduce a gold-standard set of dependency parses for CFQ, and use this to analyze the behaviour of a state-of-the art dependency parser (Qi et al., 2020) on the CFQ dataset. We find that increasing compound divergence degrades dependency parsing performance, although not as dramatically as semantic parsing performance. Additionally, we find the performance of the dependency parser does not uniformly degrade relative to compound divergence, and the parser performs differently on different splits with the same compound divergence. We explore a number of hypotheses for what causes the non-uniform degradation in dependency parsing performance, and identify a number of syntactic structures that drive the dependency parser's lower performance on the most challenging splits.

17:15-17:30 (Wicklow Hall 2b)
Unsupervised Dependency Graph Network
Yikang Shen, Shawn Tan, Alessandro Sordoni, Peng Li, Jie Zhou and Aaron Courville
Recent work has identified properties of pretrained self-attention models that mirror those of dependency parse structures. In particular, some self-attention heads correspond well to individual dependency types. Inspired by these developments, we propose a new competitive mechanism that encourages these attention heads to model different dependency relations. We introduce a new model, the Unsupervised Dependency Graph Network (UDGN), that can induce dependency structures from raw corpora and the masked language modeling task. Experiment results show that UDGN achieves very strong unsupervised dependency parsing performance without gold POS tags and any other external information. The competitive gated heads show a strong correlation with human-annotated dependency types. Furthermore, the UDGN can also achieve competitive performance on masked language modeling and sentence textual similarity tasks.

17:30-17:45 (Wicklow Hall 2b)
Semantic Composition with PSHRG for Derivation Tree Reconstruction from Graph-Based Meaning Representations
Chun Hei Lo, Wai Lam and Hong Cheng
We introduce a data-driven approach to generating derivation trees from meaning representation graphs with probabilistic synchronous hyperedge replacement grammar (PSHRG). SHRG has been used to produce meaning representation graphs from texts and syntax trees, but little is known about its viability on the reverse. In particular, we experiment on Dependency Minimal Recursion Semantics (DMRS) and adapt PSHRG as a formalism that approximates the semantic composition of DMRS graphs and simultaneously recovers the derivations that license the DMRS graphs. Consistent results are obtained as evaluated on a collection of annotated corpora. This work reveals the ability of PSHRG in formalizing a syntax–semantics interface, modelling compositional graph-to-tree translations, and channelling explainability to surface realization.

17:45-18:00 (Wicklow Hall 2b)
Meta-Learning for Fast Cross-Lingual Adaptation in Dependency Parsing
Anna Langedijk, Verna Dankers, Phillip Lippe, Sander Bos, Bryan Cardenas Guevara, Helen Yannakoudakis and Ekaterina Shutova
Meta-learning, or learning to learn, is a technique that can help to overcome resource scarcity in cross-lingual NLP problems, by enabling fast adaptation to new tasks. We apply model-agnostic meta-learning (MAML) to the task of cross-lingual dependency parsing. We train our model on a diverse set of languages to learn a parameter initialization that can adapt quickly to new languages. We find that meta-learning with pre-training can significantly improve upon the performance of language transfer and standard supervised learning baselines for a variety of unseen, typologically diverse, and low-resource languages, in a few-shot learning setup.

Generation 1

17:00-18:00 (Wicklow Hall 1)

17:00-17:15 (Wicklow Hall 1)
Fine-Grained Controllable Text Generation Using Non-Residual Prompting
Fredrik Carlsson, Joey Öhman, Fangyu Liu, Severine Verlinden, Joakim Nivre and Magnus Sahlgren
The introduction of immensely large Causal Language Models (CLMs) has rejuvenated the interest in open-ended text generation. However, controlling the generative process for these Transformer-based models is at large an unsolved problem. Earlier work has explored either plug-and-play decoding strategies, or more powerful but blunt approaches such as prompting. There hence currently exists a trade-off between fine-grained control, and the capability for more expressive high-level instructions. To alleviate this trade-off, we propose an encoder-decoder architecture that enables intermediate text prompts at arbitrary time steps. We propose a resource-efficient method for converting a pre-trained CLM into this architecture, and demonstrate its potential on various experiments, including the novel task of contextualized word inclusion. Our method provides strong results on multiple experimental settings, proving itself to be both expressive and versatile.

17:15-17:30 (Wicklow Hall 1)
Tailor: Generating and Perturbing Text with Semantic Controls
Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E Peters and Matt Gardner
Controlled text perturbation is useful for evaluating and improving model generalizability. However, current techniques rely on training a

model for every target perturbation, which is expensive and hard to generalize. We present Tailor, a semantically-controlled text generation system. Tailor builds on a pretrained seq2seq model and produces textual outputs conditioned on control codes derived from semantic representations. We craft a set of operations to modify the control codes, which in turn steer generation towards targeted attributes. These operations can be further composed into higher-level ones, allowing for flexible perturbation strategies. We demonstrate the effectiveness of these perturbations in multiple applications. First, we use Tailor to automatically create high-quality contrast sets for four distinct natural language processing (NLP) tasks. These contrast sets contain fewer spurious artifacts and are complementary to manually annotated ones in their lexical diversity. Second, we show that Tailor perturbations can improve model generalization through data augmentation. Perturbing just ∼% of training data leads to a 5.8-point gain on an NLI challenge set measuring reliance on syntactic heuristics.

17:30-17:45 (Wicklow Hall 1)
Evaluating Factuality in Text Simplification
Ashwin Devaraj, William Berkeley Sheffield, Byron C Wallace and Junyi Jessy Li
Automated simplification models aim to make input texts more readable. Such methods have the potential to make complex information accessible to a wider audience, e.g., providing access to recent medical literature which might otherwise be impenetrable for a lay reader. However, such models risk introducing errors into automatically simplified texts, for instance by inserting statements unsupported by the corresponding original text, or by omitting key information. Providing more readable but inaccurate versions of texts may in many cases be worse than providing no such access at all. The problem of factual accuracy (and the lack thereof) has received heightened attention in the context of summarization models, but the factuality of automatically simplified texts has not been investigated. We introduce a taxonomy of errors that we use to analyze both references drawn from standard simplification datasets and state-of-the-art model outputs. We find that errors often appear in both that are not captured by existing evaluation metrics, motivating a need for research into ensuring the factual accuracy of automated simplification models.

17:45-18:00 (Wicklow Hall 1)
Improving Compositional Generalization with Self-Training for Data-to-Text Generation
Sanket Vaibhav Mehta, Jinfeng Rao, Yi Tay, Mihir Kale, Ankur P Parikh and Emma Strubell
Data-to-text generation focuses on generating fluent natural language responses from structured meaning representations (MRs). Such representations are compositional and it is costly to collect responses for all possible combinations of atomic meaning schemata, thereby necessitating few-shot generalization to novel MRs. In this work, we systematically study the compositional generalization of the state-of-the-art T5 models in few-shot data-to-text tasks. We show that T5 models fail to generalize to unseen MRs, and we propose a template-based input representation that considerably improves the model's generalization capability. To further improve the model's performance, we propose an approach based on self-training using fine-tuned BLEURT for pseudo-response selection. On the commonly-used SGD and Weather benchmarks, the proposed self-training approach improves tree accuracy by $46\%+$ and reduces the slot error rates by $73\%+$ over the strong T5 baselines in few-shot settings.

Interpretability and Analysis of Models for NLP 1

17:00-17:55 (Liffey Hall 2)

17:00-17:15 (Liffey Hall 2)
[TACL] **Evaluating Explanations: How Much do Explanations from the Teacher aid Students?**
Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Soares, Michael Collins, Zachary Lipton, Graham Neubig and William Cohen

17:15-17:30 (Liffey Hall 2)
Probing for the Usage of Grammatical Number
Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau and Ryan D Cotterell
A central quest of probing is to uncover how pre-trained models encode a linguistic property within their representations. An encoding, however, might be spurious-i.e., the model might not rely on it when making predictions. In this paper, we try to find an encoding that the model actually uses, introducing a usage-based probing setup. We first choose a behavioral task which cannot be solved without using the linguistic property. Then, we attempt to remove the property by intervening on the model's representations. We contend that, if an encoding is used by the model, its removal should harm the performance on the chosen behavioral task. As a case study, we focus on how BERT encodes grammatical number, and on how it uses this encoding to solve the number agreement task. Experimentally, we find that BERT relies on a linear encoding of grammatical number to produce the correct behavioral output. We also find that BERT uses a separate encoding of grammatical number for nouns and verbs. Finally, we identify in which layers information about grammatical number is transferred from a noun to its head verb.

17:30-17:45 (Liffey Hall 2)
The Grammar-Learning Trajectories of Neural Language Models
Leshem Choshen, Guy Hacohen, Daphna Weinshall and Omri Abend
The learning trajectories of linguistic phenomena in humans provide insight into linguistic representation, beyond what can be gleaned from inspecting the behavior of an adult speaker. To apply a similar approach to analyze neural language models (NLM), it is first necessary to establish that different models are similar enough in the generalizations they make. In this paper, we show that NLMs with different initialization, architecture, and training data acquire linguistic phenomena in a similar order, despite their different end performance. These findings suggest that there is some mutual inductive bias that underlies these models' learning of linguistic phenomena. Taking inspiration from psycholinguistics, we argue that studying this inductive bias is an opportunity to study the linguistic representation implicit in NLMs.
Leveraging these findings, we compare the relative performance on different phenomena at varying learning stages with simpler reference models. Results suggest that NLMs exhibit consistent "developmental" stages. Moreover, we find the learning trajectory to be approximately one-dimensional: given an NLM with a certain overall performance, it is possible to predict what linguistic generalizations it has already acquired. Initial analysis of these stages presents phenomena clusters (notably morphological ones), whose performance progresses in unison, suggesting a potential link between the generalizations behind them.

17:45-17:55 (Liffey Hall 2)
When classifying grammatical role, BERT doesn't care about word order... except when it matters
Isabel Papadimitriou, Richard Futrell and Kyle Mahowald
Because meaning can often be inferred from lexical semantics alone, word order is often a redundant cue in natural language. For example, the words chopped, chef, and onion are more likely used to convey "The chef chopped the onion," not "The onion chopped the chef." Recent work has shown large language models to be surprisingly word order invariant, but crucially has largely considered natural prototypical inputs, where compositional meaning mostly matches lexical expectations. To overcome this confound, we probe grammatical role representation in English BERT and GPT-2, on instances where lexical expectations are not sufficient, and word order knowledge is necessary for correct classification. Such non-prototypical instances are naturally occurring English sentences with inanimate subjects or animate objects, or sentences where we systematically swap the arguments to make sentences like "The onion chopped the chef". We find that, while early layer embeddings are largely lexical, word order is in fact crucial in defining the later-layer representations of words in semantically non-prototypical positions. Our experiments isolate the effect of word order on the contextualization process, and highlight how models use context in the uncommon, but critical, instances where it matters.

Poster Session 3: Question Answering

17:00-18:00 (Forum)

17:00-18:00 (Forum)
#1 Using Interactive Feedback to Improve the Accuracy and Explainability of Question Answering Systems Post-Deployment
Zichao Li, Jackie Cheung, Xing Han Lu, Siva Reddy and Prakhar Sharma
Most research on question answering focuses on the pre-deployment stage; i.e., building an accurate model for deployment. In this paper, we ask the question: Can we improve QA systems further post-deployment based on user interactions? We focus on two kinds of improvements: 1) improving the QA system's performance itself, and 2) providing the model with the ability to explain the correctness or incorrectness of an answer. We collect a retrieval-based QA dataset, FeedbackQA, which contains interactive feedback from users. We collect this dataset by deploying a base QA system to crowdworkers who then engage with the system and provide feedback on the quality of its answers. The feedback contains both structured ratings and unstructured natural language explanations. We train a neural model with this feedback data that can generate explanations and re-score answer candidates. We show that feedback data not only improves the accuracy of the deployed QA system but also other stronger non-deployed systems. The generated explanations also help users make informed decisions about the correctness of answers.[3]

17:00-18:00 (Forum)
#2 ConditionalQA: A Complex Reading Comprehension Dataset with Conditional Answers
Haitian Sun, William W. Cohen and Ruslan Salakhutdinov
We describe a Question Answering (QA) dataset that contains complex questions with conditional answers, i.e. the answers are only applicable when certain conditions apply. We call this dataset ConditionalQA. In addition to conditional answers, the dataset also features: (1) long context documents with information that is related in logically complex ways; (2) multi-hop questions that require compositional logical reasoning; (3) a combination of extractive questions, yes/no questions, questions with multiple answers, and not-answerable questions; (4) questions asked without knowing the answers. We show that ConditionalQA is challenging for many of the existing QA models, especially in selecting answer conditions. We believe that this dataset will motivate further research in answering complex questions over long documents.

17:00-18:00 (Forum)
#3 Simulating Bandit Learning from User Feedback for Extractive Question Answering
Ge Gao, Eunsol Choi and Yoav Artzi
We study learning from user feedback for extractive question answering by simulating feedback using supervised data. We cast the problem as contextual bandit learning, and analyze the characteristics of several learning scenarios with focus on reducing data annotation. We show that systems initially trained on few examples can dramatically improve given feedback from users on model-predicted answers, and that one can use existing datasets to deploy systems in new domains without any annotation effort, but instead improving the system on-the-fly via user feedback.

17:00-18:00 (Forum)
#4 Predicting Difficulty and Discrimination of Natural Language Questions
Matthew Alexander Byrd and Shashank Srivastava
Item Response Theory (IRT) has been extensively used to numerically characterize question difficulty and discrimination for human subjects in domains including cognitive psychology and education (Primi et al., 2014; Downing, 2003). More recently, IRT has been used to similarly characterize item difficulty and discrimination for natural language models across various datasets (Lalor et al., 2019; Vania et al., 2021; Rodriguez et al., 2021). In this work, we explore predictive models for directly estimating and explaining these traits for natural language questions in a question-answering context. We use HotpotQA for illustration. Our experiments show that it is possible to predict both difficulty and discrimination parameters for new questions, and these traits are correlated with features of questions, answers, and associated contexts. Our findings can have significant implications for the creation of new datasets and tests on the one hand and strategies such as active learning and curriculum learning on the other.

17:00-18:00 (Forum)
#5 Two-Step Question Retrieval for Open-Domain QA
Yeon Seonwoo, Juhee Son, Jiho Jin, Sang-Woo Lee, Ji-Hoon Kim, Jung-Woo Ha and Alice Oh
The retriever-reader pipeline has shown promising performance in open-domain QA but suffers from a very slow inference speed. Recently proposed question retrieval models tackle this problem by indexing question-answer pairs and searching for similar questions. These models have shown a significant increase in inference speed, but at the cost of lower QA performance compared to the retriever-reader models. This paper proposes a two-step question retrieval model, SQuID (Sequential Question-Indexed Dense retrieval) and distant supervision for training.

[3] Project page: https://mcgill-nlp.github.io/feedbackqa/

SQuID uses two bi-encoders for question retrieval. The first-step retriever selects top-k similar questions, and the second-step retriever finds the most similar question from the top-k questions. We evaluate the performance and the computational efficiency of SQuID. The results show that SQuID significantly increases the performance of existing question retrieval models with a negligible loss on inference speed.

17:00-18:00 (Forum)
#6 Answer Uncertainty and Unanswerability in Multiple-Choice Machine Reading Comprehension
Vatsal Raina and Mark Gales
Machine reading comprehension (MRC) has drawn a lot of attention as an approach for assessing the ability of systems to understand natural language. Usually systems focus on selecting the correct answer to a question given a contextual paragraph. However, for many applications of multiple-choice MRC systems there are two additional considerations. For multiple-choice exams there is often a negative marking scheme; there is a penalty for an incorrect answer. In terms of an MRC system this means that the system is required to have an idea of the uncertainty in the predicted answer. The second consideration is that many multiple-choice questions have the option of none-of-the-above (NOA) indicating that none of the answers is applicable, rather than there always being the correct answer in the list of choices. This paper investigates both of these issues by making use of predictive uncertainty. Whether the system should propose an answer is a direct application of answer uncertainty. There are two possibilities when considering the NOA option. The simplest is to explicitly build a system on data that includes this option. Alternatively uncertainty can be applied to detect whether the other options include the correct answer. If the system is not sufficiently confident it will select NOA. As there is no standard corpus available to investigate these topics, the ReClor corpus is modified by removing the correct answer from a subset of possible answers. A high-performance MRC system is used to evaluate whether answer uncertainty can be applied in these situations. It is shown that uncertainty does allow questions that the system is not confident about to be detected. Additionally it is shown that uncertainty outperforms a system explicitly built with an NOA option.

17:00-18:00 (Forum)
[TACL] **#7 Break, Perturb, Build: Automatic Perturbation of Reasoning Paths Through Question Decomposition**
Mor Geva, Tomer Wolfson and Jonathan Berant

17:00-18:00 (Forum)
[TACL] **#8 Time-Aware Language Models as Temporal Knowledge Bases**
Bhuwan Dhingra, Jeremy Cole, Julian Eisenschlos, Daniel Gillick, Jacob Eisenstein and William Cohen

17:00-18:00 (Forum)
#9 Answer-level Calibration for Free-form Multiple Choice Question Answering
Sawan Kumar
Pre-trained language models have recently shown that training on large corpora using the language modeling objective enables few-shot and zero-shot capabilities on a variety of NLP tasks, including commonsense reasoning tasks. This is achieved using text interactions with the model, usually by posing the task as a natural language text completion problem. While using language model probabilities to obtain task specific scores has been generally useful, it often requires task-specific heuristics such as length normalization, or probability calibration. In this work, we consider the question answering format, where we need to choose from a set of (free-form) textual choices of unspecified lengths given a context. We present ALC (Answer-Level Calibration), where our main suggestion is to model context-independent biases in terms of the probability of a choice without the associated context and to subsequently remove it using an unsupervised estimate of similarity with the full context. We show that our unsupervised answer-level calibration consistently improves over or is competitive with baselines using standard evaluation metrics on a variety of tasks including commonsense reasoning tasks. Further, we show that popular datasets potentially favor models biased towards easy cues which are available independent of the context. We analyze such biases using an associated F1-score. Our analysis indicates that answer-level calibration is able to remove such biases and leads to a more robust measure of model capability.

17:00-18:00 (Forum)
#10 Retrieval-guided Counterfactual Generation for QA
Bhargavi Paranjape, Matthew Lamm and Ian Tenney
Deep NLP models have been shown to be brittle to input perturbations. Recent work has shown that data augmentation using counterfactuals — i.e. minimally perturbed inputs — can help ameliorate this weakness. We focus on the task of creating counterfactuals for question answering, which presents unique challenges related to world knowledge, semantic diversity, and answerability. To address these challenges, we develop a Retrieve-Generate-Filter(RGF) technique to create counterfactual evaluation and training data with minimal human supervision. Using an open-domain QA framework and question generation model trained on original task data, we create counterfactuals that are fluent, semantically diverse, and automatically labeled. Data augmentation with RGF counterfactuals improves performance on out-of-domain and challenging evaluation sets over and above existing methods, in both the reading comprehension and open-domain QA settings. Moreover, we find that RGF data leads to significant improvements in a model's robustness to local perturbations.

17:00-18:00 (Forum)
#11 Sequence-to-Sequence Knowledge Graph Completion and Question Answering
Apoorv Umang Saxena, Adrian Kochsiek and Rainer Gemulla
Knowledge graph embedding (KGE) models represent each entity and relation of a knowledge graph (KG) with low-dimensional embedding vectors. These methods have recently been applied to KG link prediction and question answering over incomplete KGs (KGQA). KGEs typically create an embedding for each entity in the graph, which results in large model sizes on real-world graphs with millions of entities. For downstream tasks these atomic entity representations often need to be integrated into a multi stage pipeline, limiting their utility. We show that an off-the-shelf encoder-decoder Transformer model can serve as a scalable and versatile KGE model obtaining state-of-the-art results for KG link prediction and incomplete KG question answering. We achieve this by posing KG link prediction as a sequence-to-sequence task and exchange the triple scoring approach taken by prior KGE methods with autoregressive decoding. Such a simple but powerful method reduces the model size up to 98

17:00-18:00 (Forum)
#12 Automated Crossword Solving
Eric Wallace, Nicholas Tomlin, Albert Xu, Kevin Yang, Eshaan Pathak, Matthew L. Ginsberg and Dan Klein

We present the Berkeley Crossword Solver, a state-of-the-art approach for automatically solving crossword puzzles. Our system works by generating answer candidates for each crossword clue using neural question answering models and then combines loopy belief propagation with local search to find full puzzle solutions. Compared to existing approaches, our system improves exact puzzle accuracy from 57

17:00-18:00 (Forum)
#13 KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering
Donghan Yu, Chenguang Zhu, Yuwei Fang, Wenhao Yu, Shuohang Wang, Yichong Xu, Xiang Ren, Yiming Yang and Michael Zeng
Current Open-Domain Question Answering (ODQA) models typically include a retrieving module and a reading module, where the retriever selects potentially relevant passages from open-source documents for a given question, and the reader produces an answer based on the retrieved passages. The recently proposed Fusion-in-Decoder (FiD) framework is a representative example, which is built on top of a dense passage retriever and a generative reader, achieving the state-of-the-art performance. In this paper we further improve the FiD approach by introducing a knowledge-enhanced version, namely KG-FiD. Our new model uses a knowledge graph to establish the structural relationship among the retrieved passages, and a graph neural network (GNN) to re-rank the passages and select only a top few for further processing. Our experiments on common ODQA benchmark datasets (Natural Questions and TriviaQA) demonstrate that KG-FiD can achieve comparable or better performance in answer prediction than FiD, with less than 40

17:00-18:00 (Forum)
#14 QAConv: Question Answering on Informative Conversations
Chien-Sheng Wu, Andrea Madotto, Wenhao Liu, Pascale Fung and Caiming Xiong
This paper introduces QAConv, a new question answering (QA) dataset that uses conversations as a knowledge source. We focus on informative conversations, including business emails, panel discussions, and work channels. Unlike open-domain and task-oriented dialogues, these conversations are usually long, complex, asynchronous, and involve strong domain knowledge. In total, we collect 34,608 QA pairs from 10,259 selected conversations with both human-written and machine-generated questions. We use a question generator and a dialogue summarizer as auxiliary tools to collect and recommend questions. The dataset has two testing scenarios: chunk mode and full mode, depending on whether the grounded partial conversation is provided or retrieved. Experimental results show that state-of-the-art pretrained QA systems have limited zero-shot performance and tend to predict our questions as unanswerable. Our dataset provides a new training and evaluation testbed to facilitate QA on conversations research.

17:00-18:00 (Forum)
#15 Turning Tables: Generating Examples from Semi-structured Tables for Endowing Language Models with Reasoning Skills
Ori Yoran, Alon Talmor and Jonathan Berant
Models pre-trained with a language modeling objective possess ample world knowledge and language skills, but are known to struggle in tasks that require reasoning. In this work, we propose to leverage semi-structured tables, and automatically generate at scale question-paragraph pairs, where answering the question requires reasoning over multiple facts in the paragraph. We add a pre-training step over this synthetic data, which includes examples that require 16 different reasoning skills such as number comparison, conjunction, and fact composition. To improve data efficiency, we sample examples from reasoning skills where the model currently errs. We evaluate our approach on three reasoning-focused reading comprehension datasets, and show that our model, PReasM, substantially outperforms T5, a popular pre-trained encoder-decoder model. Moreover, sampling examples based on model errors leads to faster training and higher performance.

17:00-18:00 (Forum)
#16 RNG-KBQA: Generation Augmented Iterative Ranking for Knowledge Base Question Answering
Xi Ye, Semih Yavuz, Kazuma Hashimoto, Yingbo Zhou and Caiming Xiong
Existing KBQA approaches, despite achieving strong performance on i.i.d. test data, often struggle in generalizing to questions involving unseen KB schema items. Prior ranking-based approaches have shown some success in generalization, but suffer from the coverage issue. We present RnG-KBQA, a Rank-and-Generate approach for KBQA, which remedies the coverage issue with a generation model while preserving a strong generalization capability. Our approach first uses a contrastive ranker to rank a set of candidate logical forms obtained by searching over the knowledge graph. It then introduces a tailored generation model conditioned on the question and the top-ranked candidates to compose the final logical form. We achieve new state-of-the-art results on GrailQA and WebQSP datasets. In particular, our method surpasses the prior state-of-the-art by a large margin on the GrailQA leaderboard. In addition, RnG-KBQA outperforms all prior approaches on the popular WebQSP benchmark, even including the ones that use the oracle entity linking. The experimental results demonstrate the effectiveness of the interplay between ranking and generation, which leads to the superior performance of our proposed approach across all settings with especially strong improvements in zero-shot generalization.

17:00-18:00 (Forum)
#17 Open Domain Question Answering with A Unified Knowledge Interface
Kaixin Ma, Hao Cheng, Xiaodong Liu, Eric Nyberg and Jianfeng Gao
The retriever-reader framework is popular for open-domain question answering (ODQA) due to its ability to use explicit knowledge. Although prior work has sought to increase the knowledge coverage by incorporating structured knowledge beyond text, accessing heterogeneous knowledge sources through a unified interface remains an open question. While data-to-text generation has the potential to serve as a universal interface for data and text, its feasibility for downstream tasks remains largely unknown. In this work, we bridge this gap and use the data-to-text method as a means for encoding structured knowledge for open-domain question answering. Specifically, we propose a verbalizer-retriever-reader framework for ODQA over data and text where verbalized tables from Wikipedia and graphs from Wikidata are used as augmented knowledge sources. We show that our Unified Data and Text QA, UDT-QA, can effectively benefit from the expanded knowledge index, leading to large gains over text-only baselines. Notably, our approach sets the single-model state-of-the-art on Natural Questions. Furthermore, our analyses indicate that verbalized knowledge is preferred for answer reasoning for both adapted and hot-swap settings.

17:00-18:00 (Forum)
#18 Hypergraph Transformer: Weakly-Supervised Multi-hop Reasoning for Knowledge-based Visual Question Answering
Yu-Jung Heo, Eun-Sol Kim, Woo Suk Choi and Byoung-Tak Zhang
Knowledge-based visual question answering (QA) aims to answer a question which requires visually-grounded external knowledge beyond image content itself. Answering complex questions that require multi-hop reasoning under weak supervision is considered as a challenging problem since i) no supervision is given to the reasoning process and ii) high-order semantics of multi-hop knowledge facts need to be captured. In this paper, we introduce a concept of hypergraph to encode high-level semantics of a question and a knowledge base, and to learn high-order associations between them. The proposed model, Hypergraph Transformer, constructs a question hypergraph and a query-aware

knowledge hypergraph, and infers an answer by encoding inter-associations between two hypergraphs and intra-associations in both hypergraph itself. Extensive experiments on two knowledge-based visual QA and two knowledge-based textual QA demonstrate the effectiveness of our method, especially for multi-hop reasoning problem. Our source code is available at https://github.com/yujungheo/kbvqa-public.

17:00-18:00 (Forum)
#19 Generated Knowledge Prompting for Commonsense Reasoning
Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi and Hannaneh Hajishirzi
It remains an open question whether incorporating external knowledge benefits commonsense reasoning while maintaining the flexibility of pretrained sequence models. To investigate this question, we develop generated knowledge prompting, which consists of generating knowledge from a language model, then providing the knowledge as additional input when answering a question. Our method does not require task-specific supervision for knowledge integration, or access to a structured knowledge base, yet it improves performance of large-scale, state-of-the-art models on four commonsense reasoning tasks, achieving state-of-the-art results on numerical commonsense (NumerSense), general commonsense (CommonsenseQA 2.0), and scientific commonsense (QASC) benchmarks. Generated knowledge prompting highlights large-scale language models as flexible sources of external knowledge for improving commonsense reasoning. Our code is available at `github.com/liujch1998/GKP`

17:00-18:00 (Forum)
#20 Calibration of Machine Reading Systems at Scale
Shehzaad Zuzar Dhuliawala, Leonard Adolphs, Rajarshi Das and Mrinmaya Sachan
In typical machine learning systems, an estimate of the probability of the prediction is used to assess the system's confidence in the prediction. This confidence measure is usually uncalibrated; i.e. the system's confidence in the prediction does not match the true probability of the predicted output. In this paper, we present an investigation into calibrating open setting machine reading systems such as open-domain question answering and claim verification systems. We show that calibrating such complex systems which contain discrete retrieval and deep reading components is challenging and current calibration techniques fail to scale to these settings. We propose simple extensions to existing calibration approaches that allows us to adapt them to these settings. Our experimental results reveal that the approach works well, and can be useful to selectively predict answers when question answering systems are posed with unanswerable or out-of-the-training distribution questions.

17:00-18:00 (Forum)
#21 BBQ: A hand-built bias benchmark for question answering
Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut and Samuel R. Bowman
It is well documented that NLP models learn social biases, but little work has been done on how these biases manifest in model outputs for applied tasks like question answering (QA). We introduce the Bias Benchmark for QA (BBQ), a dataset of question-sets constructed by the authors that highlight attested social biases against people belonging to protected classes along nine social dimensions relevant for U.S. English-speaking contexts. Our task evaluate model responses at two levels: (i) given an under-informative context, we test how strongly responses reflect social biases, and (ii) given an adequately informative context, we test whether the model's biases override a correct answer choice. We find that models often rely on stereotypes when the context is under-informative, meaning the model's outputs consistently reproduce harmful biases in this setting. Though models are more accurate when the context provides an informative answer, they still rely on stereotypes and average up to 3.4 percentage points higher accuracy when the correct answer aligns with a social bias than when it conflicts, with this difference widening to over 5 points on examples targeting gender for most models tested.

17:00-18:00 (Forum)
#22 Read before Generate! Faithful Long Form Question Answering with Machine Reading
Dan SU, Xiaoguang Li, Jindi Zhang, Lifeng Shang, Xin Jiang, Qun Liu and Pascale Fung
Long-form question answering (LFQA) aims to generate a paragraph-length answer for a given question. While current work on LFQA using large pre-trained model for generation are effective at producing fluent and somewhat relevant content, one primary challenge lies in how to generate a faithful answer that has less hallucinated content. We propose a new end-to-end framework that jointly models answer generation and machine reading. The key idea is to augment the generation model with fine-grained, answer-related salient information which can be viewed as an emphasis on faithful facts. State-of-the-art results on two LFQA datasets, ELI5 and MS MARCO, demonstrate the effectiveness of our method, in comparison with strong baselines on automatic and human evaluation metrics. A detailed analysis further proves the competency of our methods in generating fluent, relevant, and more faithful answers.

17:00-18:00 (Forum)
#23 Investigating Selective Prediction Approaches Across Several Tasks in IID, OOD, and Adversarial Settings
Neeraj Varshney, Swaroop Mishra and Chitta Baral
In order to equip NLP systems with 'selective prediction' capability, several task-specific approaches have been proposed. However, which approaches work best across tasks or even if they consistently outperform the simplest baseline MaxProb remains to be explored. To this end, we systematically study selective prediction in a large-scale setup of 17 datasets across several NLP tasks. Through comprehensive experiments under in-domain (IID), out-of-domain (OOD), and adversarial (ADV) settings, we show that despite leveraging additional resources (held-out data/computation), none of the existing approaches consistently and considerably outperforms MaxProb in all three settings. Furthermore, their performance does not translate well across tasks. For instance, Monte-Carlo Dropout outperforms all other approaches on Duplicate Detection datasets but does not fare well on NLI datasets, especially in the OOD setting. Thus, we recommend that future selective prediction approaches should be evaluated across tasks and settings for reliable estimation of their capabilities.

17:00-18:00 (Forum)
#24 Fact-Tree Reasoning for N-ary Question Answering over Knowledge Graphs
Yao Zhang, Peiyao Li, Hongru Liang, Adam Jatowt and Zhenglu Yang
Current Question Answering over Knowledge Graphs (KGQA) task mainly focuses on performing answer reasoning upon KGs with binary facts. However, it neglects the n-ary facts, which contain more than two entities. In this work, we highlight a more challenging but underexplored task: n-ary KGQA, i.e., answering n-ary facts questions upon n-ary KGs. Nevertheless, the multi-hop reasoning framework popular in binary KGQA task is not directly applicable on n-ary KGQA. We propose two feasible improvements: 1) upgrade the basic reasoning unit from entity or relation to fact, and 2) upgrade the reasoning structure from chain to tree. Therefore, we propose a novel fact-tree reasoning framework, FacTree, which integrates the above two upgrades. FacTree transforms the question into a fact tree and performs iterative fact

reasoning on the fact tree to infer the correct answer. Experimental results on the n-ary KGQA dataset we constructed and two binary KGQA benchmarks demonstrate the effectiveness of FacTree compared with state-of-the-art methods.

17:00-18:00 (Forum)
#25 Ditch the Gold Standard: Re-evaluating Conversational Question Answering
Huihan Li, Tianyu Gao, Manan Goenka and Danqi Chen
Conversational question answering aims to provide natural-language answers to users in information-seeking conversations. Existing conversational QA benchmarks compare models with pre-collected human-human conversations, using ground-truth answers provided in conversational history. It remains unclear whether we can rely on this static evaluation for model development and whether current systems can well generalize to real-world human-machine conversations. In this work, we conduct the first large-scale human evaluation of state-of-the-art conversational QA systems, where human evaluators converse with models and judge the correctness of their answers. We find that the distribution of human machine conversations differs drastically from that of human-human conversations, and there is a disagreement between human and gold-history evaluation in terms of model ranking. We further investigate how to improve automatic evaluations, and propose a question rewriting mechanism based on predicted history, which better correlates with human judgments. Finally, we analyze the impact of various modeling strategies and discuss future directions towards building better conversational question answering systems.

17:00-18:00 (Forum)
#26 Hey AI, Can You Solve Complex Tasks by Talking to Agents?
Tushar Khot, Kyle Richardson, Daniel Khashabi and Ashish Sabharwal
Training giant models from scratch for each complex task is resource- and data-inefficient. To help develop models that can leverage existing systems, we propose a new challenge: Learning to solve complex tasks by communicating with existing agents (or models) in natural language. We design a synthetic benchmark, CommaQA, with three complex reasoning tasks (explicit, implicit, numeric) designed to be solved by communicating with existing QA agents. For instance, using text and table QA agents to answer questions such as "Who had the longest javelin throw from USA?". We show that black-box models struggle to learn this task from scratch (accuracy under 50%) even with access to each agent's knowledge and gold facts supervision. In contrast, models that learn to communicate with agents outperform black-box models, reaching scores of 100% when given gold decomposition supervision. However, we show that the challenge of learning to solve complex tasks by communicating with existing agents *without relying on any auxiliary supervision or data* still remains highly elusive. We will release CommaQA, along with a compositional generalization test split, to advance research in this direction.

17:00-18:00 (Forum)
#27 Relevant CommonSense Subgraphs for "What if..." Procedural Reasoning
Chen Zheng and Parisa Kordjamshidi
We study the challenge of learning causal reasoning over procedural text to answer "What if..." questions when external commonsense knowledge is required. We propose a novel multi-hop graph reasoning model to 1) efficiently extract a commonsense subgraph with the most relevant information from a large knowledge graph; 2) predict the causal answer by reasoning over the representations obtained from the commonsense subgraph and the contextual interactions between the questions and context. We evaluate our model on WIQA benchmark and achieve state-of-the-art performance compared to the recent models.

17:00-18:00 (Forum)
[DEMO] **UKP-SQUARE: An Online Platform for Question Answering Research**
Tim Baumgärtner, Kexin Wang, Rachneet Sachdeva, Gregor Geigle, Max Eichler, Clifton Poth, Hannah Sterz, Haritz Puerto, Leonardo F. R. Ribeiro, Jonas Pfeiffer, Nils Reimers, Gözde Şahin and Iryna Gurevych
Recent advances in NLP and information retrieval have given rise to a diverse set of question answering tasks that are of different formats (e.g., extractive, abstractive), require different model architectures (e.g., generative, discriminative), and setups (e.g., with or without retrieval). Despite having a large number of powerful, specialized QA pipelines (which we refer to as Skills) that consider a single domain, model or setup, there exists no framework where users can easily explore and compare such pipelines and can extend them according to their needs. To address this issue, we present UKP-SQuARE, an extensible online QA platform for researchers which allows users to query and analyze a large collection of modern Skills via a user-friendly web interface and integrated behavioural tests. In addition, QA researchers can develop, manage, and share their custom Skills using our microservices that support a wide range of models (Transformers, Adapters, ONNX), datastores and retrieval techniques (e.g., sparse and dense). UKP-SQuARE is available on https://square.ukp-lab.de

Poster Session 3: Computational Social Science and Cultural Analytics

17:00-18:00 (Forum)

17:00-18:00 (Forum)
#28 Improving the Generalizability of Depression Detection by Leveraging Clinical Questionnaires
Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet and Arman Cohan
Automated methods have been widely used to identify and analyze mental health conditions (e.g., depression) from various sources of information, including social media. Yet, deployment of such models in real-world healthcare applications faces challenges including poor out-of-domain generalization and lack of trust in black box models. In this work, we propose approaches for depression detection that are constrained to different degrees by the presence of symptoms described in PHQ9, a questionnaire used by clinicians in the depression screening process. In dataset-transfer experiments on three social media datasets, we find that grounding the model in PHQ9's symptoms substantially improves its ability to generalize to out-of-distribution data compared to a standard BERT-based approach. Furthermore, this approach can still perform competitively on in-domain data. These results and our qualitative analyses suggest that grounding model predictions in clinically-relevant symptoms can improve generalizability while producing a model that is easier to inspect.

17:00-18:00 (Forum)
#29 Automatic Identification and Classification of Bragging in Social Media
Mali Jin, Daniel Preotiuc-Pietro, A. Seza Doğruöz and Nikolaos Aletras
Bragging is a speech act employed with the goal of constructing a favorable self-image through positive statements about oneself. It is

widespread in daily communication and especially popular in social media, where users aim to build a positive image of their persona directly or indirectly. In this paper, we present the first large scale study of bragging in computational linguistics, building on previous research in linguistics and pragmatics. To facilitate this, we introduce a new publicly available data set of tweets annotated for bragging and their types. We empirically evaluate different transformer-based models injected with linguistic information in (a) binary bragging classification, i.e., if tweets contain bragging statements or not; and (b) multi-class bragging type prediction including not bragging. Our results show that our models can predict bragging with macro F1 up to 72.42 and 35.95 in the binary and multi-class classification tasks respectively. Finally, we present an extensive linguistic and error analysis of bragging prediction to guide future research on this topic.

17:00-18:00 (Forum)
#30 Classification without (Proper) Representation: Political Heterogeneity in Social Media and Its Implications for Classification and Behavioral Analysis
Kenan Alkiek, Bohan Zhang and David Jurgens
Reddit is home to a broad spectrum of political activity, and users signal their political affiliations in multiple ways—from self-declarations to community participation. Frequently, computational studies have treated political users as a single bloc, both in developing models to infer political leaning and in studying political behavior. Here, we test this assumption of political users and show that commonly-used political-inference models do not generalize, indicating heterogeneous types of political users. The models remain imprecise at best for most users, regardless of which sources of data or methods are used. Across a 14-year longitudinal analysis, we demonstrate that the choice in definition of a political user has significant implications for behavioral analysis. Controlling for multiple factors, political users are more toxic on the platform and inter-party interactions are even more toxic—but not all political users behave this way. Last, we identify a subset of political users who repeatedly flip affiliations, showing that these users are the most controversial of all, acting as provocateurs by more frequently bringing up politics, and are more likely to be banned, suspended, or deleted.

17:00-18:00 (Forum)
#31 Suum Cuique: Studying Bias in Taboo Detection with a Community Perspective
Osama Khalid, Jonathan Rusert and Padmini Srinivasan
Prior research has discussed and illustrated the need to consider linguistic norms at the community level when studying taboo (hateful/offensive/toxic etc.) language. However, a methodology for doing so, that is firmly founded on community language norms is still largely absent. This can lead both to biases in taboo text classification and limitations in our understanding of the causes of bias. We propose a method to study bias in taboo classification and annotation where a community perspective is front and center. This is accomplished by using special classifiers tuned for each community's language. In essence, these classifiers represent community level language norms. We use these to study bias and find, for example, biases are largest against African Americans (7/10 datasets and all 3 classifiers examined). In contrast to previous papers we also study other communities and find, for example, strong biases against South Asians. In a small scale user study we illustrate our key idea which is that common utterances, i.e., those with high alignment scores with a community (community classifier confidence scores) are unlikely to be regarded taboo. Annotators who are community members contradict taboo classification decisions and annotations in a majority of instances. This paper is a significant step toward reducing false positive taboo decisions that over time harm minority communities.

17:00-18:00 (Forum)
#32 Should a Chatbot be Sarcastic? Understanding User Preferences Towards Sarcasm Generation
Silviu Vlad Oprea, Steven R. Wilson and Walid Magdy
Previous sarcasm generation research has focused on how to generate text that people perceive as sarcastic to create more human-like interactions. In this paper, we argue that we should first turn our attention to the question of when sarcasm should be generated, finding that humans consider sarcastic responses inappropriate to many input utterances. Next, we use a theory-driven framework for generating sarcastic responses, which allows us to control the linguistic devices included during generation. For each device, we investigate how much humans associate it with sarcasm, finding that pragmatic insincerity and emotional markers are devices crucial for making sarcasm recognisable.

17:00-18:00 (Forum)
#33 Reports of personal experiences and stories in argumentation: datasets and analysis
Neele Falk and Gabriella Lapesa
Reports of personal experiences or stories can play a crucial role in argumentation, as they represent an immediate and (often) relatable way to back up one's position with respect to a given topic. They are easy to understand and increase empathy: this makes them powerful in argumentation. The impact of personal reports and stories in argumentation has been studied in the Social Sciences, but it is still largely underexplored in NLP. Our work is the first step towards filling this gap: our goal is to develop robust classifiers to identify documents containing personal experiences and reports. The main challenge is the scarcity of annotated data: our solution is to leverage existing annotations to be able to scale-up the analysis. Our contribution is two-fold. First, we conduct a set of in-domain and cross-domain experiments involving three datasets (two from Argument Mining, one from the Social Sciences), modeling architectures, training setups and fine-tuning options tailored to the involved domains. We show that despite the differences among datasets and annotations, robust cross-domain classification is possible. Second, we employ linear regression for performance mining, identifying performance trends both for overall classification performance and individual classifier predictions.

17:00-18:00 (Forum)
#34 Inducing Positive Perspectives with Text Reframing
Caleb Ziems, Minzhi Li, Anthony Zhang and Diyi Yang
Sentiment transfer is one popular example of a text style transfer task, where the goal is to reverse the sentiment polarity of a text. With a sentiment reversal comes also a reversal in meaning. We introduce a different but related task called positive reframing in which we neutralize a negative point of view and generate a more positive perspective for the author without contradicting the original meaning. Our insistence on meaning preservation makes positive reframing a challenging and semantically rich task. To facilitate rapid progress, we introduce a large-scale benchmark, Positive Psychology Frames, with 8,349 sentence pairs and 12,755 structured annotations to explain positive reframing in terms of six theoretically-motivated reframing strategies. Then we evaluate a set of state-of-the-art text style transfer models, and conclude by discussing key challenges and directions for future work.

17:00-18:00 (Forum)
#35 Leveraging Wikipedia article evolution for promotional tone detection

Christine De Kock and Andreas Vlachos
Detecting biased language is useful for a variety of applications, such as identifying hyperpartisan news sources or flagging one-sided rhetoric. In this work we introduce WikiEvolve, a dataset for document-level promotional tone detection. Unlike previously proposed datasets, WikiEvolve contains seven versions of the same article from Wikipedia, from different points in its revision history; one with promotional tone, and six without it. This allows for obtaining more precise training signal for learning models from promotional tone detection. We adapt the previously proposed gradient reversal layer framework to encode two article versions simultaneously and thus leverage this additional training signal. In our experiments, our proposed adaptation of gradient reversal improves the accuracy of four different architectures on both in-domain and out-of-domain evaluation.

17:00-18:00 (Forum)
#36 Tackling Fake News Detection by Continually Improving Social Context Representations using Graph Neural Networks
Nikhil Mehta, Maria Leonor Pacheco and Dan Goldwasser
Easy access, variety of content, and fast widespread interactions are some of the reasons making social media increasingly popular. However, this rise has also enabled the propagation of fake news, text published by news sources with an intent to spread misinformation and sway beliefs. Detecting it is an important and challenging problem to prevent large scale misinformation and maintain a healthy society.
We view fake news detection as reasoning over the relations between sources, articles they publish, and engaging users on social media in a graph framework. After embedding this information, we formulate inference operators which augment the graph edges by revealing unobserved interactions between its elements, such as similarity between documents' contents and users' engagement patterns. Our experiments over two challenging fake news detection tasks show that using inference operators leads to a better understanding of the social media framework enabling fake news spread, resulting in improved performance.

17:00-18:00 (Forum)
#37 Misinfo Reaction Frames: Reasoning about Readers' Reactions to News Headlines
Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi and Yejin Choi
Even to a simple and short news headline, readers react in a multitude of ways: cognitively (e.g. inferring the writer's intent), emotionally (e.g. feeling distrust), and behaviorally (e.g. sharing the news with their friends). Such reactions are instantaneous and yet complex, as they rely on factors that go beyond interpreting factual content of news.
We propose Misinfo Reaction Frames (MRF), a pragmatic formalism for modeling how readers might react to a news headline. In contrast to categorical schema, our free-text dimensions provide a more nuanced way of understanding intent beyond being benign or malicious. We also introduce a Misinfo Reaction Frames corpus, a crowdsourced dataset of reactions to over 25k news headlines focusing on global crises: the Covid-19 pandemic, climate change, and cancer.
Empirical results confirm that it is indeed possible for neural models to predict the prominent patterns of readers' reactions to previously unseen news headlines. Additionally, our user study shows that displaying machine-generated MRF implications alongside news headlines to readers can increase their trust in real news while decreasing their trust in misinformation. Our work demonstrates the feasibility and importance of pragmatic inferences on news headlines to help enhance AI-guided misinformation detection and mitigation.

17:00-18:00 (Forum)
#38 Dynamically Refined Regularization for Improving Cross-corpora Hate Speech Detection
Tulika Bose, Nikolaos Aletras, Irina Illina and Dominique Fohr
Hate speech classifiers exhibit substantial performance degradation when evaluated on datasets different from the source. This is due to learning spurious correlations between words that are not necessarily relevant to hateful language, and hate speech labels from the training corpus. Previous work has attempted to mitigate this problem by regularizing specific terms from pre-defined static dictionaries. While this has been demonstrated to improve the generalizability of classifiers, the coverage of such methods is limited and the dictionaries require regular manual updates from human experts. In this paper, we propose to automatically identify and reduce spurious correlations using attribution methods with dynamic refinement of the list of terms that need to be regularized during training. Our approach is flexible and improves the cross-corpora performance over previous work independently and in combination with pre-defined dictionaries.

17:00-18:00 (Forum)
#39 Human Language Modeling
Nikita Soni, Matthew Matero, Niranjan Balasubramanian and H. Schwartz
Natural language is generated by people, yet traditional language modeling views words or documents as if generated independently. Here, we propose human language modeling (HuLM), a hierarchical extension to the language modeling problem where by a human- level exists to connect sequences of documents (e.g. social media messages) and capture the notion that human language is moderated by changing human states. We introduce, HaRT, a large-scale transformer model for solving HuLM, pre-trained on approximately 100,000 social media users, and demonstrate it's effectiveness in terms of both language modeling (perplexity) for social media and fine-tuning for 4 downstream tasks spanning document- and user-levels. Results on all tasks meet or surpass the current state-of-the-art.

17:00-18:00 (Forum)
#40 Listening to Affected Communities to Define Extreme Speech: Dataset and Experiments
Antonis Maronikolakis, Axel Wisiorek, Leah Nann, Haris Jabbar, Sahana Udupa and Hinrich Schuetze
Building on current work on multilingual hate speech (e.g., Ousidhoum et al. (2019)) and hate speech reduction (e.g., Sap et al. (2020)), we present XTREMESPEECH, a new hate speech dataset containing 20,297 social media passages from Brazil, Germany, India and Kenya. The key novelty is that we directly involve the affected communities in collecting and annotating the data – as opposed to giving companies and governments control over defining and combatting hate speech. This inclusive approach results in datasets more representative of actually occurring online speech and is likely to facilitate the removal of the social media content that marginalized communities view as causing the most harm. Based on XTREMESPEECH, we establish novel tasks with accompanying baselines, provide evidence that cross-country training is generally not feasible due to cultural differences between countries and perform an interpretability analysis of BERT's predictions.

17:00-18:00 (Forum)
#41 Good Night at 4 pm?! Time Expressions in Different Cultures
Vered Shwartz
We propose the task of culture-specific time expression grounding, i.e. mapping from expressions such as "morning" in English or "Manhã" in

Portuguese to specific hours in the day. We propose 3 language-agnostic methods, one of which achieves promising results on gold standard annotations that we collected for a small number of languages. We then apply this method to 27 languages and analyze the similarities across languages in the grounding of time expressions.

17:00-18:00 (Forum)
#42 Measuring the Language of Self-Disclosure across Corpora
Ann-Katrin Reuel, Sebastian Peralta, João Sedoc, Garrick Sherman and Lyle Ungar
Being able to reliably estimate self-disclosure – a key component of friendship and intimacy – from language is important for many psychology studies. We build single-task models on five self-disclosure corpora, but find that these models generalize poorly; the within-domain accuracy of predicted message-level self-disclosure of the best-performing model (mean Pearson's r=0.69) is much higher than the respective across data set accuracy (mean Pearson's r=0.32), due to both variations in the corpora (e.g., medical vs. general topics) and labeling instructions (target variables: self-disclosure, emotional disclosure, intimacy). However, some lexical features, such as expression of negative emotions and use of first person personal pronouns such as 'I' reliably predict self-disclosure across corpora. We develop a multi-task model that yields better results, with an average Pearson's r of 0.37 for out-of-corpora prediction.

Poster Session 3: Ethics in NLP

17:00-18:00 (Forum)

17:00-18:00 (Forum)
#43 An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models
Nicholas Meade, Elinor Poole-Dayan and Siva Reddy
Recent work has shown pre-trained language models capture social biases from the large amounts of text they are trained on. This has attracted attention to developing techniques that mitigate such biases. In this work, we perform an empirical survey of five recently proposed bias mitigation techniques: Counterfactual Data Augmentation (CDA), Dropout, Iterative Nullspace Projection, Self-Debias, and SentenceDebias. We quantify the effectiveness of each technique using three intrinsic bias benchmarks while also measuring the impact of these techniques on a model's language modeling ability, as well as its performance on downstream NLU tasks. We experimentally find that: (1) Self-Debias is the strongest debiasing technique, obtaining improved scores on all bias benchmarks; (2) Current debiasing techniques perform less consistently when mitigating non-gender biases; And (3) improvements on bias benchmarks such as StereoSet and CrowS-Pairs by using debiasing strategies are often accompanied by a decrease in language modeling ability, making it difficult to determine whether the bias mitigation was effective.

17:00-18:00 (Forum)
#44 French CrowS-Pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English
Aurélie Névéol, Yoann Dupont, Julien Bezançon and Karën Fort
Warning: This paper contains explicit statements of offensive stereotypes which may be upsetting. Much work on biases in natural language processing has addressed biases linked to the social and cultural experience of English speaking individuals in the United States. We seek to widen the scope of bias studies by creating material to measure social bias in language models (LMs) against specific demographic groups in France. We build on the US-centered CrowS-pairs dataset to create a multilingual stereotypes dataset that allows for comparability across languages while also characterizing biases that are specific to each country and language. We introduce 1,679 sentence pairs in French that cover stereotypes in ten types of bias like gender and age. 1,467 sentence pairs are translated from CrowS-pairs and 212 are newly crowdsourced. The sentence pairs contrast stereotypes concerning underadvantaged groups with the same sentence concerning advantaged groups. We find that four widely used language models (three French, one multilingual) favor sentences that express stereotypes in most bias categories. We report on the translation process from English into French, which led to a characterization of stereotypes in CrowS-pairs including the identification of US-centric cultural traits. We offer guidelines to further extend the dataset to other languages and cultural environments.

17:00-18:00 (Forum)
#45 Your fairness may vary: Pretrained language model fairness in toxic text classification
Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Moninder Singh and Mikhail Yurochkin
The popularity of pretrained language models in natural language processing systems calls for a careful evaluation of such models in downstream tasks, which have a higher potential for societal impact. The evaluation of such systems usually focuses on accuracy measures. Our findings in this paper call for attention to be paid to fairness measures as well. Through the analysis of more than a dozen pretrained language models of varying sizes on two toxic text classification tasks (English), we demonstrate that focusing on accuracy measures alone can lead to models with wide variation in fairness characteristics. Specifically, we observe that fairness can vary even more than accuracy with increasing training data size and different random initializations. At the same time, we find that little of the fairness variation is explained by model size, despite claims in the literature. To improve model fairness without retraining, we show that two post-processing methods developed for structured, tabular data can be successfully applied to a range of pretrained language models.
Warning: This paper contains samples of offensive text.

17:00-18:00 (Forum)
#46 Sentence-level Privacy for Document Embeddings
Casey Meehan, Khalil Mrini and Kamalika Chaudhuri
User language data can contain highly sensitive personal content. As such, it is imperative to offer users a strong and interpretable privacy guarantee when learning from their data. In this work we propose SentDP, pure local differential privacy at the sentence level for a single user document. We propose a novel technique, DeepCandidate, that combines concepts from robust statistics and language modeling to produce high (768) dimensional, general ϵ-SentDP document embeddings. This guarantees that any single sentence in a document can be substituted with any other sentence while keeping the embedding ϵ-indistinguishable. Our experiments indicate that these private document embeddings are useful for downstream tasks like sentiment analysis and topic classification and even outperform baseline methods with weaker guarantees like word-level Metric DP.

17:00-18:00 (Forum)
#47 The Dangers of Underclaiming: Reasons for Caution When Reporting How NLP Systems Fail
Samuel R. Bowman
Researchers in NLP often frame and discuss research results in ways that serve to deemphasize the field's successes, often in response to the field's widespread hype. Though well-meaning, this has yielded many misleading or false claims about the limits of our best technology. This is a problem, and it may be more serious than it looks: It harms our credibility in ways that can make it harder to mitigate present-day harms, like those involving biased systems for content moderation or resume screening. It also limits our ability to prepare for the potentially enormous impacts of more distant future advances. This paper urges researchers to be careful about these claims and suggests some research directions and communication strategies that will make it easier to avoid or rebut them.

17:00-18:00 (Forum)
#48 Ethics Sheets for AI Tasks
Saif M. Mohammad
Several high-profile events, such as the mass testing of emotion recognition systems on vulnerable sub-populations and using question answering systems to make moral judgments, have highlighted how technology will often lead to more adverse outcomes for those that are already marginalized. At issue here are not just individual systems and datasets, but also the AI tasks themselves. In this position paper, I make a case for thinking about ethical considerations not just at the level of individual models and datasets, but also at the level of AI tasks. I will present a new form of such an effort, Ethics Sheets for AI Tasks, dedicated to fleshing out the assumptions and ethical considerations hidden in how a task is commonly framed and in the choices we make regarding the data, method, and evaluation. I will also present a template for ethics sheets with 50 ethical considerations, using the task of emotion recognition as a running example. Ethics sheets are a mechanism to engage with and document ethical considerations before building datasets and systems. Similar to survey articles, a small number of carefully created ethics sheets can serve numerous researchers and developers.

17:00-18:00 (Forum)
#49 Toward Annotator Group Bias in Crowdsourcing
Haochen Liu, Joseph Davis Thekinen, Sinem Mollaoglu, Da Tang, Ji Yang, Youlong Cheng, Hui Liu and Jiliang Tang
Crowdsourcing has emerged as a popular approach for collecting annotated data to train supervised machine learning models. However, annotator bias can lead to defective annotations. Though there are a few works investigating individual annotator bias, the group effects in annotators are largely overlooked. In this work, we reveal that annotators within the same demographic group tend to show consistent group bias in annotation tasks and thus we conduct an initial study on annotator group bias. We first empirically verify the existence of annotator group bias in various real-world crowdsourcing datasets. Then, we develop a novel probabilistic graphical framework GroupAnno to capture annotator group bias with an extended Expectation Maximization (EM) algorithm. We conduct experiments on both synthetic and real-world datasets. Experimental results demonstrate the effectiveness of our model in modeling annotator group bias in label aggregation and model learning over competitive baselines.

17:00-18:00 (Forum)
#50 Reinforcement Guided Multi-Task Learning Framework for Low-Resource Stereotype Detection
Rajkumar Pujari, Erik Oveson, Priyanka Kulkarni and Elnaz Nouri
As large Pre-trained Language Models (PLMs) trained on large amounts of data in an unsupervised manner become more ubiquitous, identifying various types of bias in the text has come into sharp focus. Existing 'Stereotype Detection' datasets mainly adopt a diagnostic approach toward large PLMs. Blodgett et. al. (2021) show that there are significant reliability issues with the existing benchmark datasets. Annotating a reliable dataset requires a precise understanding of the subtle nuances of how stereotypes manifest in text. In this paper, we annotate a focused evaluation set for 'Stereotype Detection' that addresses those pitfalls by de-constructing various ways in which stereotypes manifest in text. Further, we present a multi-task model that leverages the abundance of data-rich neighboring tasks such as hate speech detection, offensive language detection, misogyny detection, etc., to improve the empirical performance on 'Stereotype Detection'. We then propose a reinforcement-learning agent that guides the multi-task learning model by learning to identify the training examples from the neighboring tasks that help the target task the most. We show that the proposed models achieve significant empirical gains over existing baselines on all the tasks.

17:00-18:00 (Forum)
#51 Entropy-based Attention Regularization Frees Unintended Bias Mitigation from Lists
Giuseppe Attanasio, Debora Nozza, Dirk Hovy and Elena Baralis
Natural Language Processing (NLP) models risk overfitting to specific terms in the training data, thereby reducing their performance, fairness, and generalizability. E.g., neural hate speech detection models are strongly influenced by identity terms like gay, or women, resulting in false positives, severe unintended bias, and lower performance. Most mitigation techniques use lists of identity terms or samples from the target domain during training. However, this approach requires a-priori knowledge and introduces further bias if important terms are neglected. Instead, we propose a knowledge-free Entropy-based Attention Regularization (EAR) to discourage overfitting to training-specific terms. An additional objective function penalizes tokens with low self-attention entropy. We fine-tune BERT via EAR: the resulting model matches or exceeds state-of-the-art performance for hate speech classification and bias metrics on three benchmark corpora in English and Italian. EAR also reveals overfitting terms, i.e., terms most likely to induce bias, to help identify their effect on the model, task, and predictions.

17:00-18:00 (Forum)
#52 Using NLP to quantify the environmental cost and diversity benefits of in-person NLP conferences
Piotr Przybyła and Matthew Shardlow
The environmental costs of research are progressively important to the NLP community and their associated challenges are increasingly debated. In this work, we analyse the carbon cost (measured as CO2-equivalent) associated with journeys made by researchers attending in-person NLP conferences. We obtain the necessary data by text-mining all publications from the ACL anthology available at the time of the study (n=60,572) and extracting information about an author's affiliation, including their address. This allows us to estimate the corresponding carbon cost and compare it to previously known values for training large models. Further, we look at the benefits of in-person conferences by demonstrating that they can increase participation diversity by encouraging attendance from the region surrounding the host country. We show how the trade-off between carbon cost and diversity of an event depends on its location and type. Our aim is to foster further discussion on the best way to address the joint issue of emissions and diversity in the future.

17:00-18:00 (Forum)
#53 FairLex: A Multilingual Benchmark for Evaluating Fairness in Legal Text Processing
Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Felix Schwemer and Anders Søgaard
We present a benchmark suite of four datasets for evaluating the fairness of pre-trained language models and the techniques used to fine-tune them for downstream tasks. Our benchmarks cover four jurisdictions (European Council, USA, Switzerland, and China), five languages (English, German, French, Italian and Chinese) and fairness across five attributes (gender, age, region, language, and legal area). In our experiments, we evaluate pre-trained language models using several group-robust fine-tuning techniques and show that performance group disparities are vibrant in many cases, while none of these techniques guarantee fairness, nor consistently mitigate group disparities. Furthermore, we provide a quantitative and qualitative analysis of our results, highlighting open challenges in the development of robustness methods in legal NLP.

17:00-18:00 (Forum)
#54 ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection
Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray and Ece Kamar
Toxic language detection systems often falsely flag text that contains minority group mentions as toxic, as those groups are often the targets of online hate. Such over-reliance on spurious correlations also causes systems to struggle with detecting implicitly toxic language. To help mitigate these issues, we create ToxiGen, a new large-scale and machine-generated dataset of 274k toxic and benign statements about 13 minority groups. We develop a demonstration-based prompting framework and an adversarial classifier-in-the-loop decoding method to generate subtly toxic and benign text with a massive pretrained language model. Controlling machine generation in this way allows ToxiGen to cover implicitly toxic text at a larger scale, and about more demographic groups, than previous resources of human-written text. We conduct a human evaluation on a challenging subset of ToxiGen and find that annotators struggle to distinguish machine-generated text from human-written language. We also find that 94.5Using three publicly-available datasets, we show that finetuning a toxicity classifier on our data improves its performance on human-written data substantially. We also demonstrate that ToxiGen can be used to fight machine-generated toxicity as finetuning improves the classifier significantly on our evaluation subset.

17:00-18:00 (Forum)
#55 SafetyKit: First Aid for Measuring Safety in Open-domain Conversational Systems
Emily Dinan, Gavin Abercrombie, A. Stevie Bergman, Shannon L. Spruit, Dirk Hovy, Y-Lan Boureau and Verena Rieser
The social impact of natural language processing and its applications has received increasing attention. In this position paper, we focus on the problem of safety for end-to-end conversational AI. We survey the problem landscape therein, introducing a taxonomy of three observed phenomena: the Instigator, Yea-Sayer, and Impostor effects. We then empirically assess the extent to which current tools can measure these effects and current systems display them. We release these tools as part of a "first aid kit" (SafetyKit) to quickly assess apparent safety concerns. Our results show that, while current tools are able to provide an estimate of the relative safety of systems in various settings, they still have several shortcomings. We suggest several future directions and discuss ethical considerations.

Poster Session 3: Information Extraction

17:00-18:00 (Forum)

17:00-18:00 (Forum)
#56 FormNet: Structural Encoding beyond Sequential Modeling in Form Document Information Extraction
Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii and Tomas Pfister
Sequence modeling has demonstrated state-of-the-art performance on natural language and document understanding tasks. However, it is challenging to correctly serialize tokens in form-like documents in practice due to their variety of layout patterns. We propose FormNet, a structure-aware sequence model to mitigate the suboptimal serialization of forms. First, we design Rich Attention that leverages the spatial relationship between tokens in a form for more precise attention score calculation. Second, we construct Super-Tokens for each word by embedding representations from their neighboring tokens through graph convolutions. FormNet therefore explicitly recovers local syntactic information that may have been lost during serialization. In experiments, FormNet outperforms existing methods with a more compact model size and less pre-training data, establishing new state-of-the-art performance on CORD, FUNSD and Payment benchmarks.

17:00-18:00 (Forum)
#57 Multilingual knowledge graph completion with self-supervised adaptive graph alignment
Zijie Huang, Tianyu Cao, Haoming Jiang, Zheng Li, Hanqing Lu, Karthik Subbian, Yizhou Yizhou, Wei Wang and Bing Yin
Predicting missing facts in a knowledge graph (KG) is crucial as modern KGs are far from complete. Due to labor-intensive human labeling, this phenomenon deteriorates when handling knowledge represented in various languages. In this paper, we explore multilingual KG completion, which leverages limited seed alignment as a bridge, to embrace the collective knowledge from multiple languages. However, language alignment used in prior works is still not fully exploited: (1) alignment pairs are treated equally to maximally push parallel entities to be close, which ignores KG capacity inconsistency; (2) seed alignment is scarce and new alignment identification is usually in a noisily unsupervised manner. To tackle these issues, we propose a novel self-supervised adaptive graph alignment (SS-AGA) method. Specifically, SS-AGA fuses all KGs as a whole graph by regarding alignment as a new edge type. As such, information propagation and noise influence across KGs can be adaptively controlled via relation-aware attention weights. Meanwhile, SS-AGA features a new pair generator that dynamically captures potential alignment pairs in a self-supervised paradigm. Extensive experiments on both the public multilingual DBPedia KG and newly-created industrial multilingual E-commerce KG empirically demonstrate the effectiveness of SS-AGA

17:00-18:00 (Forum)
#58 A Graph Enhanced BERT Model for Event Prediction
Li Du, Xiao Ding, Yue Zhang, Ting Liu and Bing Qin
Predicting the subsequent event for an existing event context is an important but challenging task, as it requires understanding the underlying relationship between events. Previous methods propose to retrieve relational features from event graph to enhance the modeling of event correlation. However, the sparsity of event graph may restrict the acquisition of relevant graph information, and hence influence the model performance. To address this issue, we consider automatically building of event graph using a BERT model. To this end, we incorporate

an additional structured variable into BERT to learn to predict the event connections in the training process. Hence, in the test process, the connection relationship for unseen events can be predicted by the structured variable. Results on two event prediction tasks: script event prediction and story ending prediction, show that our approach can outperform state-of-the-art baseline methods.

17:00-18:00 (Forum)
#59 Distantly Supervised Named Entity Recognition via Confidence-Based Multi-Class Positive and Unlabeled Learning
Kang Zhou, Yuepei Li and Qi Li
In this paper, we study the named entity recognition (NER) problem under distant supervision. Due to the incompleteness of the external dictionaries and/or knowledge bases, such distantly annotated training data usually suffer from a high false negative rate. To this end, we formulate the Distantly Supervised NER (DS-NER) problem via Multi-class Positive and Unlabeled (MPU) learning and propose a theoretically and practically novel CONFidence-based MPU (Conf-MPU) approach. To handle the incomplete annotations, Conf-MPU consists of two steps. First, a confidence score is estimated for each token of being an entity token. Then, the proposed Conf-MPU risk estimation is applied to train a multi-class classifier for the NER task. Thorough experiments on two benchmark datasets labeled by various external knowledge demonstrate the superiority of the proposed Conf-MPU over existing DS-NER methods. Our code is available at Github.

17:00-18:00 (Forum)
#60 CONTaiNER: Few-Shot Named Entity Recognition via Contrastive Learning
Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca J. Passonneau and Rui Zhang
Named Entity Recognition (NER) in Few-Shot setting is imperative for entity tagging in low resource domains. Existing approaches only learn class-specific semantic features and intermediate representations from source domains. This affects generalizability to unseen target domains, resulting in suboptimal performances. To this end, we present CONTaiNER, a novel contrastive learning technique that optimizes the inter-token distribution distance for Few-Shot NER. Instead of optimizing class-specific attributes, CONTaiNER optimizes a generalized objective of differentiating between token categories based on their Gaussian-distributed embeddings. This effectively alleviates overfitting issues originating from training domains. Our experiments in several traditional test domains (OntoNotes, CoNLL'03, WNUT '17, GUM) and a new large scale Few-Shot NER dataset (Few-NERD) demonstrate that on average, CONTaiNER outperforms previous methods by 3

17:00-18:00 (Forum)
#61 Detection, Disambiguation, Re-ranking: Autoregressive Entity Linking as a Multi-Task Problem
Khalil Mrini, Shaoliang Nie, Jiatao Gu, Sinong Wang, Maziar Sanjabi and Hamed Firooz
We propose an autoregressive entity linking model, that is trained with two auxiliary tasks, and learns to re-rank generated samples at inference time. Our proposed novelties address two weaknesses in the literature. First, a recent method proposes to learn mention detection and then entity candidate selection, but relies on predefined sets of candidates. We use encoder-decoder autoregressive entity linking in order to bypass this need, and propose to train mention detection as an auxiliary task instead. Second, previous work suggests that re-ranking could help correct prediction errors. We add a new, auxiliary task, match prediction, to learn re-ranking. Without the use of a knowledge base or candidate sets, our model sets a new state of the art in two benchmark datasets of entity linking: COMETA in the biomedical domain, and AIDA-CoNLL in the news domain. We show through ablation studies that each of the two auxiliary tasks increases performance, and that re-ranking is an important factor to the increase. Finally, our low-resource experimental results suggest that performance on the main task benefits from the knowledge learned by the auxiliary tasks, and not just from the additional training data.

17:00-18:00 (Forum)
[TACL] **#62 Multilingual Autoregressive Entity Linking**
Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel and Fabio Petroni

17:00-18:00 (Forum)
[TACL] **#63 VILA: Improving Structured Content Extraction from Scientific PDFs Using Visual Layout Groups**
Zejiang Shen, Kyle Lo, Lucy Wang, Bailey Kuehl, Daniel Weld and Doug Downey

17:00-18:00 (Forum)
#64 Simple and Effective Knowledge-Driven Query Expansion for QA-Based Product Attribute Extraction
Keiji Shinzato, Naoki Yoshinaga, Yandi Xia and Wei-Te Chen
A key challenge in attribute value extraction (AVE) from e-commerce sites is how to handle a large number of attributes for diverse products. Although this challenge is partially addressed by a question answering (QA) approach which finds a value in product data for a given query (attribute), it does not work effectively for rare and ambiguous queries. We thus propose simple knowledge-driven query expansion based on possible answers (values) of a query (attribute) for QA-based AVE. We retrieve values of a query (attribute) from the training data to expand the query. We train a model with two tricks, knowledge dropout and knowledge token mixing, which mimic the imperfection of the value knowledge in testing. Experimental results on our cleaned version of AliExpress dataset show that our method improves the performance of AVE (+6.08 macro F1), especially for rare and ambiguous attributes (+7.82 and +6.86 macro F1, respectively).

17:00-18:00 (Forum)
#65 Prix-LM: Pretraining for Multilingual Knowledge Base Construction
Wenxuan Zhou, Fangyu Liu, Ivan Vulić, Nigel Collier and Muhao Chen
Knowledge bases (KBs) contain plenty of structured world and commonsense knowledge. As such, they often complement distributional text-based information and facilitate various downstream tasks. Since their manual construction is resource- and time-intensive, recent efforts have tried leveraging large pretrained language models (PLMs) to generate additional monolingual knowledge facts for KBs. However, such methods have not been attempted for building and enriching multilingual KBs. Besides wider application, such multilingual KBs can provide richer combined knowledge than monolingual (e.g., English) KBs. Knowledge expressed in different languages may be complementary and unequally distributed: this implies that the knowledge available in high-resource languages can be transferred to low-resource ones. To achieve this, it is crucial to represent multilingual knowledge in a shared/unified space. To this end, we propose a unified representation model, Prix-LM, for multilingual KB construction and completion. We leverage two types of knowledge, monolingual triples and cross-lingual links, extracted from existing multilingual KBs, and tune a multilingual language encoder XLM-R via a causal language modeling objective. Prix-LM integrates useful multilingual and KB-based factual knowledge into a single model. Experiments on standard entity-related tasks, such as link prediction in multiple languages, cross-lingual entity linking and bilingual lexicon induction, demonstrate its effectiveness, with gains

reported over strong task-specialised baselines.

17:00-18:00 (Forum)
#66 MILIE: Modular & Iterative Multilingual Open Information Extraction
Bhushan Kotnis, Kiril Gashteovski, Daniel Onoro Rubio, Ammar Shaker, Vanesa Rodriguez-Tembras, Makoto Takamoto, Mathias Niepert and Carolin Lawrence
Open Information Extraction (OpenIE) is the task of extracting (subject, predicate, object) triples from natural language sentences. Current OpenIE systems extract all triple slots independently. In contrast, we explore the hypothesis that it may be beneficial to extract triple slots iteratively: first extract easy slots, followed by the difficult ones by conditioning on the easy slots, and therefore achieve a better overall extraction. Based on this hypothesis, we propose a neural OpenIE system, MILIE, that operates in an iterative fashion. Due to the iterative nature, the system is also modularit is possible to seamlessly integrate rule based extraction systems with a neural end-to-end system, thereby allowing rule based systems to supply extraction slots which MILIE can leverage for extracting the remaining slots. We confirm our hypothesis empirically: MILIE outperforms SOTA systems on multiple languages ranging from Chinese to Arabic. Additionally, we are the first to provide an OpenIE test dataset for Arabic and Galician.

17:00-18:00 (Forum)
#67 Continual Few-shot Relation Learning via Embedding Space Regularization and Data Augmentation
Chengwei Qin and Shafiq Joty
Existing continual relation learning (CRL) methods rely on plenty of labeled training data for learning a new task, which can be hard to acquire in real scenario as getting large and representative labeled data is often expensive and time-consuming. It is therefore necessary for the model to learn novel relational patterns with very few labeled data while avoiding catastrophic forgetting of previous task knowledge. In this paper, we formulate this challenging yet practical problem as continual few-shot relation learning (CFRL). Based on the finding that learning for new emerging few-shot tasks often results in feature distributions that are incompatible with previous tasks' learned distributions, we propose a novel method based on embedding space regularization and data augmentation. Our method generalizes to new few-shot tasks and avoids catastrophic forgetting of previous tasks by enforcing extra constraints on the relational embeddings and by adding extra relevant data in a self-supervised manner. With extensive experiments we demonstrate that our method can significantly outperform previous state-of-the-art methods in CFRL task settings.

17:00-18:00 (Forum)
#68 Towards Consistent Document-level Entity Linking: Joint Models for Entity Linking and Coreference Resolution
Klim Zaporojets, Johannes Deleu, Yiwei Jiang, Thomas Demeester and Chris Develder
We consider the task of document-level entity linking (EL), where it is important to make consistent decisions for entity mentions over the full document jointly. We aim to leverage explicit "connections" among mentions within the document itself: we propose to join EL and coreference resolution (coref) in a single structured prediction task over directed trees and use a globally normalized model to solve it. This contrasts with related works where two separate models are trained for each of the tasks and additional logic is required to merge the outputs. Experimental results on two datasets show a boost of up to +5

17:00-18:00 (Forum)
#69 Prompt for Extraction? PAIE: Prompting Argument Interaction for Event Argument Extraction
Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang and Jing Shao
In this paper, we propose an effective yet efficient model PAIE for both sentence-level and document-level Event Argument Extraction (EAE), which also generalizes well when there is a lack of training data. On the one hand, PAIE utilizes prompt tuning for extractive objectives to take the best advantages of Pre-trained Language Models (PLMs). It introduces two span selectors based on the prompt to select start/end tokens among input texts for each role. On the other hand, it captures argument interactions via multi-role prompts and conducts joint optimization with optimal span assignments via a bipartite matching loss. Also, with a flexible prompt design, PAIE can extract multiple arguments with the same role instead of conventional heuristic threshold tuning. We have conducted extensive experiments on three benchmarks, including both sentence- and document-level EAE. The results present promising improvements from PAIE (3.5% and 2.3% F1 gains in average on three benchmarks, for PAIE-base and PAIE-large respectively). Further analysis demonstrates the efficiency, generalization to few-shot settings, and effectiveness of different extractive prompt tuning strategies. Our code is available at https://github.com/mayubo2333/PAIE.

17:00-18:00 (Forum)
#70 Domain Adaptation in Multilingual and Multi-Domain Monolingual Settings for Complex Word Identification
George-Eduard Zaharia, Răzvan-Alexandru Smădu, Dumitru Clementin Cercel and Mihai Dascalu
Complex word identification (CWI) is a cornerstone process towards proper text simplification. CWI is highly dependent on context, whereas its difficulty is augmented by the scarcity of available datasets which vary greatly in terms of domains and languages. As such, it becomes increasingly more difficult to develop a robust model that generalizes across a wide array of input examples. In this paper, we propose a novel training technique for the CWI task based on domain adaptation to improve the target character and context representations. This technique addresses the problem of working with multiple domains, inasmuch as it creates a way of smoothing the differences between the explored datasets. Moreover, we also propose a similar auxiliary task, namely text simplification, that can be used to complement lexical complexity prediction. Our model obtains a boost of up to 2.42

17:00-18:00 (Forum)
#71 Leveraging Expert Guided Adversarial Augmentation For Improving Generalization in Named Entity Recognition
Aaron Reich, Jiaao Chen, Aastha Agrawal, Yanzhe Zhang and Diyi Yang
Named Entity Recognition (NER) systems often demonstrate great performance on in-distribution data, but perform poorly on examples drawn from a shifted distribution. One way to evaluate the generalization ability of NER models is to use adversarial examples, on which the specific variations associated with named entities are rarely considered. To this end, we propose leveraging expert-guided heuristics to change the entity tokens and their surrounding contexts thereby altering their entity types as adversarial attacks. Using expert-guided heuristics, we augmented the CoNLL 2003 test set and manually annotated it to construct a high-quality challenging set. We found that state-of-the-art NER systems trained on CoNLL 2003 training data drop performance dramatically on our challenging set. By training on adversarial augmented training examples and using mixup for regularization, we were able to significantly improve the performance on the challenging set as well as improve out-of-domain generalization which we evaluated by using OntoNotes data. We have publicly released our dataset and code at https://github.com/GT-SALT/Guided-Adversarial-Augmentation.

17:00-18:00 (Forum)
#72 RelationPrompt: Leveraging Prompts to Generate Synthetic Data for Zero-Shot Relation Triplet Extraction
Yew Ken Chia, Lidong Bing, Soujanya Poria and Luo Si
Despite the importance of relation extraction in building and representing knowledge, less research is focused on generalizing to unseen relations types. We introduce the task setting of Zero-Shot Relation Triplet Extraction (ZeroRTE) to encourage further research in low-resource relation extraction methods. Given an input sentence, each extracted triplet consists of the head entity, relation label, and tail entity where the relation label is not seen at the training stage. To solve ZeroRTE, we propose to synthesize relation examples by prompting language models to generate structured texts. Concretely, we unify language model prompts and structured text approaches to design a structured prompt template for generating synthetic relation samples when conditioning on relation label prompts (RelationPrompt). To overcome the limitation for extracting multiple relation triplets in a sentence, we design a novel Triplet Search Decoding method. Experiments on FewRel and Wiki-ZSL datasets show the efficacy of RelationPrompt for the ZeroRTE task and zero-shot relation classification. Our code and data are available at github.com/declare-lab/RelationPrompt.

17:00-18:00 (Forum)
#73 Label Semantics for Few Shot Named Entity Recognition
Jie Ma, Miguel Ballesteros, Srikanth Doss, Rishita Anubhai, Sunil Mallya, Yaser Al-Onaizan and Dan Roth
We study the problem of few shot learning for named entity recognition. Specifically, we leverage the semantic information in the names of the labels as a way of giving the model additional signal and enriched priors. We propose a neural architecture that consists of two BERT encoders, one to encode the document and its tokens and another one to encode each of the labels in natural language format. Our model learns to match the representations of named entities computed by the first encoder with label representations computed by the second encoder. The label semantics signal is shown to support improved state-of-the-art results in multiple few shot NER benchmarks and on-par performance in standard benchmarks. Our model is especially effective in low resource settings.

17:00-18:00 (Forum)
#74 DeepStruct: Pretraining of Language Models for Structure Prediction
Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang and Dawn Song
We introduce a method for improving the structural understanding abilities of language models. Unlike previous approaches that finetune the models with task-specific augmentation, we pretrain language models to generate structures from the text on a collection of task-agnostic corpora. Our structure pretraining enables zero-shot transfer of the learned knowledge that models have about the structure tasks. We study the performance of this approach on 28 datasets, spanning 10 structure prediction tasks including open information extraction, joint entity and relation extraction, named entity recognition, relation classification, semantic role labeling, event extraction, coreference resolution, factual probe, intent detection, and dialogue state tracking. We further enhance the pretraining with the task-specific training sets. We show that a 10B parameter language model transfers non-trivially to most tasks and obtains state-of-the-art performance on 21 of 28 datasets that we evaluate. Our code and datasets will be made publicly available.

17:00-18:00 (Forum)
[TACL] **#75 Predicting Document Coverage for Relation Extraction**
Sneha Singhania, Simon Razniewski and Gerhard Weikum

17:00-18:00 (Forum)
[DEMO] **QuickGraph: A Rapid Annotation Tool for Knowledge Graph Extraction from Technical Text**
Tyler Bikaun, Michael Stewart and Wei Liu
Acquiring high-quality annotated corpora for complex multi-task information extraction (MT-IE) is an arduous and costly process for human-annotators. Adoption of unsupervised techniques for automated annotation have thus become popular. However, these techniques rely heavily on dictionaries, gazetteers, and knowledge bases. While such resources are abundant for general domains, they are scarce for specialised technical domains. To tackle this challenge, we present QuickGraph, the first collaborative MT-IE annotation tool built with indirect weak supervision and clustering to maximise annotator productivity.
QuickGraph's main contribution is a set of novel features that enable knowledge graph extraction through rapid and consistent complex multi-task entity and relation annotation. In this paper, we discuss these key features and qualitatively compare QuickGraph to existing annotation tools.

Poster Session 3: Information Retrieval and Text Mining

17:00-18:00 (Forum)

17:00-18:00 (Forum)
[TACL] **#76 ABNIRML: Analyzing the Behavior of Neural IR Models**
Sean MacAvaney, Sergey Feldman, Nazli Goharian, Doug Downey and Arman Cohan

17:00-18:00 (Forum)
#77 Augmenting Document Representations for Dense Retrieval with Interpolation and Perturbation
Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang and Jong C. Park
Dense retrieval models, which aim at retrieving the most relevant document for an input query on a dense representation space, have gained considerable attention for their remarkable success. Yet, dense models require a vast amount of labeled training data for notable performance, whereas it is often challenging to acquire query-document pairs annotated by humans. To tackle this problem, we propose a simple but effective Document Augmentation for dense Retrieval (DAR) framework, which augments the representations of documents with their interpolation and perturbation. We validate the performance of DAR on retrieval tasks with two benchmark datasets, showing that the proposed DAR significantly outperforms relevant baselines on the dense retrieval of both the labeled and unlabeled documents.

17:00-18:00 (Forum)
#78 SDR: Efficient Neural Re-ranking using Succinct Document Representation
Nachshon Cohen, Amit Portnoy, Besnik Fetahu and Amir Ingber
BERT based ranking models have achieved superior performance on various information retrieval tasks. However, the large number of parameters and complex self-attention operations come at a significant latency overhead. To remedy this, recent works propose late-interaction architectures, which allow pre-computation of intermediate document representations, thus reducing latency. Nonetheless, having solved the immediate latency issue, these methods now introduce storage costs and network fetching latency, which limit their adoption in real-life production systems.
In this work, we propose the Succinct Document Representation (SDR) scheme that computes *highly compressed* intermediate document representations, mitigating the storage/network issue. Our approach first reduces the dimension of token representations by encoding them using a novel autoencoder architecture that uses the document's textual content in both the encoding and decoding phases. After this token encoding step, we further reduce the size of the document representations using modern quantization techniques.
Evaluation on MSMARCO's passage re-reranking task show that compared to existing approaches using compressed document representations, our method is highly efficient, achieving 4x–11.6x higher compression rates for the same ranking quality. Similarly, on the TREC CAR dataset, we achieve 7.7x higher compression rate for the same ranking quality.

17:00-18:00 (Forum)
#79 Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval
Luyu Gao and Jamie Callan
Recent research demonstrates the effectiveness of using fine-tuned language models (LM) for dense retrieval. However, dense retrievers are hard to train, typically requiring heavily engineered fine-tuning pipelines to realize their full potential. In this paper, we identify and address two underlying problems of dense retrievers: i) fragility to training data noise and ii) requiring large batches to robustly learn the embedding space. We use the recently proposed Condenser pre-training architecture, which learns to condense information into the dense vector through LM pre-training. On top of it, we propose coCondenser, which adds an unsupervised corpus-level contrastive loss to warm up the passage embedding space. Experiments on MS-MARCO, Natural Question, and Trivia QA datasets show that coCondenser removes the need for heavy data engineering such as augmentation, synthesis, or filtering, and the need for large batch training. It shows comparable performance to RocketQA, a state-of-the-art, heavily engineered system, using simple small batch fine-tuning.

17:00-18:00 (Forum)
#80 TABi: Type-Aware Bi-Encoders for Open-Domain Entity Retrieval
Megan Leszczynski, Daniel Y Fu, Mayee F Chen and Christopher Re
Entity retrieval—retrieving information about entity mentions in a query—is a key step in open-domain tasks, such as question answering or fact checking. However, state-of-the-art entity retrievers struggle to retrieve rare entities for ambiguous mentions due to biases towards popular entities. Incorporating knowledge graph types during training could help overcome popularity biases, but there are several challenges: (1) existing type-based retrieval methods require mention boundaries as input, but open-domain tasks run on unstructured text, (2) type-based methods should not compromise overall performance, and (3) type-based methods should be robust to noisy and missing types. In this work, we introduce TABi, a method to jointly train bi-encoders on knowledge graph types and unstructured text for entity retrieval for open-domain tasks. TABi leverages a type-enforced contrastive loss to encourage entities and queries of similar types to be close in the embedding space. TABi improves retrieval of rare entities on the Ambiguous Entity Retrieval (AmbER) sets, while maintaining strong overall retrieval performance on open-domain tasks in the KILT benchmark compared to state-of-the-art retrievers. TABi is also robust to incomplete type systems, improving rare entity retrieval over baselines with only 5

17:00-18:00 (Forum)
#81 LaPraDoR: Unsupervised Pretrained Dense Retriever for Zero-Shot Text Retrieval
Canwen Xu, Daya Guo, Nan Duan and Julian McAuley
In this paper, we propose LaPraDoR, a pretrained dual-tower dense retriever that does not require any supervised data for training. Specifically, we first present Iterative Contrastive Learning (ICoL) that iteratively trains the query and document encoders with a cache mechanism. ICoL not only enlarges the number of negative instances but also keeps representations of cached examples in the same hidden space. We then propose Lexicon-Enhanced Dense Retrieval (LEDR) as a simple yet effective way to enhance dense retrieval with lexical matching. We evaluate LaPraDoR on the recently proposed BEIR benchmark, including 18 datasets of 9 zero-shot text retrieval tasks. Experimental results show that LaPraDoR achieves state-of-the-art performance compared with supervised dense retrieval models, and further analysis reveals the effectiveness of our training strategy and objectives. Compared to re-ranking, our lexicon-enhanced approach can be run in milliseconds (22.5x faster) while achieving superior performance.

Student Research Workshop

17:00-18:00 (Liffey Hall 1)

17:00-17:20 (Liffey Hall 1)
Darkness can not drive out darkness: Investigating Bias in Hate SpeechDetection Models
Fatma Elsafoury
It has become crucial to develop tools for automated hate speech and abuse detection. These tools would help to stop the bullies and the haters and provide a safer environment for individuals especially from marginalized groups to freely express themselves. However, recent research shows that machine learning models are biased and they might make the right decisions for the wrong reasons. In this thesis, I set out to understand the performance of hate speech and abuse detection models and the different biases that could influence them. I show that hate speech and abuse detection models are not only subject to social bias but also to other types of bias that have not been explored before. Finally, I investigate the causal effect of the social and intersectional bias on the performance and unfairness of hate speech detection models.

17:20-17:40 (Liffey Hall 1)
What do Models Learn From Training on More Than Text? Measuring Visual Commonsense Knowledge
Lovisa Hagström and Richard Johansson
There are limitations in learning language from text alone. Therefore, recent focus has been on developing multimodal models. However, few benchmarks exist that can measure what language models learn about language from multimodal training. We hypothesize that training on a

visual modality should improve on the visual commonsense knowledge in language models. Therefore, we introduce two evaluation tasks for measuring visual commonsense knowledge in language models (code publicly available at: github.com/lovhag/measure-visual-commonsense-knowledge) and use them to evaluate different multimodal models and unimodal baselines. Primarily, we find that the visual commonsense knowledge is not significantly different between the multimodal models and unimodal baseline models trained on visual text data.

17:40-18:00 (Liffey Hall 1)
A Checkpoint on Multilingual Misogyny Identification
Arianna Muti and Alberto Barrón-Cedeño
We address the problem of identifying misogyny in tweets in mono and multilingual settings in three languages: English, Italian, and Spanish. We explore model variations considering single and multiple languages both in the pre-training of the transformer and in the training of the downstream task to explore the feasibility of detecting misogyny through a transfer learning approach across multiple languages. That is, we train monolingual transformers with monolingual data, and multilingual transformers with both monolingual and multilingual data. Our models reach state-of-the-art performance on all three languages. The single-language BERT models perform the best, closely followed by different configurations of multilingual BERT models. The performance drops in zero-shot classification across languages. Our error analysis shows that multilingual and monolingual models tend to make the same mistakes.

Social Event: Guinness Storehouse
19:30-23:00 - **Auditorium** (Guinness Storehouse)

Main Conference: Tuesday, May 24, 2022

Virtual Poster Session 1 - 07:30-08:30

VPS1: Computational Social Science and Cultural Analytics

07:30-08:30 (GatherTown)

Findings: Human Language Modeling
Speaker: Nikita Soni

Long: Doctor Recommendation in Online Health Forums via Expertise Learning
Speaker: Xiaoxin Lu

Long: Tackling Fake News Detection by Continually Improving Social Context Representations using Graph Neural Networks
Speaker: Nikhil Mehta

Outstanding Paper: Inducing Positive Perspectives with Text Reframing
Speaker: Caleb Ziems

Long: Automatic Identification and Classification of Bragging in Social Media
Speaker: Mali Jin

Long: Zoom Out and Observe: News Environment Perception for Fake News Detection
Speaker: Qiang Sheng

Long: Reports of personal experiences and stories in argumentation: datasets and analysis
Speaker: Neele Falk

Long: Leveraging Wikipedia article evolution for promotional tone detection
Speaker: Christine de Kock

Long: Should a Chatbot be Sarcastic? Understanding User Preferences Towards Sarcasm Generation
Speaker: Silviu Vlad Oprea

Long: Improving the Generalizability of Depression Detection by Leveraging Clinical Questionnaires
Speaker: Thong Nguyen

SRW: Towards Fine-grained Classification of Climate Change related Social Media Text
Speaker: Roopal Vaid

VPS1: Dialogue and Interactive Systems

07:30-08:30 (GatherTown)

Long: UniTranSeR: A Unified Transformer Semantic Representation Framework for Multimodal Task-Oriented Dialog System
Speaker: Zhiyuan Ma

Long: Dynamic Schema Graph Fusion Network for Multi-Domain Dialogue State Tracking
Speaker: Yue Feng

Findings: A Slot Is Not Built in One Utterance: Spoken Language Dialogs with Sub-Slots
Speaker: Sai Zhang

Long: Multi-Party Empathetic Dialogue Generation: A New Task for Dialog Systems
Speaker: LingYu Zhu

Long: MISC: A Mixed Strategy-Aware Model integrating COMET for Emotional Support Conversation
Speaker: Quan Tu

Findings: Towards Large-Scale Interpretable Knowledge Graph Reasoning for Dialogue Systems
Speaker: Yi-Lin Tuan

Long: Perceiving the World: Question-guided Reinforcement Learning for Text-based Games
Speaker: Yunqiu Xu

Long: DEAM: Dialogue Coherence Evaluation using AMR-based Semantic Manipulations
Speaker: Sarik Ghazarian

Findings: HybriDialogue: An Information-Seeking Dialogue Dataset Grounded on Tabular and Textual Data
Speaker: Kai Nakamura

Findings: Inverse is Better! Fast and Accurate Prompt for Few-shot Slot Tagging
Speaker: Yutai Hou

Findings: Hierarchical Inductive Transfer for Continual Dialogue Learning
Speaker: Shaoxiong Feng

Long: Where to Go for the Holidays: Towards Mixed-Type Dialogs for Clarification of User Goals
Speaker: Zeming Liu

Long: Continual Prompt Tuning for Dialog State Tracking
Speaker: Qi Zhu

Findings: DS-TOD: Efficient Domain Specialization for Task-Oriented Dialog
Speaker: Chia-Chien Hung

Short: Learning-by-Narrating: Narrative Pre-Training for Zero-Shot Dialogue Comprehension
Speaker: Chao Zhao

Long: GlobalWoZ: Globalizing MultiWoZ to Develop Multilingual Task-Oriented Dialogue Systems
Speaker: Bosheng Ding

Short: Mismatch between Multi-turn Dialogue and its Evaluation Metric in Dialogue State Tracking
Speaker: Takyoung Kim

Short: Towards Fair Evaluation of Dialogue State Tracking by Flexible Incorporation of Turn-level Performances
Speaker: Suvodip Dey

Findings: TegTok: Augmenting Text Generation via Task-specific and Open-world Knowledge
Speaker: Chao-Hong Tan

Long: Multimodal Dialogue Response Generation
Speaker: Qingfeng Sun

Long: Achieving Conversational Goals with Unsupervised Post-hoc Knowledge Injection
Speaker: Bodhisattwa Prasad Majumder

Long: There Are a Thousand Hamlets in a Thousand People's Eyes: Enhancing Knowledge-grounded Dialogue with Personal Memory
Speaker: Tingchen Fu

Long: Multi-Task Pre-Training for Plug-and-Play Task-Oriented Dialogue System
Speaker: Yixuan Su

Findings: On Controlling Fallback Responses for Grounded Dialogue Generation
Speaker: Hongyuan Lu

Long: DialogVED: A Pre-trained Latent Variable Encoder-Decoder Model for Dialog Response Generation
Speaker: Wei Chen

Long: An Interpretable Neuro-Symbolic Reasoning Framework for Task-Oriented Dialogue Generation
Speaker: Shiquan Yang

Long: CICERO: A Dataset for Contextualized Commonsense Inference in Dialogues
Speaker: Deepanway Ghosal

Long: HeterMPC: A Heterogeneous Graph Neural Network for Response Generation in Multi-Party Conversations
Speaker: Jia-Chen Gu

Long: KNN-Contrastive Learning for Out-of-Domain Intent Classification
Speaker: yunhua zhou

Findings: Mitigating Contradictions in Dialogue Based on Contrastive Learning
Speaker: Weizhao Li

Short: UniGDD: A Unified Generative Framework for Goal-Oriented Document-Grounded Dialogue
Speaker: Chang Gao

Long: When did you become so smart, oh wise one?! Sarcasm Explanation in Multi-modal Multi-party Dialogues
Speaker: Shivani Kumar

Long: SalesBot: Transitioning from Chit-Chat to Task-Oriented Dialogues
Speaker: Ssu Chiu

Long: Achieving Reliable Human Assessment of Open-Domain Dialogue Systems
Speaker: Tianbo Ji

Findings: Addressing Resource and Privacy Constraints in Semantic Parsing Through Data Augmentation
Speaker: Kevin Yang

Long: ChatMatch: Evaluating Chatbots by Autonomous Chat Tournaments
Speaker: Ruolan Yang

Short: Probing the Robustness of Trained Metrics for Conversational Dialogue Systems
Speaker: Jan Deriu

Short: Rethinking and Refining the Distinct Metric
Speaker: Siyang Liu

Findings: Dialogue Summaries as Dialogue States (DS2), Template-Guided Summarization for Few-shot Dialogue State Tracking
Speaker: Jamin Shin

Findings: Rethinking Offensive Text Detection as a Multi-Hop Reasoning Problem
Speaker: Qiang Zhang

Long: Situated Dialogue Learning through Procedural Environment Generation
Speaker: Prithviraj Ammanabrolu

Findings: What is wrong with you?: Leveraging User Sentiment for Automatic Dialog Evaluation
Speaker: Sarik Ghazarian

TACL: TopiOCQA: Open-domain Conversational Question Answering with Topic Switching
Speaker: Vaibhav Adlakha

SRW: Building a Dialogue Corpus Annotated with Expressed and Experienced Emotions
Speaker: Tatsuya Ide

SRW: Integrating Question Rewrites in Conversational Question Answering: A Reinforcement Learning Approach
Speaker: Etsuko Ishii

VPS1: Discourse and Pragmatics & Ethics in NLP"

07:30-08:30 (GatherTown)

Long: RST Discourse Parsing with Second-Stage EDU-Level Pre-training
Speaker: Nan Yu

Long: Entity-based Neural Local Coherence Modeling
Speaker: Sungho Jeon

TACL: Out-of-Domain Discourse Dependency Parsing via Bootstrapping: An Empirical Analysis on Its Effectiveness and Limitation
Speaker: Noriki Nishida

VPS1: Ethics in NLP

07:30-08:30 (GatherTown)

Findings: Mitigating Gender Bias in Distilled Language Models via Counterfactual Role Reversal
Speaker: Umang Gupta

Long: Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts
Speaker: Yue Guo

Long: Toward Annotator Group Bias in Crowdsourcing
Speaker: Haochen Liu, Haochen Liu

Long: An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models
Speaker: Nicholas Meade

Long: The Moral Integrity Corpus: A Benchmark for Ethical Dialogue Systems
Speaker: Caleb Ziems

Long: SafetyKit: First Aid for Measuring Safety in Open-domain Conversational Systems
Speaker: Emily Dinan

Findings: Assessing Multilingual Fairness in Pre-trained Multimodal Representations
Speaker: Jialu Wang

Long: Measuring Fairness of Text Classifiers via Prediction Sensitivity
Speaker: Satyapriya Krishna

Long: Reinforcement Guided Multi-Task Learning Framework for Low-Resource Stereotype Detection
Speaker: Rajkumar Pujari

Long: Ethics Sheets for AI Tasks
Speaker: Saif Mohammad

Long: French CrowS-Pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English
Speaker: Aurélie Névéol

VPS1: Generation

07:30-08:30 (GatherTown)

Findings: Compilable Neural Code Generation with Compiler Feedback
Speaker: Xin Wang

Long: Rare Tokens Degenerate All Tokens: Improving Neural Text Generation via Adaptive Gradient Gating for Rare Token Embeddings
Speaker: Sangwon Yu

Findings: Multi-Scale Distribution Deep Variational Autoencoder for Explanation Generation
Speaker: ZeFeng Cai

Short: High probability or low information? The probability–quality paradox in language generation
Speaker: Clara Meister

Long: Quality Controlled Paraphrase Generation
Speaker: Elron Bandel

Long: Spurious Correlations in Reference-Free Evaluation of Text Generation
Speaker: Esin Durmus

Short: Multilingual Pre-training with Language and Task Adaptation for Multilingual Text Style Transfer
Speaker: Huiyuan Lai

Findings: Multi-task Learning for Paraphrase Generation With Keyword and Part-of-Speech Reconstruction
Speaker: Xuhang Xie

Long: Hybrid Semantics for Goal-Directed Natural Language Generation
Speaker: Connor Baumler

Long: PLANET: Dynamic Content Planning in Autoregressive Transformers for Long-form Text Generation
Speaker: Zhe Hu

Long: Generating Scientific Claims for Zero-Shot Scientific Fact Checking
Speaker: Dustin Wright

Long: Hierarchical Sketch Induction for Paraphrase Generation
Speaker: Tom Hosking

Long: Tailor: Generating and Perturbing Text with Semantic Controls
Speaker: Alexis Ross

Findings: CaM-Gen: Causally Aware Metric-Guided Text Generation
Speaker: Navita Goyal

Long: Improving Compositional Generalization with Self-Training for Data-to-Text Generation
Speaker: Sanket Vaibhav Mehta

Findings: MReD: A Meta-Review Dataset for Structure-Controllable Text Generation
Speaker: Chenhui Shen

Findings: Learning to Describe Solutions for Bug Reports Based on Developer Discussions
Speaker: Sheena Panthaplackel

Long: Non-neural Models Matter: a Re-evaluation of Neural Referring Expression Generation Systems
Speaker: Fahime Same, Fahime Same

Long: Fine-Grained Controllable Text Generation Using Non-Residual Prompting
Speaker: Fredrik Carlsson

Short: Rewarding Semantic Similarity under Optimized Alignments for AMR-to-Text Generation
Speaker: Lisa Jin

Findings: GCPG: A General Framework for Controllable Paraphrase Generation
Speaker: Kexin Yang

VPS1: Information Extraction

07:30-08:30 (GatherTown)

Findings: RelationPrompt: Leveraging Prompts to Generate Synthetic Data for Zero-Shot Relation Triplet Extraction
Speaker: Yew Ken Chia

Findings: LEVEN: A Large-Scale Chinese Legal Event Detection Dataset
Speaker: Feng Yao

Long: An Unsupervised Multiple-Task and Multiple-Teacher Model for Cross-lingual Named Entity Recognition
Speaker: Zhuoran Li

Long: Parallel Instance Query Network for Named Entity Recognition
Speaker: Yongliang Shen

Short: Event-Event Relation Extraction using Probabilistic Box Embedding
Speaker: EunJeong Hwang

Long: Divide and Denoise: Learning from Noisy Labels in Fine-Grained Entity Typing with Cluster-Wise Loss Correction
Speaker: Kunyuan Pang

Findings: A Transformational Biencoder with In-Domain Negative Sampling for Zero-Shot Entity Linking
Speaker: Kai Sun

Short: Complex Evolutional Pattern Learning for Temporal Knowledge Graph Reasoning
Speaker: Zixuan Li

Long: Alignment-Augmented Consistent Translation for Multilingual Open Information Extraction
Speaker: Keshav Kolluru

Findings: Decomposed Meta-Learning for Few-Shot Named Entity Recognition
Speaker: Tingting Ma

Findings: Learning Reasoning Patterns for Relational Triple Extraction with Mutual Generation of Text and Graph
Speaker: Yubo Chen

Short: PARE: A Simple and Strong Baseline for Monolingual and Multilingual Distantly Supervised Relation Extraction
Speaker: Vipul Rathore

Long: Good Examples Make A Faster Learner: Simple Demonstration-based Learning for Low-resource NER
Speaker: Dong-Ho Lee

Findings: Improving Relation Extraction through Syntax-induced Pre-training with Dependency Masking
Speaker: Yuanhe Tian, Yuanhe Tian

Long: Show Me More Details: Discovering Hierarchies of Procedures from Semi-structured Web Data
Speaker: Shuyan Zhou

Long: Saliency as Evidence: Event Detection with Trigger Saliency Attribution
Speaker: Jian Liu, Jian Liu

Findings: HiCLRE: A Hierarchical Contrastive Learning Framework for Distantly Supervised Relation Extraction
Speaker: Dongyang Li

Short: A Simple but Effective Pluggable Entity Lookup Table for Pre-trained Language Models
Speaker: Deming Ye

Long: Packed Levitated Marker for Entity and Relation Extraction
Speaker: Deming Ye

Long: Dynamic Prefix-Tuning for Generative Template-based Event Extraction
Speaker: Xiao Liu

Long: Prix-LM: Pretraining for Multilingual Knowledge Base Construction
Speaker: Wenxuan Zhou

Long: Few-shot Named Entity Recognition with Self-describing Networks
Speaker: jiawei chen

Long: Unified Structure Generation for Universal Information Extraction
Speaker: Yaojie Lu

Long: Pre-training to Match for Unified Low-shot Relation Extraction
Speaker: Fangchao Liu

Long: Nested Named Entity Recognition as Latent Lexicalized Constituency Parsing
Speaker: Chao Lou

Long: Prompt for Extraction? PAIE: Prompting Argument Interaction for Event Argument Extraction
Speaker: Yubo Ma

Findings: Consistent Representation Learning for Continual Relation Extraction
Speaker: Kang Zhao

Long: MILIE: Modular & Iterative Multilingual Open Information Extraction
Speaker: Bhushan Kotnis

Long: Boundary Smoothing for Named Entity Recognition
Speaker: Enwei Zhu

Long: Rethinking Negative Sampling for Handling Missing Entity Annotations
Speaker: Yangming Li

Long: Distantly Supervised Named Entity Recognition via Confidence-Based Multi-Class Positive and Unlabeled Learning
Speaker: Kang Zhou

Findings: Improving Candidate Retrieval with Entity Profile Generation for Wikidata Entity Linking
Speaker: Tuan Lai, Tuan Lai

Short: Towards Consistent Document-level Entity Linking: Joint Models for Entity Linking and Coreference Resolution
Speaker: Klim Zaporojets

TACL: Multilingual Autoregressive Entity Linking
Speaker: Nicola De Cao

TACL: Predicting Document Coverage for Relation Extraction
Speaker: Sneha Singhania

SRW: RFBFN: A Relation-First Blank Filling Network for Joint Relational Triple Extraction
Speaker: Zhe Li

SRW: TeluguNER: Leveraging Multi-Domain Named Entity Recognition with Deep Transformers
Speaker: Suma Reddy Duggenpudi

VPS1: Information Retrieval and Text Mining

07:30-08:30 (GatherTown)

Long: Sentence-aware Contrastive Learning for Open-Domain Passage Retrieval
Speaker: Bohong Wu

Long: Accelerating Code Search with Deep Hashing and Code Classification
Speaker: Wenchao Gu

Long: Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval
Speaker: Luyu Gao

Short: Augmenting Document Representations for Dense Retrieval with Interpolation and Perturbation
Speaker: Soyeong Jeong

Long: An Effective and Efficient Entity Alignment Decoding Algorithm via Third-Order Tensor Isomorphism
Speaker: Xin Mao

Long: UCTopic: Unsupervised Contrastive Learning for Phrase Representations and Topic Mining
Speaker: Jiacheng Li

Long: SDR: Efficient Neural Re-ranking using Succinct Document Representation
Speaker: Nachshon Cohen

TACL: ABNIRML: Analyzing the Behavior of Neural IR Models
Speaker: Sean MacAvaney

VPS1: Interpretability and Analysis of Models for NLP

07:30-08:30 (GatherTown)

Short: Are Shortest Rationales the Best Explanations for Human Understanding?
Speaker: Hua Shen

Short: Does BERT Know that the IS-A Relation Is Transitive?
Speaker: Ruixi Lin

Findings: MoEfication: Transformer Feed-forward Layers are Mixtures of Experts
Speaker: Zhengyan Zhang

Findings: From BERT's Point of View: Revealing the Prevailing Contextual Differences
Speaker: Carolin Schuster

Long: Probing as Quantifying the Inductive Bias of Pre-trained Representations
Speaker: Alexander Immer

Findings: Finding the Dominant Winning Ticket in Pre-Trained Language Models
Speaker: Zhuocheng Gong

Long: Are Prompt-based Models Clueless?
Speaker: Pride Kavumba

Long: On the Sensitivity and Stability of Model Interpretations in NLP
Speaker: Fan Yin

Long: ProtoTEx: Explaining Model Decisions with Prototype Tensors
Speaker: Anubrata Das

Long: Adaptive Testing and Debugging of NLP Models
Speaker: Marco Tulio Ribeiro

Long: Can Transformer be Too Compositional? Analysing Idiom Processing in Neural Machine Translation
Speaker: Verna Dankers

Long: An Investigation of the (In)effectiveness of Counterfactually Augmented Data
Speaker: Nitish Joshi

Findings: Mitigating the Inconsistency Between Word Saliency and Model Confidence with Pathological Contrastive Training
Speaker: Pengwei Zhan

Long: The Paradox of the Compositionality of Natural Language: A Neural Machine Translation Case Study
Speaker: Verna Dankers

Short: How Distributed are Distributed Representations? An Observation on the Locality of Syntactic Information in Verb Agreement Tasks
Speaker: Bingzhi Li

Long: Flooding-X: Improving BERT's Resistance to Adversarial Attacks via Loss-Restricted Fine-Tuning
Speaker: Qin Liu

Long: Finding Structural Knowledge in Multimodal-BERT
Speaker: Victor Milewski

Long: Can Prompt Probe Pretrained Language Models? Understanding the Invisible Risks from a Causal View
Speaker: Boxi Cao

Long: Logic Traps in Evaluating Attribution Scores
Speaker: Yiming Ju

Findings: Exploring the Impact of Negative Samples of Contrastive Learning: A Case Study of Sentence Embedding
Speaker: Rui Cao

Long: Toward Interpretable Semantic Textual Similarity via Optimal Transport-based Contrastive Sentence Learning
Speaker: Seonghyeon Lee

Findings: The Inefficiency of Language Models in Scholarly Retrieval: An Experimental Walk-through
Speaker: Shruti Singh

Long: Can Explanations Be Useful for Calibrating Black Box Models?
Speaker: Xi Ye

Long: An Empirical Study of Memorization in NLP
Speaker: Xiaosen Zheng

Short: When classifying grammatical role, BERT doesn't care about word order... except when it matters
Speaker: Isabel Papadimitriou

Long: Interpreting Character Embeddings With Perceptual Representations: The Case of Shape, Sound, and Color
Speaker: Sidsel Boldsen

Long: Word Order Does Matter and Shuffled Language Models Know It
Speaker: Mostafa Abdou

Long: An Empirical Study on Explanations in Out-of-Domain Settings
Speaker: George Chrysostomou

Long: MPII: Multi-Level Mutual Promotion for Inference and Interpretation
Speaker: Yan Liu

Short: On the Effect of Isotropy on VAE Representations of Text
Speaker: Lan Zhang

Long: Memorisation versus Generalisation in Pre-trained Language Models
Speaker: Michael Tanzer

Short: How reparametrization trick broke differentially-private text representation learning
Speaker: nan

Long: Can Pre-trained Language Models Interpret Similes as Smart as Human?
Speaker: Qianyu He

Outstanding Paper: Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity
Speaker: Yao Lu

Findings: Probing BERT's priors with serial reproduction chains
Speaker: Takateru Yamakoshi

Long: The Grammar-Learning Trajectories of Neural Language Models
Speaker: Leshem Choshen

Long: Learning Disentangled Representations of Negation and Uncertainty
Speaker: Jake Vasilakes

Long: Knowledge Neurons in Pretrained Transformers
Speaker: Damai Dai

Findings: On Length Divergence Bias in Textual Matching Models
Speaker: Lan Jiang

Long: Causal Probing for Grammatical Number: From Encoding to Usage
Speaker: Karim Lasri

TACL: Word Acquisition in Neural Language Models
Speaker: Tyler Chang

TACL: Evaluating Explanations: How Much do Explanations from the Teacher aid Students?
Speaker: Danish Pruthi

VPS1: Language Groundings, Speech and Multimodality

07:30-08:30 (GatherTown)

Long: Cross-Utterance Conditioned VAE for Non-Autoregressive Text-to-Speech
Speaker: Yang Li

Short: Voxel-informed Language Grounding
Speaker: Rodolfo Corona Rodriguez

Long: Multimodal fusion via cortical network inspired losses
Speaker: Shiv Shankar

Long: Modeling Temporal-Modal Entity Graph for Procedural Multimodal Machine Comprehension
Speaker: Huibin Zhang

Long: Searching for fingerspelled content in American Sign Language
Speaker: Bowen Shi

Long: Skill Induction and Planning with Latent Language
Speaker: Pratyusha Sharma

Short: Sample, Translate, Recombine: Leveraging Audio Alignments for Data Augmentation in End-to-end Speech Translation
Speaker: Tsz Kin Lam

Long: Multi-Modal Sarcasm Detection via Cross-Modal Graph Convolutional Network
Speaker: Bin Liang, Bin Liang

Findings: End-to-End Speech Translation for Code Switched Speech
Speaker: Orion Weller

Long: Leveraging Visual Knowledge in Language Tasks: An Empirical Study on Intermediate Pre-training for Cross-Modal Knowledge Transfer
Speaker: Woojeong Jin

Long: A Good Prompt Is Worth Millions of Parameters: Low-resource Prompt-based Learning for Vision-Language Models
Speaker: Woojeong Jin

Long: FIBER: Fill-in-the-Blanks as a Challenging Video Understanding Evaluation Framework
Speaker: Santiago Castro Serra

Long: Cross-Modal Discrete Representation Learning
Speaker: Alexander Liu

Long: Contrastive Visual Semantic Pretraining Magnifies the Semantics of Natural Language Representations
Speaker: Robert Wolfe

Long: Image Retrieval from Contextual Descriptions
Speaker: Benno Krojer

Findings: Co-VQA : Answering by Interactive Sub Question Sequence
Speaker: Ruonan Wang

Short: XDBERT: Distilling Visual Information to BERT from Cross-Modal Systems to Improve Language Understanding
Speaker: Chan-Jan Hsu

Findings: DU-VLG: Unifying Vision-and-Language Generation via Dual Sequence-to-Sequence Pre-training
Speaker: Luyang Huang

Long: Visual-Language Navigation Pretraining via Prompt-based Environmental Self-exploration
Speaker: Xiwen Liang

Long: ReCLIP: A Strong Zero-Shot Baseline for Referring Expression Comprehension
Speaker: Sanjay Subramanian

Long: HOLM: Hallucinating Objects with Language Models for Referring Expression Recognition in Partially-Observed Scenes
Speaker: Volkan Cirik

Long: SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing
Speaker: Junyi Ao

Findings: UNIMO-2: End-to-End Unified Vision-Language Grounded Learning
Speaker: Wei Li

Long: On Vision Features in Multimodal Machine Translation
Speaker: Bei Li

Long: CARETS: A Consistency And Robustness Evaluative Test Suite for VQA
Speaker: Carlos Jimenez

Long: Analyzing Generalization of Vision and Language Navigation to Unseen Outdoor Areas
Speaker: Raphael Schumann

Long: LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding
Speaker: Jiapeng Wang

Long: Learning the Beauty in Songs: Neural Singing Voice Beautifier
Speaker: Jinglin Liu

Short: Understanding Game-Playing Agents with Natural Language Annotations
Speaker: Nicholas Tomlin

Long: Revisiting Over-Smoothness in Text to Speech
Speaker: Yi Ren

Long: VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena
Speaker: Letitia Parcalabescu

Long: SUPERB-SG: Enhanced Speech processing Universal PERformance Benchmark for Semantic and Generative Capabilities
Speaker: Hsiang-Sheng Tsai

Findings: Attention as Grounding: Exploring Textual and Cross-Modal Attention on Entities and Relations in Language-and-Vision Transformer
Speaker: Nikolai Ilinykh

Long: Inferring Rewards from Language in Context
Speaker: Jessy Lin

Long: End-to-End Modeling via Information Tree for One-Shot Natural Language Spatial Video Grounding
Speaker: Mengze Li

TACL: Word Representation Learning in Multimodal Pre-Trained Transformers: An Intrinsic Evaluation
Speaker: Sandro Pezzelle

TACL: Retrieve Fast, Rerank Smart: Cooperative and Joint Approaches for Improved Cross-Modal Retrieval
Speaker: Gregor Geigle

VPS1: Linguistic Theories, Cognitive Modeling and Psycholinguistics

07:30-08:30 (GatherTown)

Short: Analyzing Wrap-Up Effects through an Information-Theoretic Lens
Speaker: Clara Meister

Long: Metaphors in Pre-Trained Language Models: Probing and Generalization Across Datasets and Languages
Speaker: Ehsan Aghazadeh

Long: Decoding Part-of-Speech from Human EEG Signals
Speaker: Alex Murphy

Long: Speaker Information Can Guide Models to Better Inductive Biases: A Case Study On Predicting Code-Switching
Speaker: Alissa Ostapenko

Long: Learning Functional Distributional Semantics with Visual Data
Speaker: Yinhong Liu

Long: Word Segmentation as Unsupervised Constituency Parsing
Speaker: Raquel G. Alhama

Long: Do Transformer Models Show Similar Attention Patterns to Task-Specific Human Gaze?
Speaker: Oliver Eberle

Long: Context Matters: A Pragmatic Study of PLMs' Negation Understanding
Speaker: Reto Gubelmann

Long: Neural reality of argument structure constructions
Speaker: Bai Li

Long: Do self-supervised speech models develop human-like perception biases?
Speaker: Juliette Millet

TACL: Quantifying Cognitive Factors in Lexical Decline
Speaker: Ella Rabinovich

TACL: A Biologically Plausible Parser
Speaker: Daniel Mitropolsky

SRW-157
Speaker: nan

VPS1: Machine Learning for NLP

07:30-08:30 (GatherTown)

Long: AdapLeR: Speeding up Inference by Adaptive Length Reduction
Speaker: Ali Modarressi

Findings: Input-specific Attention Subnetworks for Adversarial Detection
Speaker: Emil Biju

Long: GLM: General Language Model Pretraining with Autoregressive Blank Infilling
Speaker: Zhengxiao Du

Long: Learn to Adapt for Generalized Zero-Shot Text Classification
Speaker: Yiwen Zhang

Long: Improving Meta-learning for Low-resource Text Classification and Generation via Memory Imitation
Speaker: Yingxiu Zhao

Long: Early Stopping Based on Unlabeled Samples in Text Classification
Speaker: HongSeok Choi

Long: Prompt-Based Rule Discovery and Boosting for Interactive Weakly-Supervised Learning
Speaker: Rongzhi Zhang

Short: P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks
Speaker: Xiao Liu

Short: The Power of Prompt Tuning for Low-Resource Semantic Parsing
Speaker: Nathan Schucher

Long: Structured Pruning Learns Compact and Accurate Models
Speaker: Mengzhou Xia

Long: Domain Knowledge Transferring for Pre-trained Language Model via Calibrated Activation Boundary Distillation
Speaker: Dongha Choi

Long: Distributionally Robust Finetuning BERT for Covariate Drift in Spoken Language Understanding
Speaker: Samuel Broscheit

Long: Enhancing Chinese Pre-trained Language Model via Heterogeneous Linguistics Graph
Speaker: Yanzeng Li

Long: Word2Box: Capturing Set-Theoretic Semantics of Words using Box Embeddings
Speaker: Shib Dasgupta

Short: LM-BFF-MS: Improving Few-Shot Fine-tuning of Language Models based on Multiple Soft Demonstration Memory
Speaker: Eunhwan Park

Long: Learning Disentangled Textual Representations via Statistical Measures of Similarity
Speaker: Pierre Colombo

Long: Efficient Hyper-parameter Search for Knowledge Graph Embedding
Speaker: Yongqi Zhang

Findings: Metadata Shaping: A Simple Approach for Knowledge-Enhanced Language Models
Speaker: Simran Arora

Findings: Enhancing Natural Language Representation with Large-Scale Out-of-Domain Commonsense
Speaker: Wanyun Cui

Long: Leveraging Relaxed Equilibrium by Lazy Transition for Sequence Modeling
Speaker: Xi Ai

Long: On Continual Model Refinement in Out-of-Distribution Data Streams
Speaker: Bill Yuchen Lin

Long: Imputing Out-of-Vocabulary Embeddings with LOVE Makes Language Models Robust with Little Cost
Speaker: Lihu Chen

Short: Revisiting the Compositional Generalization Abilities of Neural Sequence Models
Speaker: Arkil Patel

Long: Generative Pretraining for Paraphrase Evaluation
Speaker: Jack Weston

Findings: A Simple Hash-Based Early Exiting Approach For Language Understanding and Generation
Speaker: Tianxiang Sun, Tianxiang Sun

Findings: Syntax-guided Contrastive Learning for Pre-trained Language Model
Speaker: Shuai Zhang

Outstanding Paper: Compression of Generative Pre-trained Language Models via Quantization
Speaker: Chaofan Tao

Long: SPoT: Better Frozen Model Adaptation through Soft Prompt Transfer
Speaker: Tu Vu

Findings: ELLE: Efficient Lifelong Pre-training for Emerging Data
Speaker: Yujia Qin

Long: E-LANG: Energy-Based Joint Inferencing of Super and Swift Language Models
Speaker: Mohammad Akbari

Long: Fully Hyperbolic Neural Networks
Speaker: Weize Chen

Short: DMix: Adaptive Distance-aware Interpolative Mixup
Speaker: Ramit Sawhney

Long: RotateQVS: Representing Temporal Information as Rotations in Quaternion Vector Space for Temporal Knowledge Graph Completion
Speaker: Kai Chen

Findings: Word-level Perturbation Considering Word Length and Compositional Subwords
Speaker: Tatsuya Hiraoka

Long: SHIELD: Defending Textual Neural Networks against Multiple Black-Box Adversarial Attacks with Stochastic Multi-Expert Patcher
Speaker: Thai Le

Findings: Prompt Tuning for Discriminative Pre-trained Language Models
Speaker: Yuan Yao

Long: Prototypical Verbalizer for Prompt-based Few-shot Tuning
Speaker: ganqu cui

Long: BERT Learns to Teach: Knowledge Distillation with Meta Learning
Speaker: Wangchunshu Zhou

Long: StableMoE: Stable Routing Strategy for Mixture of Experts
Speaker: Damai Dai

Short: Contrastive Learning-Enhanced Nearest Neighbor Mechanism for Multi-Label Text Classification
Speaker: Xi'ao Su

Findings: Improving the Adversarial Robustness of NLP Models by Information Bottleneck
Speaker: Cenyuan Zhang

Short: NoisyTune: A Little Noise Can Help You Finetune Pretrained Language Models Better
Speaker: Chuhan Wu

Long: Transkimmer: Transformer Learns to Layer-wise Skim
Speaker: Yue Guan, Yue Guan

Long: SkipBERT: Efficient Inference with Shallow Layer Skipping
Speaker: Jue Wang

Long: mLUKE: The Power of Entity Representations in Multilingual Pretrained Language Models
Speaker: Ryokan Ri

Long: ABC: Attention with Bounded-memory Control
Speaker: Hao Peng

Long: Cluster & Tune: Boost Cold Start Performance in Text Classification
Speaker: Eyal Shnarch

Long: That Is a Suspicious Reaction!: Interpreting Logits Variation to Detect NLP Adversarial Attacks
Speaker: Edoardo Mosca

Short: A Flexible Multi-Task Model for BERT Serving
Speaker: Tianwen Wei

Long: LinkBERT: Pretraining Language Models with Document Links
Speaker: Michihiro Yasunaga

Findings: Attention Mechanism with Energy-Friendly Operations
Speaker: Yu Wan

Long: Coherence boosting: When your pretrained language model is not paying enough attention
Speaker: Nikolay Malkin

Long: Uncertainty Estimation of Transformer Predictions for Misclassification Detection
Speaker: Artem Vazhentsev

Long: Deduplicating Training Data Makes Language Models Better
Speaker: Katherine Lee

Long: Sparsifying Transformer Models with Trainable Representation Pooling
Speaker: Michał Pietruszka

Long: Uncertainty Determines the Adequacy of the Mode and the Tractability of Decoding in Sequence-to-Sequence Models
Speaker: Felix Stahlberg

Long: FlipDA: Effective and Robust Data Augmentation for Few-Shot Learning
Speaker: Jing Zhou

TACL: Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP
Speaker: Timo Schick

TACL: CANINE: Pre-training an Efficient Tokenization-Free Encoder for Language Representation
Speaker: Jonathan Clark

TACL: Towards General Natural Language Understanding with Probabilistic Worldbuilding
Speaker: Abulhair Saparov

VPS1: Machine Translation and Multilinguality

07:30-08:30 (GatherTown)

Long: Towards Making the Most of Cross-Lingual Transfer for Zero-Shot Neural Machine Translation
Speaker: Guanhua Chen

Long: Learning When to Translate for Streaming Speech
Speaker: Qianqian Dong

Short: On Efficiently Acquiring Annotations for Multilingual Models
Speaker: Joel Ruben Antony Moniz

Long: Language-agnostic BERT Sentence Embedding
Speaker: Fangxiaoyu Feng

Findings: Automatic Song Translation for Tonal Languages
Speaker: Fenfei Guo

Long: Match the Script, Adapt if Multilingual: Analyzing the Effect of Multilingual Pretraining on Cross-lingual Transferability
Speaker: Yoshinari Fujinuma

Long: Composable Sparse Fine-Tuning for Cross-Lingual Transfer
Speaker: Alan Ansell

Long: Enhancing Cross-lingual Natural Language Inference by Prompt-learning from Cross-lingual Templates
Speaker: Kunxun Qi

Long: Overcoming Catastrophic Forgetting beyond Continual Learning: Balanced Training for Neural Machine Translation
Speaker: Chenze Shao

Long: Learning Confidence for Transformer-based Neural Machine Translation
Speaker: Yu Lu

Long: Conditional Bilingual Mutual Information Based Adaptive Training for Neural Machine Translation
Speaker: Songming Zhang

Long: Modeling Dual Read/Write Paths for Simultaneous Machine Translation
Speaker: Shaolei Zhang

Long: Understanding and Improving Sequence-to-Sequence Pretraining for Neural Machine Translation
Speaker: Wenxuan Wang

Long: MSCTD: A Multimodal Sentiment Chat Translation Dataset
Speaker: Yunlong Liang

Findings: DaLC: Domain Adaptation Learning Curve Prediction for Neural Machine Translation
Speaker: Cheonbok Park

Long: Investigating Failures of Automatic Translation in the Case of Unambiguous Gender
Speaker: Adi Renduchintala

Long: Multilingual Document-Level Translation Enables Zero-Shot Transfer From Sentences to Documents
Speaker: Biao Zhang

Long: Scheduled Multi-task Learning for Neural Chat Translation
Speaker: Yunlong Liang

Long: Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder Translation Models
Speaker: Lorenzo Lupo

Short: As Little as Possible, as Much as Necessary: Detecting Over- and Undertranslations with Contrastive Conditioning
Speaker: Jannis Vamvas

Long: Cross-Lingual Ability of Multilingual Masked Language Models: A Study of Language Structure
Speaker: Yuan Chai

Findings: Prompt-Driven Neural Machine Translation
Speaker: Yafu Li

Long: Can Synthetic Translations Improve Bitext Quality?
Speaker: Eleftheria Briakou

Short: S^4-Tuning: A Simple Cross-lingual Sub-network Tuning Method
Speaker: Runxin Xu

Long: Multi Task Learning For Zero Shot Performance Prediction of Multilingual Models
Speaker: Kabir Ahuja

Long: Neural Machine Translation with Phrase-Level Universal Visual Representations
Speaker: Qingkai Fang

Long: MSP: Multi-Stage Prompting for Making Pre-trained Language Models Better Translators
Speaker: Zhixing Tan

Long: XLM-E: Cross-lingual Language Model Pre-training via ELECTRA
Speaker: Zewen Chi

Long: Universal Conditional Masked Language Pre-training for Neural Machine Translation
Speaker: Pengfei Li

Long: Bridging the Data Gap between Training and Inference for Unsupervised Neural Machine Translation
Speaker: Zhiwei He

Long: Accurate Online Posterior Alignments for Principled Lexically-Constrained Decoding
Speaker: Soumya Chatterjee

Long: STEMM: Self-learning with Speech-text Manifold Mixup for Speech Translation
Speaker: Qingkai Fang

Long: Integrating Vectorized Lexical Constraints for Neural Machine Translation
Speaker: Shuo Wang

Findings: First the Worst: Finding Better Gender Translations During Beam Search
Speaker: Danielle Saunders

Long: Learning Adaptive Segmentation Policy for End-to-End Simultaneous Translation
Speaker: Ruiqing Zhang

Outstanding Paper: Learning to Generalize to More: Continuous Semantic Augmentation for Neural Machine Translation
Speaker: Xiangpeng Wei

Long: UniTE: Unified Translation Evaluation
Speaker: Yu Wan

Long: EAG: Extract and Generate Multi-way Aligned Corpus for Complete Multi-lingual Neural Machine Translation
Speaker: Yulin Xu

Long: ODE Transformer: An Ordinary Differential Equation-Inspired Model for Sequence Generation
Speaker: Bei Li

Findings: CrossAligner & Co: Zero-Shot Transfer Methods for Task-Oriented Cross-lingual Natural Language Understanding
Speaker: Milan Gritta

Findings: Structural Supervision for Word Alignment and Machine Translation
Speaker: Lei Li

TACL: Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation
Speaker: Markus Freitag

TACL: Samanantar: The Largest Publicly Available Parallel Corpora Collection For 11 Indic Languages
Speaker: Gowtham Ramesh

TACL: Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets
Speaker: Julia Kreutzer

TACL: ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models
Speaker: Linting Xue

TACL: The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation
Speaker: Naman Goyal

SRW: Restricted or Not: A General Training Framework for Neural Machine Translation
Speaker: Zuchao Li

VPS1: NLP Applications

07:30-08:30 (GatherTown)

Long: Towards Comprehensive Patent Approval Predictions:Beyond Traditional Document Classification
Speaker: Xiaochen Gao

Long: Controllable Dictionary Example Generation: Generating Example Sentences for Specific Targeted Audiences
Speaker: Xingwei He

Findings: Reinforced Cross-modal Alignment for Radiology Report Generation
Speaker: Han Qin

Long: Legal Judgment Prediction via Event Extraction with Constraints
Speaker: Yi Feng

Findings: Cross-Modal Cloze Task: A New Task to Brain-to-Word Decoding
Speaker: Shuxian Zou

Short: Automatic Detection of Entity-Manipulated Text using Factual Knowledge
Speaker: Ganesh Jawahar, Ganesh Jawahar

Long: FORTAP: Using Formulas for Numerical-Reasoning-Aware Table Pretraining
Speaker: Zhoujun Cheng

Short: Buy Tesla, Sell Ford: Assessing Implicit Stock Market Preference in Pre-trained Language Models
Speaker: Cheng Yu Chuang

Long: Differentiable Multi-Agent Actor-Critic for Multi-Step Radiology Report Summarization
Speaker: Sanjeev Kumar Karn

Long: Few-Shot Tabular Data Enrichment Using Fine-Tuned Transformer Architectures
Speaker: Asaf Harari

Long: Exploring and Adapting Chinese GPT to Pinyin Input Method
Speaker: Minghuan Tan, Minghuan Tan

Long: Predicting Intervention Approval in Clinical Trials through Multi-Document Summarization
Speaker: Georgios Katsimpras

Findings: A Novel Framework Based on Medical Concept Driven Attention for Explainable Medical Code Prediction via External Knowledge
Speaker: Tao Wang

Long: CAKE: A Scalable Commonsense-Aware Framework For Multi-View Knowledge Graph Completion
Speaker: Guanglin Niu

Long: KenMeSH: Knowledge-enhanced End-to-end Biomedical Text Labelling
Speaker: Xindi Wang

Long: PromDA: Prompt-based Data Augmentation for Low-Resource NLU Tasks
Speaker: Yufei Wang

Long: FairLex: A Multilingual Benchmark for Evaluating Fairness in Legal Text Processing
Speaker: Ilias Chalkidis

Long: Fine- and Coarse-Granularity Hybrid Self-Attention for Efficient BERT
Speaker: Jing Zhao

Findings: Extracting Person Names from User Generated Text: Named-Entity Recognition for Combating Human Trafficking
Speaker: Yifei Li

Long: Improving Generalizability in Implicitly Abusive Language Detection with Concept Activation Vectors
Speaker: Isar Nejadgholi

Findings: CRASpell: A Contextual Typo Robust Approach to Improve Chinese Spelling Correction
Speaker: Shulin Liu

Long: What does the sea say to the shore? A BERT based DST style approach for speaker to dialogue attribution in novels
Speaker: Animesh Prasad

Long: Continual Pre-training of Language Models for Math Problem Understanding with Syntax-Aware Memory Network
Speaker: Zheng Gong

Short: HYPHEN: Hyperbolic Hawkes Attention For Text Streams
Speaker: Shivam Agarwal

Long: Learning to Reason Deductively: Math Word Problem Solving as Complex Relation Extraction
Speaker: Zhanming Jie

Findings: Type-Driven Multi-Turn Corrections for Grammatical Error Correction
Speaker: Shaopeng Lai

Long: ReACC: A Retrieval-Augmented Code Completion Framework
Speaker: Shuai Lu

Long: The AI Doctor Is In: A Survey of Task-Oriented Dialogue Systems for Healthcare Applications
Speaker: Mina Valizadeh

Long: Letters From the Past: Modeling Historical Sound Change Through Diachronic Character Embeddings
Speaker: Sidsel Boldsen

Long: Clickbait Spoiling via Question Answering and Passage Retrieval
Speaker: Matthias Hagen

Long: Interpretability for Language Learners Using Example-Based Grammatical Error Correction
Speaker: Masahiro Kaneko

Long: UniXcoder: Unified Cross-Modal Pre-training for Code Representation
Speaker: Daya Guo

Long: UniXcoder: Unified Cross-Modal Pre-training for Code Representation
Speaker: Daya Guo

Long: Adversarial Authorship Attribution for Deobfuscation
Speaker: Wanyue Zhai

Long: On the Robustness of Offensive Language Classifiers
Speaker: Jonathan Rusert

Findings: Unsupervised Preference-Aware Language Identification
Speaker: Xingzhang Ren

CL: Linguistic Parameters of Spontaneous Speech for identifying Mild Cognitive Impairment and Alzheimer's Disease
Speaker: Veronika Vincze

Long: Your Answer is Incorrect... Would you like to know why? Introducing a Bilingual Short Answer Feedback Dataset
Speaker: Anna Filighera

Long: RNSum: A Large-Scale Dataset for Automatic Release Note Generation via Commit Logs Summarization
Speaker: Hisashi Kamezawa

Long: Modeling Persuasive Discourse to Adaptively Support Students' Argumentative Writing
Speaker: Thiemo Wambsganss

TACL: A Neighbourhood Framework for Resource-Lean Content Flagging
Speaker: Momchil Hardalov

CL: Novelty Detection: A Perspective from Natural Language Processing
Speaker: Tirthankar Ghosal

VPS1: Phonology, Morphology and Word Segmentation

07:30-08:30 (GatherTown)

Short: An Embarrassingly Simple Method to Mitigate Undesirable Properties of Pretrained Language Model Tokenizers
Speaker: Valentin Hofmann

Short: WLASL-LEX: a Dataset for Recognising Phonological Properties in American Sign Language
Speaker: Federico Tavella

Findings: More Than Words: Collocation Retokenization for Latent Dirichlet Allocation Models
Speaker: Jin Cheevaprawatdomrong

VPS1: Question Answering

07:30-08:30 (GatherTown)

Long: Learning to Imagine: Integrating Counterfactual Thinking in Neural Discrete Reasoning
Speaker: Moxin Li

Long: Hypergraph Transformer: Weakly-Supervised Multi-hop Reasoning for Knowledge-based Visual Question Answering
Speaker: Yu-Jung Heo

Long: KaFSP: Knowledge-Aware Fuzzy Semantic Parsing for Conversational Question Answering over a Large-Scale Knowledge Base
Speaker: Junzhuo Li

Findings: Fact-Tree Reasoning for N-ary Question Answering over Knowledge Graphs
Speaker: YAO ZHANG

Long: Tracing Origins: Coreference-aware Machine Reading Comprehension
Speaker: Zhuosheng Zhang

Short: Predicting Difficulty and Discrimination of Natural Language Questions
Speaker: Matthew Byrd

Findings: Answer Uncertainty and Unanswerability in Multiple-Choice Machine Reading Comprehension
Speaker: Vatsal Raina

Long: Open Domain Question Answering with A Unified Knowledge Interface
Speaker: Kaixin Ma

Long: Retrieval-guided Counterfactual Generation for QA
Speaker: Bhargavi Paranjape

Long: On the Robustness of Question Rewriting Systems to Questions of Varying Hardness
Speaker: Hai Ye

Findings: Logic-Driven Context Extension and Data Augmentation for Logical Reasoning of Text
Speaker: Siyuan Wang

Long: Sequence-to-Sequence Knowledge Graph Completion and Question Answering
Speaker: Apoorv Saxena

Long: Automated Crossword Solving
Speaker: Eric Wallace

Findings: Relevant CommonSense Subgraphs for "What if..." Procedural Reasoning
Speaker: Chen Zheng

Long: Generated Knowledge Prompting for Commonsense Reasoning
Speaker: Jiacheng Liu

Long: ConditionalQA: A Complex Reading Comprehension Dataset with Conditional Answers
Speaker: William Cohen

Findings: BBQ: A hand-built bias benchmark for question answering
Speaker: Alicia Parrish

Short: A Copy-Augmented Generative Model for Open-Domain Question Answering
Speaker: shuang liu

Long: MMCoQA: Conversational Question Answering over Text, Tables, and Images
Speaker: Yongqi Li

Long: KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering
Speaker: Donghan Yu

Long: Feeding What You Need by Understanding What You Learned
Speaker: Xiaoqiang Wang

Long: RNG-KBQA: Generation Augmented Iterative Ranking for Knowledge Base Question Answering
Speaker: Xi Ye

Findings: MERIt: Meta-Path Guided Contrastive Learning for Logical Reasoning
Speaker: Fangkai Jiao

Long: AdaLoGN: Adaptive Logic Graph Network for Reasoning-Based Machine Reading Comprehension
Speaker: Xiao Li

Long: Program Transfer for Answering Complex Questions over Knowledge Bases
Speaker: Shulin Cao

TACL: Break, Perturb, Build: Automatic Perturbation of Reasoning Paths Through Question Decomposition
Speaker: Mor Geva

TACL: Time-Aware Language Models as Temporal Knowledge Bases
Speaker: Bhuwan Dhingra

VPS1: Resources and Evaluation

07:30-08:30 (GatherTown)

Long: Quantified Reproducibility Assessment of NLP Results
Speaker: Anya Belz

Long: AlephBERT: Language Model Pre-training and Evaluation from Sub-Word to Sentence Level
Speaker: Amit Seker

Long: QuoteR: A Benchmark of Quote Recommendation for Writing
Speaker: Fanchao Qi

Long: FewNLU: Benchmarking State-of-the-Art Methods for Few-Shot Natural Language Understanding
Speaker: Yanan Zheng

Long: Premise-based Multimodal Reasoning: Conditional Inference on Joint Textual and Visual Clues
Speaker: Qingxiu Dong

Long: HiTab: A Hierarchical Table Dataset for Question Answering and Natural Language Generation
Speaker: Zhoujun Cheng

Findings: Mukayese: Turkish NLP Strikes Back
Speaker: Ali Safaya

Long: Sense Embeddings are also Biased – Evaluating Social Biases in Static and Contextualised Sense Embeddings
Speaker: Yi Zhou

Long: IAM: A Comprehensive and Large-Scale Dataset for Integrated Argument Mining Tasks
Speaker: Liying Cheng

Long: A Taxonomy of Empathetic Questions in Social Dialogs
Speaker: Ekaterina Svikhnushina

Short: k-Rater Reliability: The Correct Unit of Reliability for Aggregated Human Annotations
Speaker: Ka Wong

Long: ILDAE: Instance-Level Difficulty Analysis of Evaluation Data
Speaker: Neeraj Varshney

Long: Cross-Task Generalization via Natural Language Crowdsourcing Instructions
Speaker: Swaroop Mishra

Long: NumGLUE: A Suite of Fundamental yet Challenging Mathematical Reasoning Tasks
Speaker: Swaroop Mishra

Long: VALUE: Understanding Dialect Disparity in NLU
Speaker: Caleb Ziems

Long: LexGLUE: A Benchmark Dataset for Legal Language Understanding in English
Speaker: Ilias Chalkidis

Best Resource: DiBiMT: A Novel Benchmark for Measuring Word Sense Disambiguation Biases in Machine Translation
Speaker: Niccolò Campolungo

Long: BenchIE: A Framework for Multi-Faceted Fact-Based Open Information Extraction Evaluation
Speaker: Kiril Gashteovski

Long: Identifying Moments of Change from Longitudinal User Text
Speaker: Adam Tsakalidis

Long: Impact of Evaluation Methodologies on Code Summarization
Speaker: Pengyu Nie

Long: FaVIQ: FAct Verification from Information-seeking Questions
Speaker: Jungsoo Park

Long: RoMe: A Robust Metric for Evaluating Natural Language Generation
Speaker: Md Rashad Al Hasan Rony

Short: PriMock57: A Dataset Of Primary Care Mock Consultations
Speaker: Alex Papadopoulos Korfiatis

Long: Human Evaluation and Correlation with Automatic Metrics in Consultation Note Generation
Speaker: Francesco Moramarco

Long: Evaluating Extreme Hierarchical Multi-label Classification
Speaker: Enrique Amigó

Long: KQA Pro: A Dataset with Explicit Compositional Programs for Complex Question Answering over Knowledge Base
Speaker: Shulin Cao

Long: Learning to Rank Visual Stories From Human Ranking Data
Speaker: Chi-Yang Hsu

Long: A Token-level Reference-free Hallucination Detection Benchmark for Free-form Text Generation
Speaker: Tianyu Liu

Long: A Statutory Article Retrieval Dataset in French
Speaker: Antoine Louis

Long: ParaDetox: Detoxification with Parallel Data
Speaker: Varvara Logacheva

Long: TwittIrish: A Universal Dependencies Treebank of Tweets in Modern Irish
Speaker: Lauren Cassidy

Findings: HLDC: Hindi Legal Documents Corpus
Speaker: Arnav Kapoor

Long: RELiC: Retrieving Evidence for Literary Claims
Speaker: Katherine Thai

Outstanding Paper: Active Evaluation: Efficient NLG Evaluation with Few Pairwise Comparisons
Speaker: Akash Kumar Mohankumar

TACL: Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics
Speaker: Paula Czarnowska, Paula Czarnowska

TACL: Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations
Speaker: Aida Mostafazadeh Davani

VPS1: Semantics

07:30-08:30 (GatherTown)

Findings: Going "Deeper": Structured Sememe Prediction via Transformer with Tree Attention
Speaker: Yining Ye

Findings: Table-based Fact Verification with Self-adaptive Mixture of Experts
Speaker: Yuxuan Zhou

Long: FaiRR: Faithful and Robust Deductive Reasoning over Natural Language
Speaker: Soumya Sanyal

Long: Principled Paraphrase Generation with Parallel Corpora
Speaker: Aitor Ormazabal

Long: Leveraging Similar Users for Personalized Language Modeling with Limited Data
Speaker: Charles Welch

Findings: S^2SQL: Injecting Syntax to Question-Schema Interaction Graph Encoder for Text-to-SQL Parsers
Speaker: Binyuan Huii

Long: ExtEnD: Extractive Entity Disambiguation
Speaker: Edoardo Barba, Edoardo Barba

Short: Exploiting Language Model Prompts Using Similarity Measures: A Case Study on the Word-in-Context Task
Speaker: Mohsen Tabasi

Long: Generating Data to Mitigate Spurious Correlations in Natural Language Inference Datasets
Speaker: Yuxiang Wu

Long: Improving Event Representation via Simultaneous Weakly Supervised Contrastive Learning and Clustering
Speaker: Jun Gao

Long: Right for the Right Reason: Evidence Extraction for Trustworthy Tabular Reasoning
Speaker: Vivek Gupta

Long: LAGr: Label Aligned Graphs for Better Systematic Generalization in Semantic Parsing
Speaker: Dora Jambor

Long: Improving Word Translation via Two-Stage Contrastive Learning
Speaker: Yaoyiran Li

Findings: ASCM: An Answer Space Clustered Prompting Method without Answer Engineering
Speaker: Zhen Wang

Findings: Seeking Patterns, Not just Memorizing Procedures: Contrastive Learning for Solving Math Word Problems
Speaker: Zhongli Li

Long: EPT-X: An Expression-Pointer Transformer model that generates eXplanations for numbers
Speaker: Bugeun Kim

Long: Probing for Predicate Argument Structures in Pretrained Language Models
Speaker: Simone Conia

Findings: Lacking the Embedding of a Word? Look it up into a Traditional Dictionary
Speaker: Elena Sofia Ruzzetti

Short: Fire Burns, Sword Cuts: Commonsense Inductive Bias for Exploration in Text-based Games
Speaker: Dongwon Kelvin Ryu

Long: Bridging the Generalization Gap in Text-to-SQL Parsing with Schema Expansion
Speaker: Chen Zhao

Long: Predicate-Argument Based Bi-Encoder for Paraphrase Identification
Speaker: Qiwei Peng

Long: Entailment Graph Learning with Textual Entailment and Soft Transitivity
Speaker: Zhibin Chen

Long: Graph Pre-training for AMR Parsing and Generation
Speaker: Xuefeng Bai

Long: Just Rank: Rethinking Evaluation with Word and Sentence Similarities
Speaker: Bin Wang

Long: Debiased Contrastive Learning of Unsupervised Sentence Representations
Speaker: Kun Zhou

Findings: Divide and Conquer: Text Semantic Matching with Disentangled Keywords and Intents
Speaker: Yicheng Zou

Long: Learning to Generate Programs for Table Fact Verification via Structure-Aware Semantic Parsing
Speaker: Suixin Ou

Long: Modeling Syntactic-Semantic Dependency Correlations in Semantic Role Labeling Using Mixture Models
Speaker: Junjie Chen

Long: Towards Better Characterization of Paraphrases
Speaker: Timothy Liu

TACL: Weisfeiler-Leman in the BAMBOO: Novel AMR Graph Metrics and a Benchmark for AMR Graph Similarity
Speaker: Juri Opitz

TACL: Sentence Similarity Based on Contexts
Speaker: Jiwei Li

TACL: It's not Rocket Science: Interpreting Figurative Language in Narratives
Speaker: Tuhin Chakrabarty

VPS1: Sentiment Analysis, Stylistic Analysis, and Argument Mining

07:30-08:30 (GatherTown)

Findings: Towards Unifying the Label Space for Aspect- and Sentence-based Sentiment Analysis
Speaker: Yiming Zhang

Long: JointCL: A Joint Contrastive Learning Framework for Zero-Shot Stance Detection
Speaker: Bin Liang, Bin Liang

Short: Have my arguments been replied to? Argument Pair Extraction as Machine Reading Comprehension
Speaker: Jianzhu Bao

Long: So Different Yet So Alike! Constrained Unsupervised Text Style Transfer
Speaker: Abhinav Ramesh Kashyap

Findings: Multi-Granularity Semantic Aware Graph Model for Reducing Position Bias in Emotion Cause Pair Extraction
Speaker: Yinan Bao

Long: Vision-Language Pre-Training for Multimodal Aspect-Based Sentiment Analysis
Speaker: Yan Ling

Findings: BiSyn-GAT+: Bi-Syntax Aware Graph Attention Network for Aspect-based Sentiment Analysis
Speaker: Shuo Liang

Long: Effective Token Graph Modeling using a Novel Labeling Strategy for Structured Sentiment Analysis
Speaker: Wenxuan Shi

Findings: ECO v1: Towards Event-Centric Opinion Mining
Speaker: Ruoxi Xu

Long: A Rationale-Centric Framework for Human-in-the-loop Machine Learning
Speaker: Jinghui Lu

Findings: Incorporating Dynamic Semantics into Pre-Trained Language Model for Aspect-based Sentiment Analysis
Speaker: Kai Zhang

Long: DoCoGen: Domain Counterfactual Generation for Low Resource Domain Adaptation
Speaker: Nitay Calderon

Long: Can Unsupervised Knowledge Transfer from Social Discussions Help Argument Mining?
Speaker: Subhabrata Dutta

Long: Fair and Argumentative Language Modeling for Computational Argumentation
Speaker: Carolin Holtermann

Long: The Moral Debater: A Study on the Computational Generation of Morally Framed Arguments
Speaker: Milad Alshomary

VPS1: Special Theme on Language Diversity: From Low Resource to Endangered

07:30-08:30 (GatherTown)

Long: Computational Historical Linguistics and Language Diversity in South Asia
Speaker: Aryaman Arora, Aryaman Arora

Long: How can NLP Help Revitalize Endangered Languages? A Case Study and Roadmap for the Cherokee Language
Speaker: Shiyue Zhang

Findings: OCR Improves Machine Translation for Low-Resource Languages
Speaker: Oana Ignat

Long: Interactive Word Completion for Plains Cree
Speaker: William Lane

Long: Dataset Geography: Mapping Language Data to Language Users
Speaker: FAHIM FAISAL

Short: Machine Translation for Livonian: Catering to 20 Speakers
Speaker: Matīss Rikters

Long: Phone-ing it in: Towards Flexible Multi-Modal Language Model Training by Phonetic Representations of Data
Speaker: Colin Leong

Long: Systematic Inequalities in Language Technology Performance across the World's Languages
Speaker: Damian Blasi

Long: From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology
Speaker: Mark Dingemanse

Short: Sub-Word Alignment is Still Useful: A Vest-Pocket Method for Enhancing Low-Resource Machine Translation
Speaker: Minhan Xu

Long: AmericasNLI: Evaluating Zero-shot Natural Language Understanding of Pretrained Multilingual Models in Truly Low-resource Languages
Speaker: Abteen Ebrahimi

Long: Challenges and Strategies in Cross-Cultural NLP
Speaker: Daniel Hershcovich, Daniel Hershcovich

Best Theme: Requirements and Motivations of Low-Resource Speech Synthesis for Language Revitalization
Speaker: Aidan Pine

Short: Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism
Speaker: Lane Schwartz

Long: Make the Best of Cross-lingual Transfer: Evidence from POS Tagging with over 100 Languages
Speaker: Wietse de Vries

Long: Local Languages, Third Spaces, and other High-Resource Scenarios
Speaker: Steven Bird

VPS1: Summarization

07:30-08:30 (GatherTown)

Long: Attention Temperature Matters in Abstractive Summarization Distillation
Speaker: Shengqiang Zhang

Long: Discriminative Marginalized Probabilistic Neural Method for Multi-Document Summarization of Medical Literature
Speaker: Gianluca Moro

Long: Neural Label Search for Zero-Shot Multi-Lingual Extractive Summarization
Speaker: Ruipeng Jia

Long: Unsupervised Extractive Opinion Summarization Using Sparse Coding
Speaker: Somnath Basu Roy Chowdhury

Long: A Well-Composed Text is Half Done! Composition Sampling for Diverse Conditional Generation
Speaker: Shashi Narayan

Long: Faithful or Extractive? On Mitigating the Faithfulness-Abstractiveness Trade-off in Abstractive Summarization
Speaker: Faisal Ladhak

Long: Other Roles Matter! Enhancing Role-Oriented Dialogue Summarization via Role Interactions
Speaker: Haitao Lin

Long: EntSUM: A Data Set for Entity-Centric Extractive Summarization
Speaker: Mounica Maddela

Long: SummaReranker: A Multi-Task Mixture-of-Experts Re-ranking Framework for Abstractive Summarization
Speaker: Mathieu Ravaut

Long: The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature
Speaker: Yulia Otmakhova

Long: PRIMERA: Pyramid-based Masked Sentence Pre-training for Multi-document Summarization
Speaker: Wen Xiao

Findings: Comparative Opinion Summarization via Collaborative Decoding
Speaker: Hayate Iso

Long: ASPECTNEWS: Aspect-Oriented Summarization of News Documents
Speaker: Ojas Ahuja

Long: Learning Non-Autoregressive Models from Search for Unsupervised Sentence Summarization
Speaker: Puyuan Liu

Findings: Should We Trust This Summary? Bayesian Abstractive Summarization to The Rescue
Speaker: Alexios Gidiotis

TACL: Planning with Learned Entity Prompts for Abstractive Summarization
Speaker: Shashi Narayan

TACL: SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization
Speaker: Philippe Laban

TACL: Document Summarization with Latent Queries
Speaker: Yumo Xu

VPS1: Syntax: Tagging, Chunking and Parsing

07:30-08:30 (GatherTown)

Findings: SyMCoM - Syntactic Measure of Code Mixing A Study Of English-Hindi Code-Mixing
Speaker: Prashant Kodali

Long: Semi-supervised Domain Adaptation for Dependency Parsing with Dynamic Matching Network
Speaker: Ying Li

Long: Investigating Non-local Features for Neural Constituency Parsing
Speaker: Leyang Cui

Long: Headed-Span-Based Projective Dependency Parsing
Speaker: Songlin Yang

Long: MELM: Data Augmentation with Masked Entity Language Modeling for Low-Resource NER
Speaker: Ran Zhou

Long: Bottom-Up Constituency Parsing and Nested Named Entity Recognition with Pointer Networks
Speaker: Songlin Yang

Findings: Auxiliary tasks to boost Biaffine Semantic Dependency Parsing
Speaker: Marie Candito

Long: Unsupervised Dependency Graph Network
Speaker: Yikang Shen

Findings: Cross-domain Named Entity Recognition via Graph Matching
Speaker: Junhao Zheng

Long: Semantic Composition with PSHRG for Derivation Tree Reconstruction from Graph-Based Meaning Representations
Speaker: Chun Hei Lo

Findings: Bridging Pre-trained Language Models and Hand-crafted Features for Unsupervised POS Tagging
Speaker: Houquan Zhou

Long: Phrase-aware Unsupervised Constituency Parsing
Speaker: Xiaotao Gu

Long: Probing for Labeled Dependency Trees
Speaker: Max Müller-Eberstein

Long: Meta-Learning for Fast Cross-Lingual Adaptation in Dependency Parsing
Speaker: Anna Langedijk

CL: Improved N-Best Extraction with an Evaluation on Language Data
Speaker: Johanna Björklund

Next Big Ideas Talks
09:00-10:30 - **Auditorium** (Auditorium)

Coffee Break
10:30-11:00 - **Auditorium** (Forum)

Session 4 - 11:00-12:30

Interpretability and Analysis of Models for NLP 2
11:00-12:30 (The Liffey A)

11:00-11:15 (The Liffey A)
Can Explanations Be Useful for Calibrating Black Box Models?
Xi Ye and Greg Durrett
NLP practitioners often want to take existing trained models and apply them to data from new domains. While fine-tuning or few-shot learning can be used to adapt a base model, there is no single recipe for making these techniques work; moreover, one may not have access to the original model weights if it is deployed as a black box. We study how to improve a black box model's performance on a new domain by leveraging explanations of the model's behavior. Our approach first extracts a set of features combining human intuition about the task with model attributions generated by black box interpretation techniques, then uses a simple calibrator, in the form of a classifier, to predict whether the base model was correct or not. We experiment with our method on two tasks, extractive question answering and natural language inference, covering adaptation from several pairs of domains with limited target-domain data. The experimental results across all the domain pairs show that explanations are useful for calibrating these models, boosting accuracy when predictions do not have to be returned on every example. We further show that the calibration model transfers to some extent between tasks.

11:15-11:30 (The Liffey A)
An Empirical Study on Explanations in Out-of-Domain Settings
George Chrysostomou and Nikolaos Aletras
Recent work in Natural Language Processing has focused on developing approaches that extract faithful explanations, either via identifying the most important tokens in the input (i.e. post-hoc explanations) or by designing inherently faithful models that first select the most important tokens and then use them to predict the correct label (i.e. select-then-predict models). Currently, these approaches are largely evaluated on in-domain settings. Yet, little is known about how post-hoc explanations and inherently faithful models perform in out-of-domain settings. In this paper, we conduct an extensive empirical study that examines: (1) the out-of-domain faithfulness of post-hoc explanations, generated by five feature attribution methods; and (2) the out-of-domain performance of two inherently faithful models over six datasets. Contrary to our expectations, results show that in many cases out-of-domain post-hoc explanation faithfulness measured by sufficiency and comprehensiveness is higher compared to in-domain. We find this misleading and suggest using a random baseline as a yardstick for evaluating post-hoc explanation faithfulness. Our findings also show that select-then predict models demonstrate comparable predictive performance in out-of-domain settings to full-text trained models.

11:30-11:45 (The Liffey A)
Rewire-then-Probe: A Contrastive Recipe for Probing Biomedical Knowledge of Pre-trained Language Models
Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Anne Collins and Nigel Collier
Knowledge probing is crucial for understanding the knowledge transfer mechanism behind the pre-trained language models (PLMs). Despite the growing progress of probing knowledge for PLMs in the general domain, specialised areas such as the biomedical domain are vastly under-explored. To facilitate this, we release a well-curated biomedical knowledge probing benchmark, MedLAMA, constructed based on the Unified Medical Language System (UMLS) Metathesaurus. We test a wide spectrum of state-of-the-art PLMs and probing approaches on our benchmark, reaching at most 3% of acc@10. While highlighting various sources of domain-specific challenges that amount to this underwhelming performance, we illustrate that the underlying PLMs have a higher potential for probing tasks. To achieve this, we propose Contrastive-Probe, a novel self-supervised contrastive probing approach, that adjusts the underlying PLMs without using any probing data. While Contrastive-Probe pushes the acc@10 to 28%, the performance gap still remains notable. Our human expert evaluation suggests that the probing performance of our Contrastive-Probe is still under-estimated as UMLS still does not include the full spectrum of factual knowledge. We hope MedLAMA and Contrastive-Probe facilitate further developments of more suited probing techniques for this domain. Our code and dataset are publicly available at https://github.com/cambridgeltl/medlama.

11:45-12:00 (The Liffey A)
Is Attention Explanation? An Introduction to the Debate
Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François and Patrick Watrin
The performance of deep learning models in NLP and other fields of machine learning has led to a rise in their popularity, and so the need for explanations of these models becomes paramount. Attention has been seen as a solution to increase performance, while providing some explanations. However, a debate has started to cast doubt on the explanatory power of attention in neural networks. Although the debate has created a vast literature thanks to contributions from various areas, the lack of communication is becoming more and more tangible. In this paper, we provide a clear overview of the insights on the debate by critically confronting works from these different areas. This holistic vision can be of great interest for future works in all the communities concerned by this debate. We sum up the main challenges spotted in these areas, and we conclude by discussing the most promising future avenues on attention as an explanation.

12:00-12:15 (The Liffey A)
Probing as Quantifying Inductive Bias
Alexander Immer, Lucas Torroba Hennigen, Vincent Fortuin and Ryan D Cotterell
Pre-trained contextual representations have led to dramatic performance improvements on a range of downstream tasks. Such performance improvements have motivated researchers to quantify and understand the linguistic information encoded in these representations. In general, researchers quantify the amount of linguistic information through probing, an endeavor which consists of training a supervised model to predict a linguistic property directly from the contextual representations. Unfortunately, this definition of probing has been subject to extensive criticism in the literature, and has been observed to lead to paradoxical and counter-intuitive results. In the theoretical portion of this paper, we take the position that the goal of probing ought to be measuring the amount of inductive bias that the representations encode on a specific task. We further describe a Bayesian framework that operationalizes this goal and allows us to quantify the representations' inductive bias. In the empirical portion of the paper, we apply our framework to a variety of NLP tasks. Our results suggest that our proposed framework alleviates many previous problems found in probing. Moreover, we are able to offer concrete evidence that—for some tasks—fastText can offer a better inductive bias than BERT.

12:15-12:30 (The Liffey A)
Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity
Yao Lu, Max Bartolo, Alastair Philip Moore, Sebastian Riedel and Pontus Stenetorp
When primed with only a handful of training samples, very large, pretrained language models such as GPT-3 have shown competitive results when compared to fully-supervised, fine-tuned, large, pretrained language models. We demonstrate that the order in which the samples are provided can make the difference between near state-of-the-art and random guess performance: essentially some permutations are "fantastic" and some not. We analyse this phenomenon in detail, establishing that: it is present across model sizes (even for the largest current models), it is not related to a specific subset of samples, and that a given good permutation for one model is not transferable to another. While one could use a development set to determine which permutations are performant, this would deviate from the true few-shot setting as it requires additional annotated data. Instead, we use the generative nature of language models to construct an artificial development set and based on entropy statistics of the candidate permutations on this set, we identify performant prompts. Our method yields a 13

Machine Learning for NLP 3

11:00-12:30 (The Liffey B)

11:00-11:15 (The Liffey B)
[TACL] **PADA: Example-based Prompt Learning for on-the-fly Adaptation to Unseen Domains**
Roi Reichart, Eyal Ben-David and Nadav Oved

11:15-11:30 (The Liffey B)
[TACL] **Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP**
Timo Schick, Sahana Udupa and Hinrich Schütze

11:30-11:45 (The Liffey B)
Deduplicating Training Data Makes Language Models Better
Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch and Nicholas Carlini
We find that existing language modeling datasets contain many near-duplicate examples and long repetitive substrings. As a result, over 1We develop two tools that allow us to deduplicate training datasets—for example removing from C4 a single 61 word English sentence that is repeated over 60,000 times. Deduplication allows us to train models that emit memorized text ten times less frequently and require fewer training steps to achieve the same or better accuracy. We can also reduce train-test overlap, which affects over 4Code for deduplication is

released at https://github.com/google-research/deduplicate-text-datasets.

11:45-12:00 (The Liffey B)
Sparsifying Transformer Models with Trainable Representation Pooling
Michał Pietruszka, Łukasz Borchmann and Łukasz Garncarek
We propose a novel method to sparsify attention in the Transformer model by learning to select the most-informative token representations during the training process, thus focusing on the task-specific parts of an input. A reduction of quadratic time and memory complexity to sublinear was achieved due to a robust trainable top-k operator. Our experiments on a challenging long document summarization task show that even our simple baseline performs comparably to the current SOTA, and with trainable pooling we can retain its top quality, while being $1.8\times$ faster during training, $4.5\times$ faster during inference, and up to $13\times$ more computationally efficient in the decoder.

12:00-12:15 (The Liffey B)
∞-former: Infinite Memory Transformer
Pedro Henrique Martins, Zita Marinho and Andre Martins
Transformers are unable to model long-term memories effectively, since the amount of computation they need to perform grows with the context length. While variations of efficient transformers have been proposed, they all have a finite memory capacity and are forced to drop old information. In this paper, we propose the ∞-former, which extends the vanilla transformer with an unbounded long-term memory. By making use of a continuous-space attention mechanism to attend over the long-term memory, the ∞-former's attention complexity becomes independent of the context length, trading off memory length with precision. In order to control where precision is more important, ∞-former maintains "sticky memories," being able to model arbitrarily long contexts while keeping the computation budget fixed. Experiments on a synthetic sorting task, language modeling, and document grounded dialogue generation demonstrate the ∞-former's ability to retain information from long sequences.

12:15-12:30 (The Liffey B)
Imputing Out-of-Vocabulary Embeddings with LOVE Makes LanguageModels Robust with Little Cost
Lihu Chen, Gael Varoquaux and Fabian M. Suchanek
State-of-the-art NLP systems represent inputs with word embeddings, but these are brittle when faced with Out-of-Vocabulary (OOV) words. To address this issue, we follow the principle of mimick-like models to generate vectors for unseen words, by learning the behavior of pre-trained embeddings using only the surface form of words. We present a simple contrastive learning framework, LOVE, which extends the word representation of an existing pre-trained language model (such as BERT) and makes it robust to OOV with few additional parameters. Extensive evaluations demonstrate that our lightweight model achieves similar or even better performances than prior competitors, both on original datasets and on corrupted variants. Moreover, it can be used in a plug-and-play fashion with FastText and BERT, where it significantly improves their robustness.

NLP Applications 2

11:00-12:30 (Wicklow Hall 2a)

11:00-11:15 (Wicklow Hall 2a)
What does the sea say to the shore? A BERT based DST style approach for speaker to dialogue attribution in novels
Carolina Cuesta-Lazaro, Animesh Prasad and Trevor Wood
We present a complete pipeline to extract characters in a novel and link them to their direct-speech utterances. Our model is divided into three independent components: extracting direct-speech, compiling a list of characters, and attributing those characters to their utterances. Although we find that existing systems can perform the first two tasks accurately, attributing characters to direct speech is a challenging problem due to the narrator's lack of explicit character mentions, and the frequent use of nominal and pronominal coreference when such explicit mentions are made. We adapt the progress made on Dialogue State Tracking to tackle a new problem: attributing speakers to dialogues. This is the first application of deep learning to speaker attribution, and it shows that is possible to overcome the need for the hand-crafted features and rules used in the past. Our full pipeline improves the performance of state-of-the-art models by a relative 50% in F1-score.

11:15-11:30 (Wicklow Hall 2a)
TableFormer: Robust Transformer Modeling for Table-Text Encoding
Jingfeng Yang, Aditya Gupta, Shyam Upadhyay, Luheng He, Rahul Goel and Shachi Paul
Understanding tables is an important aspect of natural language understanding. Existing models for table understanding require linearization of the table structure, where row or column order is encoded as an unwanted bias. Such spurious biases make the model vulnerable to row and column order perturbations. Additionally, prior work has not thoroughly modeled the table structures or table-text alignments, hindering the table-text understanding ability. In this work, we propose a robust and structurally aware table-text encoding architecture TableFormer, where tabular structural biases are incorporated completely through learnable attention biases. TableFormer is (1) strictly invariant to row and column orders, and, (2) could understand tables better due to its tabular inductive biases. Our evaluations showed that TableFormer outperforms strong baselines in all settings on SQA, WTQ and TabFact table reasoning datasets, and achieves state-of-the-art performance on SQA, especially when facing answer-invariant row and column order perturbations (6

11:30-11:45 (Wicklow Hall 2a)
Your Answer is Incorrect... Would you like to know why? Introducing a Bilingual Short Answer Feedback Dataset
Anna Filighera, Siddharth Singh Parihar, Tim Steuer, Tobias Meuser and Sebastian Ochs
Handing in a paper or exercise and merely receiving "bad" or "incorrect" as feedback is not very helpful when the goal is to improve. Unfortunately, this is currently the kind of feedback given by Automatic Short Answer Grading (ASAG) systems. One of the reasons for this is a lack of content-focused elaborated feedback datasets. To encourage research on explainable and understandable feedback systems, we present the Short Answer Feedback dataset (SAF). Similar to other ASAG datasets, SAF contains learner responses and reference answers to German and English questions. However, instead of only assigning a label or score to the learners' answers, SAF also contains elaborated feedback explaining the given score. Thus, SAF enables supervised training of models that grade answers and explain where and why mistakes were made. This paper discusses the need for enhanced feedback models in real-world pedagogical scenarios, describes the dataset annotation

process, gives a comprehensive analysis of SAF, and provides T5-based baselines for future comparison.

11:45-12:00 (Wicklow Hall 2a)
Differentiable Multi-Agent Actor-Critic for Multi-Step Radiology Report Summarization
Sanjeev Kumar Karn, Ning Liu, Hinrich Schuetze and Oladimeji Farri
The IMPRESSIONS section of a radiology report about an imaging study is a summary of the radiologist's reasoning and conclusions, and it also aids the referring physician in confirming or excluding certain diagnoses. A cascade of tasks are required to automatically generate an abstractive summary of the typical information-rich radiology report. These tasks include acquisition of salient content from the report and generation of a concise, easily consumable IMPRESSIONS section. Prior research on radiology report summarization has focused on single-step end-to-end models – which subsume the task of salient content acquisition. To fully explore the cascade structure and explainability of radiology report summarization, we introduce two innovations. First, we design a two-step approach: extractive summarization followed by abstractive summarization. Second, we additionally break down the extractive part into two independent tasks: extraction of salient (1) sentences and (2) keywords. Experiments on a publicly available radiology report dataset show our novel approach leads to a more precise summary compared to single-step and to two-step-with-single-extractive-process baselines with an overall improvement in F1 score of 3-4

12:00-12:15 (Wicklow Hall 2a)
Adversarial Authorship Attribution for Deobfuscation
Wanyue Zhai, Jonathan Rusert, Zubair Shafiq and Padmini Srinivasan
Recent advances in natural language processing have enabled powerful privacy-invasive authorship attribution. To counter authorship attribution, researchers have proposed a variety of rule-based and learning-based text obfuscation approaches. However, existing authorship obfuscation approaches do not consider the adversarial threat model. Specifically, they are not evaluated against adversarially trained authorship attributors that are aware of potential obfuscation. To fill this gap, we investigate the problem of adversarial authorship attribution for deobfuscation. We show that adversarially trained authorship attributors are able to degrade the effectiveness of existing obfuscators from 20-30

12:15-12:30 (Wicklow Hall 2a)
[CL] **Novelty Detection: A Perspective from Natural Language Processing**
Tirthankar Ghosal, Tanik Saikh, Tameesh Biswas, Asif Ekbal and Pushpak Bhattacharyya

Resources and Evaluation 2

11:00-12:30 (Liffey Hall 1)

11:00-11:15 (Liffey Hall 1)
FaVIQ: FAct Verification from Information-seeking Questions
Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer and Hannaneh Hajishirzi
Despite significant interest in developing general purpose fact checking models, it is challenging to construct a large-scale fact verification dataset with realistic real-world claims. Existing claims are either authored by crowdworkers, thereby introducing subtle biases that are difficult to control for, or manually verified by professional fact checkers, causing them to be expensive and limited in scale. In this paper, we construct a large-scale challenging fact verification dataset called FAVIQ, consisting of 188k claims derived from an existing corpus of ambiguous information-seeking questions. The ambiguities in the questions enable automatically constructing true and false claims that reflect user confusions (e.g., the year of the movie being filmed vs. being released). Claims in FAVIQ are verified to be natural, contain little lexical bias, and require a complete understanding of the evidence for verification. Our experiments show that the state-of-the-art models are far from solving our new task. Moreover, training on our data helps in professional fact-checking, outperforming models trained on the widely used dataset FEVER or in-domain data by up to 17

11:15-11:30 (Liffey Hall 1)
A Taxonomy of Empathetic Questions in Social Dialogs
Ekaterina Svikhnushina, Iuliana Voinea, Anuradha Welivita and Pearl Pu
Effective question-asking is a crucial component of a successful conversational chatbot. It could help the bots manifest empathy and render the interaction more engaging by demonstrating attention to the speaker's emotions. However, current dialog generation approaches do not model this subtle emotion regulation technique due to the lack of a taxonomy of questions and their purpose in social chitchat. To address this gap, we have developed an empathetic question taxonomy (EQT), with special attention paid to questions' ability to capture communicative acts and their emotion-regulation intents. We further design a crowd-sourcing task to annotate a large subset of the EmpatheticDialogues dataset with the established labels. We use the crowd-annotated data to develop automatic labeling tools and produce labels for the whole dataset. Finally, we employ information visualization techniques to summarize co-occurrences of question acts and intents and their role in regulating interlocutor's emotion. These results reveal important question-asking strategies in social dialogs. The EQT classification scheme can facilitate computational analysis of questions in datasets. More importantly, it can inform future efforts in empathetic question generation using neural or hybrid methods.

11:30-11:45 (Liffey Hall 1)
RELiC: Retrieving Evidence for Literary Claims
Katherine Thai, Yapei Chang, Kalpesh Krishna and Mohit Iyyer
Humanities scholars commonly provide evidence for claims that they make about a work of literature (e.g., a novel) in the form of quotations from the work. We collect a large-scale dataset (RELiC) of 78K literary quotations and surrounding critical analysis and use it to formulate the novel task of literary evidence retrieval, in which models are given an excerpt of literary analysis surrounding a masked quotation and asked to retrieve the quoted passage from the set of all passages in the work. Solving this retrieval task requires a deep understanding of complex literary and linguistic phenomena, which proves challenging to methods that overwhelmingly rely on lexical and semantic similarity matching. We implement a RoBERTa-based dense passage retriever for this task that outperforms existing pretrained information retrieval baselines; however, experiments and analysis by human domain experts indicate that there is substantial room for improvement.

11:45-12:00 (Liffey Hall 1)
A Statutory Article Retrieval Dataset in French

Antoine Louis and Gerasimos Spanakis
Statutory article retrieval is the task of automatically retrieving law articles relevant to a legal question. While recent advances in natural language processing have sparked considerable interest in many legal tasks, statutory article retrieval remains primarily untouched due to the scarcity of large-scale and high-quality annotated datasets. To address this bottleneck, we introduce the Belgian Statutory Article Retrieval Dataset (BSARD), which consists of 1,100+ French native legal questions labeled by experienced jurists with relevant articles from a corpus of 22,600+ Belgian law articles. Using BSARD, we benchmark several state-of-the-art retrieval approaches, including lexical and dense architectures, both in zero-shot and supervised setups. We find that fine-tuned dense retrieval models significantly outperform other systems. Our best performing baseline achieves 74.8

12:00-12:15 (Liffey Hall 1)
Ditch the Gold Standard: Re-evaluating Conversational Question Answering
Huihan Li, Tianyu Gao, Manan Goenka and Danqi Chen
Conversational question answering aims to provide natural-language answers to users in information-seeking conversations. Existing conversational QA benchmarks compare models with pre-collected human-human conversations, using ground-truth answers provided in conversational history. It remains unclear whether we can rely on this static evaluation for model development and whether current systems can well generalize to real-world human-machine conversations. In this work, we conduct the first large-scale human evaluation of state-of-the-art conversational QA systems, where human evaluators converse with models and judge the correctness of their answers. We find that the distribution of human machine conversations differs drastically from that of human-human conversations, and there is a disagreement between human and gold-history evaluation in terms of model ranking. We further investigate how to improve automatic evaluations, and propose a question rewriting mechanism based on predicted history, which better correlates with human judgments. Finally, we analyze the impact of various modeling strategies and discuss future directions towards building better conversational question answering systems.

12:15-12:30 (Liffey Hall 1)
Detecting Unassimilated Borrowings in Spanish: An Annotated Corpus and Approaches to Modeling
Elena Álvarez-Mellado and Constantine Lignos
This work presents a new resource for borrowing identification and analyzes the performance and errors of several models on this task. We introduce a new annotated corpus of Spanish newswire rich in unassimilated lexical borrowings—words from one language that are introduced into another without orthographic adaptation—and use it to evaluate how several sequence labeling models (CRF, BiLSTM-CRF, and Transformer-based models) perform. The corpus contains 370,000 tokens and is larger, more borrowing-dense, OOV-rich, and topic-varied than previous corpora available for this task. Our results show that a BiLSTM-CRF model fed with subword embeddings along with either Transformer-based embeddings pretrained on codeswitched data or a combination of contextualized word embeddings outperforms results obtained by a multilingual BERT-based model.

Special Theme 2

11:00-12:30 (Liffey Hall 2)

11:00-11:15 (Liffey Hall 2)
One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia
Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau and Sebastian Ruder
NLP research is impeded by a lack of resources and awareness of the challenges presented by underrepresented languages and dialects. Focusing on the languages spoken in Indonesia, the second most linguistically diverse and the fourth most populous nation of the world, we provide an overview of the current state of NLP research for Indonesia's 700+ languages. We highlight challenges in Indonesian NLP and how these affect the performance of current NLP systems. Finally, we provide general recommendations to help develop NLP technology not only for languages of Indonesia but also other underrepresented languages.

11:15-11:30 (Liffey Hall 2)
Towards Afrocentric NLP for African Languages: Where We Are and Where We Can Go
Ife Adebara and Muhammad Abdul-Mageed
Aligning with ACL 2022 special Theme on "Language Diversity: from Low Resource to Endangered Languages", we discuss the major linguistic and sociopolitical challenges facing development of NLP technologies for African languages. Situating African languages in a typological framework, we discuss how the particulars of these languages can be harnessed. To facilitate future research, we also highlight current efforts, communities, venues, datasets, and tools. Our main objective is to motivate and advocate for an Afrocentric approach to technology development. With this in mind, we recommend *what* technologies to build and *how* to build, evaluate, and deploy them based on the needs of local African communities.

11:30-11:45 (Liffey Hall 2)
Weakly Supervised Word Segmentation for Computational Language Documentation
Shu Okabe, Laurent BesacierUGA and François Yvon
Word and morpheme segmentation are fundamental steps of language documentation as they allow to discover lexical units in a language for which the lexicon is unknown. However, in most language documentation scenarios, linguists do not start from a blank page: they may already have a pre-existing dictionary or have initiated manual segmentation of a small part of their data. This paper studies how such a weak supervision can be taken advantage of in Bayesian non-parametric models of segmentation. Our experiments on two very low resource languages (Mboshi and Japhug), whose documentation is still in progress, show that weak supervision can be beneficial to the segmentation quality. In addition, we investigate an incremental learning scenario where manual segmentations are provided in a sequential manner. This work opens the way for interactive annotation tools for documentary linguists.

11:45-12:00 (Liffey Hall 2)
Make the Best of Cross-lingual Transfer: Evidence from POS Tagging with over 100 Languages
Wietse de Vries, Martijn Wieling and Malvina Nissim
Cross-lingual transfer learning with large multilingual pre-trained models can be an effective approach for low-resource languages with no

labeled training data. Existing evaluations of zero-shot cross-lingual generalisability of large pre-trained models use datasets with English training data, and test data in a selection of target languages. We explore a more extensive transfer learning setup with 65 different source languages and 105 target languages for part-of-speech tagging. Through our analysis, we show that pre-training of both source and target language, as well as matching language families, writing systems, word order systems, and lexical-phonetic distance significantly impact cross-lingual performance. The findings described in this paper can be used as indicators of which factors are important for effective zero-shot cross-lingual transfer to zero- and low-resource languages.

12:00-12:15 (Liffey Hall 2)
Expanding Pretrained Models to Thousands More Languages via Lexicon-based Adaptation
Xinyi Wang, Sebastian Ruder and Graham Neubig
The performance of multilingual pretrained models is highly dependent on the availability of monolingual or parallel text present in a target language. Thus, the majority of the world's languages cannot benefit from recent progress in NLP as they have no or limited textual data. To expand possibilities of using NLP technology in these under-represented languages, we systematically study strategies that relax the reliance on conventional language resources through the use of bilingual lexicons, an alternative resource with much better language coverage. We analyze different strategies to synthesize textual or labeled data using lexicons, and how this data can be combined with monolingual or parallel text when available. For 19 under-represented languages across 3 tasks, our methods lead to consistent improvements of up to 5 and 15 points with and without extra monolingual text respectively. Overall, our study highlights how NLP methods can be adapted to thousands more languages that are under-served by current technology.

12:15-12:30 (Liffey Hall 2)
Interactive Word Completion for Plains Cree
William Abbott Lane, Atticus Galvin Harrigan and Antti Arppe
The composition of richly-inflected words in morphologically complex languages can be a challenge for language learners developing literacy. Accordingly, Lane and Bird (2020) proposed a finite state approach which maps prefixes in a language to a set of possible completions up to the next morpheme boundary, for the incremental building of complex words. In this work, we develop an approach to morph-based auto-completion based on a finite state morphological analyzer of Plains Cree (nêhiyawêwin), showing the portability of the concept to a much larger, more complete morphological transducer. Additionally, we propose and compare various novel ranking strategies on the morph auto-complete output. The best weighting scheme ranks the target completion in the top 10 results in 64.9

Summarization 1

11:00-12:30 (Wicklow Hall 1)

11:00-11:15 (Wicklow Hall 1)
EntSUM: A Data Set for Entity-Centric Extractive Summarization
Mounica Maddela, Mayank Kulkarni and Daniel Preotiuc-Pietro
Controllable summarization aims to provide summaries that take into account user-specified aspects and preferences to better assist them with their information need, as opposed to the standard summarization setup which build a single generic summary of a document. We introduce a human-annotated data set EntSUM for controllable summarization with a focus on named entities as the aspects to control. We conduct an extensive quantitative analysis to motivate the task of entity-centric summarization and show that existing methods for controllable summarization fail to generate entity-centric summaries. We propose extensions to state-of-the-art summarization approaches that achieve substantially better results on our data set. Our analysis and results show the challenging nature of this task and of the proposed data set.

11:15-11:30 (Wicklow Hall 1)
Learning Non-Autoregressive Models from Search for Unsupervised Sentence Summarization
Puyuan Liu, Chenyang Huang and Lili Mou
Text summarization aims to generate a short summary for an input text. In this work, we propose a Non-Autoregressive Unsupervised Summarization (NAUS) approach, which does not require parallel data for training. Our NAUS first performs edit-based search towards a heuristically defined score, and generates a summary as pseudo-groundtruth. Then, we train an encoder-only non-autoregressive Transformer based on the search result. We also propose a dynamic programming approach for length-control decoding, which is important for the summarization task. Experiments on two datasets show that NAUS achieves state-of-the-art performance for unsupervised summarization, yet largely improving inference efficiency. Further, our algorithm is able to perform explicit length-transfer summary generation.

11:30-11:45 (Wicklow Hall 1)
Efficient Unsupervised Sentence Compression by Fine-tuning Transformers with Reinforcement Learning
Demian Gholipour Ghalandari, Chris Hokamp and Georgiana Ifrim
Sentence compression reduces the length of text by removing non-essential content while preserving important facts and grammaticality. Unsupervised objective driven methods for sentence compression can be used to create customized models without the need for ground-truth training data, while allowing flexibility in the objective function(s) that are used for learning and inference. Recent unsupervised sentence compression approaches use custom objectives to guide discrete search; however, guided search is expensive at inference time. In this work, we explore the use of reinforcement learning to train effective sentence compression models that are also fast when generating predictions. In particular, we cast the task as binary sequence labelling and fine-tune a pre-trained transformer using a simple policy gradient approach. Our approach outperforms other unsupervised models while also being more efficient at inference time.

11:45-12:00 (Wicklow Hall 1)
A Multi-Document Coverage Reward for RELAXed Multi-Document Summarization
Jacob Parnell, Inigo Jauregi Unanue and Massimo Piccardi
Multi-document summarization (MDS) has made significant progress in recent years, in part facilitated by the availability of new, dedicated datasets and capacious language models. However, a standing limitation of these models is that they are trained against limited references and with plain maximum-likelihood objectives. As for many other generative tasks, reinforcement learning (RL) offers the potential to improve the training of MDS models; yet, it requires a carefully-designed reward that can ensure appropriate leverage of both the reference summaries and the input documents. For this reason, in this paper we propose fine-tuning an MDS baseline with a reward that balances a reference-based

metric such as ROUGE with coverage of the input documents. To implement the approach, we utilize RELAX (Grathwohl et al., 2018), a contemporary gradient estimator which is both low-variance and unbiased, and we fine-tune the baseline in a few-shot style for both stability and computational efficiency. Experimental results over the Multi-News and WCEP MDS datasets show significant improvements of up to +0.95 pp average ROUGE score and +3.17 pp METEOR score over the baseline, and competitive results with the literature. In addition, they show that the coverage of the input documents is increased, and evenly across all documents.

12:00-12:15 (Wicklow Hall 1)
[TACL] **Planning with Learned Entity Prompts for Abstractive Summarization**
Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Ryan McDonald and Vitaly Nikolaev

12:15-12:30 (Wicklow Hall 1)
[TACL] **Document Summarization with Latent Queries**
Yumo Xu and Mirella Lapata

Dialogue and Interactive Systems 2

11:00-12:30 (Wicklow Hall 2b)

11:00-11:15 (Wicklow Hall 2b)
DialFact: A Benchmark for Fact-Checking in Dialogue
Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu and Caiming Xiong
Fact-checking is an essential tool to mitigate the spread of misinformation and disinformation. We introduce the task of fact-checking in dialogue, which is a relatively unexplored area. We construct DialFact, a testing benchmark dataset of 22,245 annotated conversational claims, paired with pieces of evidence from Wikipedia. There are three sub-tasks in DialFact: 1) Verifiable claim detection task distinguishes whether a response carries verifiable factual information; 2) Evidence retrieval task retrieves the most relevant Wikipedia snippets as evidence; 3) Claim verification task predicts a dialogue response to be supported, refuted, or not enough information. We found that existing fact-checking models trained on non-dialogue data like FEVER fail to perform well on our task, and thus, we propose a simple yet data-efficient solution to effectively improve fact-checking performance in dialogue. We point out unique challenges in DialFact such as handling the colloquialisms, coreferences, and retrieval ambiguities in the error analysis to shed light on future research in this direction.

11:15-11:30 (Wicklow Hall 2b)
Situated Dialogue Learning through Procedural Environment Generation
Prithviraj Ammanabrolu, Renee Jia and Mark Riedl
We teach goal-driven agents to interactively act and speak in situated environments by training on generated curriculums. Our agents operate in LIGHT (Urbanek et al. 2019)—a large-scale crowd-sourced fantasy text adventure game wherein an agent perceives and interacts with the world through textual natural language. Goals in this environment take the form of character-based quests, consisting of personas and motivations. We augment LIGHT by learning to procedurally generate additional novel textual worlds and quests to create a curriculum of steadily increasing difficulty for training agents to achieve such goals. In particular, we measure curriculum difficulty in terms of the rarity of the quest in the original training distribution—an easier environment is one that is more likely to have been found in the unaugmented dataset. An ablation study shows that this method of learning from the tail of a distribution results in significantly higher generalization abilities as measured by zero-shot performance on never-before-seen quests.

11:30-11:45 (Wicklow Hall 2b)
Achieving Conversational Goals with Unsupervised Post-hoc Knowledge Injection
Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick and Julian McAuley
A limitation of current neural dialog models is that they tend to suffer from a lack of specificity and informativeness in generated responses, primarily due to dependence on training data that covers a limited variety of scenarios and conveys limited knowledge. One way to alleviate this issue is to extract relevant knowledge from external sources at decoding time and incorporate it into the dialog response. In this paper, we propose a post-hoc knowledge-injection technique where we first retrieve a diverse set of relevant knowledge snippets conditioned on both the dialog history and an initial response from an existing dialog model. We construct multiple candidate responses, individually injecting each retrieved snippet into the initial response using a gradient-based decoding method, and then select the final response with an unsupervised ranking step. Our experiments in goal-oriented and knowledge-grounded dialog settings demonstrate that human annotators judge the outputs from the proposed method to be more engaging and informative compared to responses from prior dialog systems. We further show that knowledge-augmentation promotes success in achieving conversational goals in both experimental settings.

11:45-12:00 (Wicklow Hall 2b)
When did you become so smart, oh wise one?! Sarcasm Explanation in Multi-modal Multi-party Dialogues
Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar and Tanmoy Chakraborty
Indirect speech such as sarcasm achieves a constellation of discourse goals in human communication. While the indirectness of figurative language warrants speakers to achieve certain pragmatic goals, it is challenging for AI agents to comprehend such idiosyncrasies of human communication. Though sarcasm identification has been a well-explored topic in dialogue analysis, for conversational systems to truly grasp a conversation's innate meaning and generate appropriate responses, simply detecting sarcasm is not enough; it is vital to explain its underlying sarcastic connotation to capture its true essence. In this work, we study the discourse structure of sarcastic conversations and propose a novel task – Sarcasm Explanation in Dialogue (SED). Set in a multimodal and code-mixed setting, the task aims to generate natural language explanations of satirical conversations. To this end, we curate WITS, a new dataset to support our task. We propose MAF (Modality Aware Fusion), a multimodal context-aware attention and global information fusion module to capture multimodality and use it to benchmark WITS. The proposed attention module surpasses the traditional multimodal fusion baselines and reports the best performance on almost all metrics. Lastly, we carry out detailed analysis both quantitatively and qualitatively.

12:00-12:15 (Wicklow Hall 2b)
[TACL] **TopiOCQA: Open-domain Conversational Question Answering with Topic Switching**
Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm Vries and Siva Reddy

12:15-12:25 (Wicklow Hall 2b)
Can Visual Dialogue Models Do Scorekeeping? Exploring How Dialogue Representations Incrementally Encode Shared Knowledge
Brielen Madureira and David Schlangen
Cognitively plausible visual dialogue models should keep a mental scoreboard of shared established facts in the dialogue context. We propose a theory-based evaluation method for investigating to what degree models pretrained on the VisDial dataset incrementally build representations that appropriately do scorekeeping. Our conclusion is that the ability to make the distinction between shared and privately known statements along the dialogue is moderately present in the analysed models, but not always incrementally consistent, which may partially be due to the limited need for grounding interactions in the original task.

Poster Session 4: Linguistic Theories, Cognitive Modeling and Psycholinguistics

11:00-12:30 (Forum)

11:00-12:30 (Forum)
#1 Do self-supervised speech models develop human-like perception biases?
Juliette Millet and Ewan Dunbar
Self-supervised models for speech processing form representational spaces without using any external labels. Increasingly, they appear to be a feasible way of at least partially eliminating costly manual annotations, a problem of particular concern for low-resource languages. But what kind of representational spaces do these models construct? Human perception specializes to the sounds of listeners' native languages. Does the same thing happen in self-supervised models? We examine the representational spaces of three kinds of state of the art self-supervised models: wav2vec, HuBERT and contrastive predictive coding (CPC), and compare them with the perceptual spaces of French-speaking and English-speaking human listeners, both globally and taking account of the behavioural differences between the two language groups. We show that the CPC model shows a small native language effect, but that wav2vec and HuBERT seem to develop a universal speech perception space which is not language specific. A comparison against the predictions of supervised phone recognisers suggests that all three self-supervised models capture relatively fine-grained perceptual phenomena, while supervised models are better at capturing coarser, phone-level effects, and effects of listeners' native language, on perception.

11:00-12:30 (Forum)
#2 Decoding Part-of-Speech from Human EEG Signals
Alex Graeme Murphy, Bernd Bohnet, Ryan McDonald and Uta Noppeney
This work explores techniques to predict Part-of-Speech (PoS) tags from neural signals measured at millisecond resolution with electroencephalography (EEG) during text reading. We first show that information about word length, frequency and word class is encoded by the brain at different post-stimulus latencies. We then demonstrate that pre-training on averaged EEG data and data augmentation techniques boost PoS decoding accuracy for single EEG trials. Finally, applying optimised temporally-resolved decoding techniques we show that Transformers substantially outperform linear-SVMs on PoS tagging of unigram and bigram data.

11:00-12:30 (Forum)
[TACL] **#3 Quantifying Cognitive Factors in Lexical Decline**
Ella Rabinovich, David Francis, Samir Farhan, David Mortensen and Suzanne Stevenson

11:00-12:30 (Forum)
[TACL] **#4 A Biologically Plausible Parser**
Daniel Mitropolsky, Michael Collins and Christos Papadimitriou

11:00-12:30 (Forum)
#5 Speaker Information Can Guide Models to Better Inductive Biases: A Case Study On Predicting Code-Switching
Alissa Ostapenko, Shuly Wintner, Melinda Fricke and Yulia Tsvetkov
Natural language processing (NLP) models trained on people-generated data can be unreliable because, without any constraints, they can learn from spurious correlations that are not relevant to the task. We hypothesize that enriching models with speaker information in a controlled, educated way can guide them to pick up on relevant inductive biases. For the speaker-driven task of predicting code-switching points in English–Spanish bilingual dialogues, we show that adding sociolinguistically-grounded speaker features as prepended prompts significantly improves accuracy. We find that by adding influential phrases to the input, speaker-informed models learn useful and explainable linguistic information. To our knowledge, we are the first to incorporate speaker characteristics in a neural model for code-switching, and more generally, take a step towards developing transparent, personalized models that use speaker information in a controlled way.

11:00-12:30 (Forum)
#6 Characterizing Idioms: Conventionality and Contingency
Michaela Socolof, Jackie CK Cheung, Michael Wagner and Timothy J. O'Donnell
Idioms are unlike most phrases in two important ways. First, words in an idiom have non-canonical meanings. Second, the non-canonical meanings of words in an idiom are contingent on the presence of other words in the idiom. Linguistic theories differ on whether these properties depend on one another, as well as whether special theoretical machinery is needed to accommodate idioms. We define two measures that correspond to the properties above, and we show that idioms fall at the expected intersection of the two dimensions, but that the dimensions themselves are not correlated. Our results suggest that introducing special machinery to handle idioms may not be warranted.

11:00-12:30 (Forum)
#7 Do Transformer Models Show Similar Attention Patterns to Task-Specific Human Gaze?
Oliver Eberle, Stephanie Brandl, Jonas Pilot and Anders Søgaard
Learned self-attention functions in state-of-the-art NLP models often correlate with human attention. We investigate whether self-attention in large-scale pre-trained language models is as predictive of human eye fixation patterns during task-reading as classical cognitive models of human attention. We compare attention functions across two task-specific reading datasets for sentiment analysis and relation extraction. We find the predictiveness of large-scale pre-trained self-attention for human attention depends on 'what is in the tail', e.g., the syntactic nature of rare contexts. Further, we observe that task-specific fine-tuning does not increase the correlation with human task-specific reading. Through an input reduction experiment we give complementary insights on the sparsity and fidelity trade-off, showing that lower-entropy attention vectors are more faithful.

11:00-12:30 (Forum)
#8 Analyzing Wrap-Up Effects through an Information-Theoretic Lens
Clara Isabel Meister, Tiago Pimentel, Thomas Hikaru Clark, Ryan D Cotterell and Roger P. Levy
Numerous analyses of reading time (RT) data have been undertaken in the effort to learn more about the internal processes that occur during reading comprehension. However, data measured on words at the end of a sentence–or even clause–is often omitted due to the confounding factors introduced by so-called "wrap-up effects," which manifests as a skewed distribution of RTs for these words. Consequently, the understanding of the cognitive processes that might be involved in these effects is limited. In this work, we attempt to learn more about these processes by looking for the existence–or absence–of a link between wrap-up effects and information theoretic quantities, such as word and context information content. We find that the information distribution of prior context is often predictive of sentence- and clause-final RTs (while not of sentence-medial RTs), which lends support to several prior hypotheses about the processes involved in wrap-up effects.

11:00-12:30 (Forum)
#9 Context Matters: A Pragmatic Study of PLMs' Negation Understanding
Reto Gubelmann and Siegfried Handschuh
In linguistics, there are two main perspectives on negation: a semantic and a pragmatic view. So far, research in NLP on negation has almost exclusively adhered to the semantic view. In this article, we adopt the pragmatic paradigm to conduct a study of negation understanding focusing on transformer-based PLMs. Our results differ from previous, semantics-based studies and therefore help to contribute a more comprehensive – and, given the results, much more optimistic – picture of the PLMs' negation understanding.

11:00-12:30 (Forum)
#10 Word Segmentation as Unsupervised Constituency Parsing
Raquel G. Alhama
Word identification from continuous input is typically viewed as a segmentation task. Experiments with human adults suggest that familiarity with syntactic structures in their native language also influences word identification in artificial languages; however, the relation between syntactic processing and word identification is yet unclear. This work takes one step forward by exploring a radically different approach of word identification, in which segmentation of a continuous input is viewed as a process isomorphic to unsupervised constituency parsing. Besides formalizing the approach, this study reports simulations of human experiments with DIORA (Drozdov et al., 2020), a neural unsupervised constituency parser. Results show that this model can reproduce human behavior in word identification experiments, suggesting that this is a viable approach to study word identification and its relation to syntactic processing.

11:00-12:30 (Forum)
#11 Learning Functional Distributional Semantics with Visual Data
Yinhong Liu and Guy Emerson
Functional Distributional Semantics is a recently proposed framework for learning distributional semantics that provides linguistic interpretability. It models the meaning of a word as a binary classifier rather than a numerical vector. In this work, we propose a method to train a Functional Distributional Semantics model with grounded visual data. We train it on the Visual Genome dataset, which is closer to the kind of data encountered in human language acquisition than a large text corpus. On four external evaluation datasets, our model outperforms previous work on learning semantics from Visual Genome.

11:00-12:30 (Forum)
[CL] **#12 Assessing corpus evidence for formal and psycholinguistic constraints on nonprojectivity**
Himanshu Yadav, Samar Husain and Richard Futrell

11:00-12:30 (Forum)
#13 Slangvolution: A Causal Analysis of Semantic Change and Frequency Dynamics in Slang
Daphna Keidar, Andreas Opedal, Zhijing Jin and Mrinmaya Sachan
Languages are continuously undergoing changes, and the mechanisms that underlie these changes are still a matter of debate. In this work, we approach language evolution through the lens of causality in order to model not only how various distributional factors associate with language change, but how they causally affect it. In particular, we study slang, which is an informal language that is typically restricted to a specific group or social setting. We analyze the semantic change and frequency shift of slang words and compare them to those of standard, nonslang words. With causal discovery and causal inference techniques, we measure the effect that word type (slang/nonslang) has on both semantic change and frequency shift, as well as its relationship to frequency, polysemy and part of speech. Our analysis provides some new insights in the study of language change, e.g., we show that slang words undergo less semantic change but tend to have larger frequency shifts over time.

Poster Session 4: Syntax: Tagging, Chunking and Parsing

11:00-12:30 (Forum)

11:00-12:30 (Forum)
#14 Compositional Generalization in Dependency Parsing
Emily Goodwin, Siva Reddy, Timothy J. O'Donnell and Dzmitry Bahdanau
Compositionality— the ability to combine familiar units like words into novel phrases and sentences— has been the focus of intense interest in artificial intelligence in recent years. To test compositional generalization in semantic parsing, Keysers et al. (2020) introduced Compositional Freebase Queries (CFQ). This dataset maximizes the similarity between the test and train distributions over primitive units, like words, while maximizing the compound divergence: the dissimilarity between test and train distributions over larger structures, like phrases. Dependency parsing, however, lacks a compositional generalization benchmark. In this work, we introduce a gold-standard set of dependency parses for CFQ, and use this to analyze the behaviour of a state-of-the art dependency parser (Qi et al., 2020) on the CFQ dataset. We find that increasing compound divergence degrades dependency parsing performance, although not as dramatically as semantic parsing performance. Additionally, we find the performance of the dependency parser does not uniformly degrade relative to compound divergence, and the parser performs differently on different splits with the same compound divergence. We explore a number of hypotheses for what causes the non-uniform degradation in dependency parsing performance, and identify a number of syntactic structures that drive the dependency parser's lower performance on the most challenging splits.

11:00-12:30 (Forum)
[CL] **#15 Improved N-Best Extraction with an Evaluation on Language Data**
Johanna Björklund, Frank Drewes, Anna Jonsson

11:00-12:30 (Forum)
#16 Revisiting the Effects of Leakage on Dependency Parsing
Nathaniel Krasner, Miriam Wanner and Antonios Anastasopoulos
Recent work by Søgaard (2020) showed that, treebank size aside, overlap between training and test graphs (termed *leakage*) explains more of the observed variation in dependency parsing performance than other explanations. In this work we revisit this claim, testing it on more models and languages. We find that it only holds for zero-shot cross-lingual settings. We then propose a more fine-grained measure of such leakage which, unlike the original measure, not only explains but also correlates with observed performance variation. Code and data are available here: https://github.com/miriamwanner/reu-nlp-project

11:00-12:30 (Forum)
#17 Semantic Composition with PSHRG for Derivation Tree Reconstruction from Graph-Based Meaning Representations
Chun Hei Lo, Wai Lam and Hong Cheng
We introduce a data-driven approach to generating derivation trees from meaning representation graphs with probabilistic synchronous hyperedge replacement grammar (PSHRG). SHRG has been used to produce meaning representation graphs from texts and syntax trees, but little is known about its viability on the reverse. In particular, we experiment on Dependency Minimal Recursion Semantics (DMRS) and adapt PSHRG as a formalism that approximates the semantic composition of DMRS graphs and simultaneously recovers the derivations that license the DMRS graphs. Consistent results are obtained as evaluated on a collection of annotated corpora. This work reveals the ability of PSHRG in formalizing a syntax–semantics interface, modelling compositional graph-to-tree translations, and channelling explainability to surface realization.

11:00-12:30 (Forum)
#18 Meta-Learning for Fast Cross-Lingual Adaptation in Dependency Parsing
Anna Langedijk, Verna Dankers, Phillip Lippe, Sander Bos, Bryan Cardenas Guevara, Helen Yannakoudakis and Ekaterina Shutova
Meta-learning, or learning to learn, is a technique that can help to overcome resource scarcity in cross-lingual NLP problems, by enabling fast adaptation to new tasks. We apply model-agnostic meta-learning (MAML) to the task of cross-lingual dependency parsing. We train our model on a diverse set of languages to learn a parameter initialization that can adapt quickly to new languages. We find that meta-learning with pre-training can significantly improve upon the performance of language transfer and standard supervised learning baselines for a variety of unseen, typologically diverse, and low-resource languages, in a few-shot learning setup.

11:00-12:30 (Forum)
#19 MELM: Data Augmentation with Masked Entity Language Modeling for Low-Resource NER
Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si and Chunyan Miao
Data augmentation is an effective solution to data scarcity in low-resource scenarios. However, when applied to token-level tasks such as NER, data augmentation methods often suffer from token-label misalignment, which leads to unsatsifactory performance. In this work, we propose Masked Entity Language Modeling (MELM) as a novel data augmentation framework for low-resource NER. To alleviate the token-label misalignment issue, we explicitly inject NER labels into sentence context, and thus the fine-tuned MELM is able to predict masked entity tokens by explicitly conditioning on their labels. Thereby, MELM generates high-quality augmented data with novel entities, which provides rich entity regularity knowledge and boosts NER performance. When training data from multiple languages are available, we also integrate MELM with code-mixing for further improvement. We demonstrate the effectiveness of MELM on monolingual, cross-lingual and multilingual NER across various low-resource levels. Experimental results show that our MELM consistently outperforms the baseline methods.

11:00-12:30 (Forum)
#20 Probing for Labeled Dependency Trees
Max Müller-Eberstein, Rob Van Der Goot and Barbara Plank
Probing has become an important tool for analyzing representations in Natural Language Processing (NLP). For graphical NLP tasks such as dependency parsing, linear probes are currently limited to extracting undirected or unlabeled parse trees which do not capture the full task. This work introduces DepProbe, a linear probe which can extract labeled and directed dependency parse trees from embeddings while using fewer parameters and compute than prior methods. Leveraging its full task coverage and lightweight parametrization, we investigate

its predictive power for selecting the best transfer language for training a full biaffine attention parser. Across 13 languages, our proposed method identifies the best source treebank 94

11:00-12:30 (Forum)
#21 Zero-Shot Dependency Parsing with Worst-Case Aware Automated Curriculum Learning
Miryam De Lhoneux, Sheng Zhang and Anders Søgaard
Large multilingual pretrained language models such as mBERT and XLM-RoBERTa have been found to be surprisingly effective for cross-lingual transfer of syntactic parsing models Wu and Dredze (2019), but only between related languages. However, source and training languages are rarely related, when parsing truly low-resource languages. To close this gap, we adopt a method from multi-task learning, which relies on automated curriculum learning, to dynamically optimize for parsing performance on *outlier* languages. We show that this approach is significantly better than uniform and size-proportional sampling in the zero-shot setting.

11:00-12:30 (Forum)
#22 Co-training an Unsupervised Constituency Parser with Weak Supervision
Nickil Maveli and Shay B Cohen
We introduce a method for unsupervised parsing that relies on bootstrapping classifiers to identify if a node dominates a specific span in a sentence. There are two types of classifiers, an inside classifier that acts on a span, and an outside classifier that acts on everything outside of a given span. Through self-training and co-training with the two classifiers, we show that the interplay between them helps improve the accuracy of both, and as a result, effectively parse. A seed bootstrapping technique prepares the data to train these classifiers. Our analyses further validate that such an approach in conjunction with weak supervision using prior branching knowledge of a known language (left/right-branching) and minimal heuristics injects strong inductive bias into the parser, achieving 63.1 F_1 on the English (PTB) test set. In addition, we show the effectiveness of our architecture by evaluating on treebanks for Chinese (CTB) and Japanese (KTB) and achieve new state-of-the-art results.

11:00-12:30 (Forum)
#23 Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise
Noëmi Aepli and Rico Sennrich
Cross-lingual transfer between a high-resource language and its dialects or closely related language varieties should be facilitated by their similarity. However, current approaches that operate in the embedding space do not take surface similarity into account. This work presents a simple yet effective strategy to improve cross-lingual transfer between closely related varieties. We propose to augment the data of the high-resource source language with character-level noise to make the model more robust towards spelling variations. Our strategy shows consistent improvements over several languages and tasks: Zero-shot transfer of POS tagging and topic identification between language varieties from the Finnic, West and North Germanic, and Western Romance language branches. Our work provides evidence for the usefulness of simple surface-level noise in improving transfer between language varieties.

11:00-12:30 (Forum)
#24 Coloring the Blank Slate: Pre-training Imparts a Hierarchical Inductive Bias to Sequence-to-sequence Models
Aaron Mueller, Robert Frank, Tal Linzen, Luheng Wang and Sebastian Schuster
Relations between words are governed by hierarchical structure rather than linear ordering. Sequence-to-sequence (seq2seq) models, despite their success in downstream NLP applications, often fail to generalize in a hierarchy-sensitive manner when performing syntactic transformations—for example, transforming declarative sentences into questions. However, syntactic evaluations of seq2seq models have only observed models that were not pre-trained on natural language data before being trained to perform syntactic transformations, in spite of the fact that pre-training has been found to induce hierarchical linguistic generalizations in language models; in other words, the syntactic capabilities of seq2seq models may have been greatly understated. We address this gap using the pre-trained seq2seq models T5 and BART, as well as their multilingual variants mT5 and mBART. We evaluate whether they generalize hierarchically on two transformations in two languages: question formation and passivization in English and German. We find that pre-trained seq2seq models generalize hierarchically when performing syntactic transformations, whereas models trained from scratch on syntactic transformations do not. This result presents evidence for the learnability of hierarchical syntactic information from non-annotated natural language text while also demonstrating that seq2seq models are capable of syntactic generalization, though only after exposure to much more language data than human learners receive.

11:00-12:30 (Forum)
#25 Pretraining with Artificial Language: Studying Transferable Knowledge in Language Models
Ryokan Ri and Yoshimasa Tsuruoka
We investigate what kind of structural knowledge learned in neural network encoders is transferable to processing natural language. We design *artificial languages* with structural properties that mimic natural language, pretrain encoders on the data, and see how much performance the encoder exhibits on downstream tasks in natural language. Our experimental results show that pretraining with an artificial language with a nesting dependency structure provides some knowledge transferable to natural language. A follow-up probing analysis indicates that its success in the transfer is related to the amount of encoded contextual information and what is transferred is the knowledge of *position-aware context dependence* of language. Our results provide insights into how neural network encoders process human languages and the source of cross-lingual transferability of recent multilingual language models.

Poster Session 4: Semantics

11:00-12:30 (Forum)

11:00-12:30 (Forum)
#26 Zero-Shot Cross-lingual Semantic Parsing
Tom Sherborne and Mirella Lapata
Recent work in cross-lingual semantic parsing has successfully applied machine translation to localize parsers to new languages. However, these advances assume access to high-quality machine translation systems and word alignment tools. We remove these assumptions and study cross-lingual semantic parsing as a zero-shot problem, without parallel data (i.e., utterance-logical form pairs) for new languages. We propose a multi-task encoder-decoder model to transfer parsing knowledge to additional languages using only English-logical form paired data and

in-domain natural language corpora in each new language. Our model encourages language-agnostic encodings by jointly optimizing for logical-form generation with auxiliary objectives designed for cross-lingual latent representation alignment. Our parser performs significantly above translation-based baselines and, in some cases, competes with the supervised upper-bound.

11:00-12:30 (Forum)
#27 Improving Word Translation via Two-Stage Contrastive Learning
Yaoyiran Li, Fangyu Liu, Nigel Collier, Anna Korhonen and Ivan Vulić
Word translation or bilingual lexicon induction (BLI) is a key cross-lingual task, aiming to bridge the lexical gap between different languages. In this work, we propose a robust and effective two-stage contrastive learning framework for the BLI task. At Stage C1, we propose to refine standard cross-lingual linear maps between static word embeddings (WEs) via a contrastive learning objective; we also show how to integrate it into the self-learning procedure for even more refined cross-lingual maps. In Stage C2, we conduct BLI-oriented contrastive fine-tuning of mBERT, unlocking its word translation capability. We also show that static WEs induced from the 'C2-tuned' mBERT complement static WEs from Stage C1. Comprehensive experiments on standard BLI datasets for diverse languages and different experimental setups demonstrate substantial gains achieved by our framework. While the BLI method from Stage C1 already yields substantial gains over all state-of-the-art BLI methods in our comparison, even stronger improvements are met with the full two-stage framework: e.g., we report gains for 112/112 BLI setups, spanning 28 language pairs.

11:00-12:30 (Forum)
#28 Predicate-Argument Based Bi-Encoder for Paraphrase Identification
Qiwei Peng, David Weir, Julie Weeds and Yekun Chai
Paraphrase identification involves identifying whether a pair of sentences express the same or similar meanings. While cross-encoders have achieved high performances across several benchmarks, bi-encoders such as SBERT have been widely applied to sentence pair tasks. They exhibit substantially lower computation complexity and are better suited to symmetric tasks. In this work, we adopt a bi-encoder approach to the paraphrase identification task, and investigate the impact of explicitly incorporating predicate-argument information into SBERT through weighted aggregation. Experiments on six paraphrase identification datasets demonstrate that, with a minimal increase in parameters, the proposed model is able to outperform SBERT/SRoBERTa significantly. Further, ablation studies reveal that the predicate-argument based component plays a significant role in the performance gain.

11:00-12:30 (Forum)
#29 Modeling Syntactic-Semantic Dependency Correlations in Semantic Role Labeling Using Mixture Models
Junjie Chen, Xiangheng He and Yusuke Miyao
In this paper, we propose a mixture model-based end-to-end method to model the syntactic-semantic dependency correlation in Semantic Role Labeling (SRL). Semantic dependencies in SRL are modeled as a distribution over semantic dependency labels conditioned on a predicate and an argument word. The semantic label distribution varies depending on Shortest Syntactic Dependency Path (SSDP) hop patterns. We target the variation of semantic label distributions using a mixture model, separately estimating semantic label distributions for different hop patterns and probabilistically clustering hop patterns with similar semantic label distributions. Experiments show that the proposed method successfully learns a cluster assignment reflecting the variation of semantic label distributions. Modeling the variation improves performance in predicting short distance semantic dependencies, in addition to the improvement on long distance semantic dependencies that previous syntax-aware methods have achieved. The proposed method achieves a small but statistically significant improvement over baseline methods in English, German, and Spanish and obtains competitive performance with state-of-the-art methods in English.

11:00-12:30 (Forum)
#30 Towards Better Characterization of Paraphrases
Timothy Liu and De Wen Soh
To effectively characterize the nature of paraphrase pairs without expert human annotation, we proposes two new metrics: word position deviation (WPD) and lexical deviation (LD). WPD measures the degree of structural alteration, while LD measures the difference in vocabulary used. We apply these metrics to better understand the commonly-used MRPC dataset and study how it differs from PAWS, another paraphrase identification dataset. We also perform a detailed study on MRPC and propose improvements to the dataset, showing that it improves generalizability of models trained on the dataset. Lastly, we apply our metrics to filter the output of a paraphrase generation model and show how it can be used to generate specific forms of paraphrases for data augmentation or robustness testing of NLP models.

11:00-12:30 (Forum)
#31 Lacking the Embedding of a Word? Look it up into a Traditional Dictionary
Elena Sofia Ruzzetti, Leonardo Ranaldi, Michele Mastromattei, Francesca Fallucchi, Noemi Scarpato and Fabio Massimo Zanzotto
Word embeddings are powerful dictionaries, which may easily capture language variations. However, these dictionaries fail to give sense to rare words, which are surprisingly often covered by traditional dictionaries. In this paper, we propose to use definitions retrieved in traditional dictionaries to produce word embeddings for rare words. For this purpose, we introduce two methods: Definition Neural Network (DefiNNet) and Define BERT (DefBERT). In our experiments, DefiNNet and DefBERT significantly outperform state-of-the-art as well as baseline methods devised for producing embeddings of unknown words. In fact, DefiNNet significantly outperforms FastText, which implements a method for the same task-based on n-grams, and DefBERT significantly outperforms the BERT method for OOV words. Then, definitions in traditional dictionaries are useful to build word embeddings for rare words.

11:00-12:30 (Forum)
#32 Learning from Missing Relations: Contrastive Learning with Commonsense Knowledge Graphs for Commonsense Inference
Yong-Ho Jung, Jun-Hyung Park, Joon-Young Choi, Mingyu Lee, Junho Kim, Kang-Min Kim and SangKeun Lee
Commonsense inference poses a unique challenge to reason and generate the physical, social, and causal conditions of a given event. Existing approaches to commonsense inference utilize commonsense transformers, which are large-scale language models that learn commonsense knowledge graphs. However, they suffer from a lack of coverage and expressive diversity of the graphs, resulting in a degradation of the representation quality. In this paper, we focus on addressing missing relations in commonsense knowledge graphs, and propose a novel contrastive learning framework called SOLAR. Our framework contrasts sets of semantically similar and dissimilar events, learning richer inferential knowledge compared to existing approaches. Empirical results demonstrate the efficacy of SOLAR in commonsense inference of diverse commonsense knowledge graphs. Specifically, SOLAR outperforms the state-of-the-art commonsense transformer on commonsense inference with ConceptNet by 1.84

11:00-12:30 (Forum)
#33 Cross-lingual Inference with A Chinese Entailment Graph
Tianyi Li, Sabine Weber, Mohammad Javad Hosseini, Liane Guillou and Mark Steedman
Predicate entailment detection is a crucial task for question-answering from text, where previous work has explored unsupervised learning of entailment graphs from typed open relation triples. In this paper, we present the first pipeline for building Chinese entailment graphs, which involves a novel high-recall open relation extraction (ORE) method and the first Chinese fine-grained entity typing dataset under the FIGER type ontology. Through experiments on the Levy-Holt dataset, we verify the strength of our Chinese entailment graph, and reveal the cross-lingual complementarity: on the parallel Levy-Holt dataset, an ensemble of Chinese and English entailment graphs outperforms both monolingual graphs, and raises unsupervised SOTA by 4.7 AUC points.

11:00-12:30 (Forum)
[TACL] **#34 Is My Model Using The Right Evidence? Systematic Probes for Examining Evidence-Based Tabular Reasoning**
Vivek Gupta, Riyaz Bhat, Atreya Ghosal, Manish Shrivastava, Maneesh Singh and Vivek Srikumar

11:00-12:30 (Forum)
[TACL] **#35 It's not Rocket Science: Interpreting Figurative Language in Narratives**
Tuhin Chakrabarty, Yejin Choi and Vered Shwartz

11:00-12:30 (Forum)
[TACL] **#36 Weisfeiler-Leman in the BAMBOO: Novel AMR Graph Metrics and a Benchmark for AMR Graph Similarity**
Juri Opitz, Anette Frank and Angel Daza

11:00-12:30 (Forum)
#37 Leveraging Similar Users for Personalized Language Modeling with Limited Data
Charles Welch, Chenxi Gu, Jonathan K Kummerfeld, Veronica Perez-Rosas and Rada Mihalcea
Personalized language models are designed and trained to capture language patterns specific to individual users. This makes them more accurate at predicting what a user will write. However, when a new user joins a platform and not enough text is available, it is harder to build effective personalized language models. We propose a solution for this problem, using a model trained on users that are similar to a new user. In this paper, we explore strategies for finding the similarity between new users and existing ones and methods for using the data from existing users who are a good match. We further explore the trade-off between available data for new users and how well their language can be modeled.

11:00-12:30 (Forum)
#38 ExtEnD: Extractive Entity Disambiguation
Edoardo Barba, Luigi Procopio and Roberto Navigli
Local models for Entity Disambiguation (ED) have today become extremely powerful, in most part thanks to the advent of large pre-trained language models. However, despite their significant performance achievements, most of these approaches frame ED through classification formulations that have intrinsic limitations, both computationally and from a modeling perspective. In contrast with this trend, here we propose ExtEnD, a novel local formulation for ED where we frame this task as a text extraction problem, and present two Transformer-based architectures that implement it. Based on experiments in and out of domain, and training over two different data regimes, we find our approach surpasses all its competitors in terms of both data efficiency and raw performance. ExtEnD outperforms its alternatives by as few as 6 F1 points on the more constrained of the two data regimes and, when moving to the other higher-resourced regime, sets a new state of the art on 4 out of 4 benchmarks under consideration, with average improvements of 0.7 F1 points overall and 1.1 F1 points out of domain. In addition, to gain better insights from our results, we also perform a fine-grained evaluation of our performances on different classes of label frequency, along with an ablation study of our architectural choices and an error analysis. We release our code and models for research purposes at https://github.com/SapienzaNLP/extend.

11:00-12:30 (Forum)
#39 WatClaimCheck: A new Dataset for Claim Entailment and Inference
Kashif Khan, Ruizhe Wang and Pascal Poupart
We contribute a new dataset for the task of automated fact checking and an evaluation of state of the art algorithms. The dataset includes claims (from speeches, interviews, social media and news articles), review articles published by professional fact checkers and premise articles used by those professional fact checkers to support their review and verify the veracity of the claims. An important challenge in the use of premise articles is the identification of relevant passages that will help to infer the veracity of a claim. We show that transferring a dense passage retrieval model trained with review articles improves the retrieval quality of passages in premise articles. We report results for the prediction of claim veracity by inference from premise articles.

11:00-12:30 (Forum)
#40 Generating Data to Mitigate Spurious Correlations in Natural Language Inference Datasets
Yuxiang Wu, Matt Gardner, Pontus Stenetorp and Pradeep Dasigi
Natural language processing models often exploit spurious correlations between task-independent features and labels in datasets to perform well only within the distributions they are trained on, while not generalising to different task distributions. We propose to tackle this problem by generating a debiased version of a dataset, which can then be used to train a debiased, off-the-shelf model, by simply replacing its training data. Our approach consists of 1) a method for training data generators to generate high-quality, label-consistent data samples; and 2) a filtering mechanism for removing data points that contribute to spurious correlations, measured in terms of z-statistics. We generate debiased versions of the SNLI and MNLI datasets, and we evaluate on a large suite of debiased, out-of-distribution, and adversarial test sets. Results show that models trained on our debiased datasets generalise better than those trained on the original datasets in all settings. On the majority of the datasets, our method outperforms or performs comparably to previous state-of-the-art debiasing strategies, and when combined with an orthogonal technique, product-of-experts, it improves further and outperforms previous best results of SNLI-hard and MNLI-hard.

11:00-12:30 (Forum)
#41 Right for the Right Reason: Evidence Extraction for Trustworthy Tabular Reasoning
Vivek Gupta, Shuo Zhang, Alakananda Vempala, Yujie He, Temma Choji and Vivek Srikumar
When pre-trained contextualized embedding-based models developed for unstructured data are adapted for structured tabular data, they perform admirably. However, recent probing studies show that these models use spurious correlations, and often predict inference labels by focusing on false evidence or ignoring it altogether. To study this issue, we introduce the task of Trustworthy Tabular Reasoning, where a model needs to extract evidence to be used for reasoning, in addition to predicting the label. As a case study, we propose a two-stage sequential prediction approach, which includes an evidence extraction and an inference stage. First, we crowdsource evidence row labels and develop several unsupervised and supervised evidence extraction strategies for InfoTabS, a tabular NLI benchmark. Our evidence extraction strategy outperforms earlier baselines. On the downstream tabular inference task, using only the automatically extracted evidence as the premise, our approach outperforms prior benchmarks.

11:00-12:30 (Forum)
#42 Probing for Predicate Argument Structures in Pretrained Language Models
Simone Conia and Roberto Navigli
Thanks to the effectiveness and wide availability of modern pretrained language models (PLMs), recently proposed approaches have achieved remarkable results in dependency- and span-based, multilingual and cross-lingual Semantic Role Labeling (SRL). These results have prompted researchers to investigate the inner workings of modern PLMs with the aim of understanding how, where, and to what extent they encode information about SRL. In this paper, we follow this line of research and probe for predicate argument structures in PLMs. Our study shows that PLMs do encode semantic structures directly into the contextualized representation of a predicate, and also provides insights into the correlation between predicate senses and their structures, the degree of transferability between nominal and verbal structures, and how such structures are encoded across languages. Finally, we look at the practical implications of such insights and demonstrate the benefits of embedding predicate argument structure information into an SRL model.

11:00-12:30 (Forum)
#43 EPT-X: An Expression-Pointer Transformer model that generates eXplanations for numbers
Bugeun Kim, Kyung Seo Ki, Sangkyu Rhim and Gahgene Gweon
In this paper, we propose a neural model EPT-X (Expression-Pointer Transformer with Explanations), which utilizes natural language explanations to solve an algebraic word problem. To enhance the explainability of the encoding process of a neural model, EPT-X adopts the concepts of plausibility and faithfulness which are drawn from math word problem solving strategies by humans. A plausible explanation is one that includes contextual information for the numbers and variables that appear in a given math word problem. A faithful explanation is one that accurately represents the reasoning process behind the model's solution equation. The EPT-X model yields an average baseline performance of 69.59

11:00-12:30 (Forum)
#44 Fully-Semantic Parsing and Generation: the BabelNet Meaning Representation
Abelardo Carlos Martinez Lorenzo, Marco Maru and Roberto Navigli
A language-independent representation of meaning is one of the most coveted dreams in Natural Language Understanding. With this goal in mind, several formalisms have been proposed as frameworks for meaning representation in Semantic Parsing. And yet, the dependencies these formalisms share with respect to language-specific repositories of knowledge make the objective of closing the gap between high- and low-resourced languages hard to accomplish. In this paper, we present the BabelNet Meaning Representation (BMR), an interlingual formalism that abstracts away from language-specific constraints by taking advantage of the multilingual semantic resources of BabelNet and VerbAtlas. We describe the rationale behind the creation of BMR and put forward BMR 1.0, a dataset labeled entirely according to the new formalism. Moreover, we show how BMR is able to outperform previous formalisms thanks to its fully-semantic framing, which enables top-notch multilingual parsing and generation. We release the code at https://github.com/SapienzaNLP/bmr.

11:00-12:30 (Forum)
#45 Just Rank: Rethinking Evaluation with Word and Sentence Similarities
Bin Wang, C.-c. Jay Kuo and Haizhou Li
Word and sentence embeddings are useful feature representations in natural language processing. However, intrinsic evaluation for embeddings lags far behind, and there has been no significant update since the past decade. Word and sentence similarity tasks have become the de facto evaluation method. It leads models to overfit to such evaluations, negatively impacting embedding models' development. This paper first points out the problems using semantic similarity as the gold standard for word and sentence embedding evaluations. Further, we propose a new intrinsic evaluation method called EvalRank, which shows a much stronger correlation with downstream tasks. Extensive experiments are conducted based on 60+ models and popular datasets to certify our judgments. Finally, the practical evaluation toolkit is released for future benchmarking purposes.

11:00-12:30 (Forum)
#46 Few-Shot Learning with Siamese Networks and Label Tuning
Thomas Müller, Guillermo Pérez-Torró and Marc Franco-Salvador
We study the problem of building text classifiers with little or no training data, commonly known as zero and few-shot text classification. In recent years, an approach based on neural textual entailment models has been found to give strong results on a diverse range of tasks. In this work, we show that with proper pre-training, Siamese Networks that embed texts and labels offer a competitive alternative. These models allow for a large reduction in inference cost: constant in the number of labels rather than linear. Furthermore, we introduce label tuning, a simple and computationally efficient approach that allows to adapt the models in a few-shot setup by only changing the label embeddings. While giving lower performance than model fine-tuning, this approach has the architectural advantage that a single encoder can be shared by many different tasks.

11:00-12:30 (Forum)
#47 LAGr: Label Aligned Graphs for Better Systematic Generalization in Semantic Parsing
Dora Jambor and Dzmitry Bahdanau
Semantic parsing is the task of producing structured meaning representations for natural language sentences. Recent research has pointed out that the commonly-used sequence-to-sequence (seq2seq) semantic parsers struggle to generalize systematically, i.e. to handle examples that require recombining known knowledge in novel settings. In this work, we show that better systematic generalization can be achieved

by producing the meaning representation directly as a graph and not as a sequence. To this end we propose LAGr (Label Aligned Graphs), a general framework to produce semantic parses by independently predicting node and edge labels for a complete multi-layer input-aligned graph. The strongly-supervised LAGr algorithm requires aligned graphs as inputs, whereas weakly-supervised LAGr infers alignments for originally unaligned target graphs using approximate maximum-a-posteriori inference. Experiments demonstrate that LAGr achieves significant improvements in systematic generalization upon the baseline seq2seq parsers in both strongly- and weakly-supervised settings.

11:00-12:30 (Forum)
#48 Principled Paraphrase Generation with Parallel Corpora
Aitor Ormazabal, Mikel Artetxe, Aitor Soroa, Gorka Labaka and Eneko Agirre
Round-trip Machine Translation (MT) is a popular choice for paraphrase generation, which leverages readily available parallel corpora for supervision. In this paper, we formalize the implicit similarity function induced by this approach, and show that it is susceptible to non-paraphrase pairs sharing a single ambiguous translation. Based on these insights, we design an alternative similarity metric that mitigates this issue by requiring the entire translation distribution to match, and implement a relaxation of it through the Information Bottleneck method. Our approach incorporates an adversarial term into MT training in order to learn representations that encode as much information about the reference translation as possible, while keeping as little information about the input as possible. Paraphrases can be generated by decoding back to the source from this representation, without having to generate pivot translations. In addition to being more principled and efficient than round-trip MT, our approach offers an adjustable parameter to control the fidelity-diversity trade-off, and obtains better results in our experiments.

11:00-12:30 (Forum)
#49 FaiRR: Faithful and Robust Deductive Reasoning over Natural Language
Soumya Sanyal, Harman Singh and Xiang Ren
Transformers have been shown to be able to perform deductive reasoning on a logical rulebase containing rules and statements written in natural language. Recent works show that such models can also produce the reasoning steps (i.e., the proof graph) that emulate the model's logical reasoning process. Currently, these black-box models generate both the proof graph and intermediate inferences within the same model and thus may be unfaithful. In this work, we frame the deductive logical reasoning task by defining three modular components: rule selection, fact selection, and knowledge composition. The rule and fact selection steps select the candidate rule and facts to be used and then the knowledge composition combines them to generate new inferences. This ensures model faithfulness by assured causal relation from the proof step to the inference reasoning. To test our framework, we propose FaiRR (Faithful and Robust Reasoner) where the above three components are independently modeled by transformers. We observe that FaiRR is robust to novel language perturbations, and is faster at inference than previous works on existing reasoning datasets. Additionally, in contrast to black-box generative models, the errors made by FaiRR are more interpretable due to the modular approach.

11:00-12:30 (Forum)
#50 Open Relation Modeling: Learning to Define Relations between Entities
Jie Huang, Kevin Chang, Jinjun Xiong and Wen-mei Hwu
Relations between entities can be represented by different instances, e.g., a sentence containing both entities or a fact in a Knowledge Graph (KG). However, these instances may not well capture the general relations between entities, may be difficult to understand by humans, even may not be found due to the incompleteness of the knowledge source. In this paper, we introduce the Open Relation Modeling problem - given two entities, generate a coherent sentence describing the relation between them. To solve this problem, we propose to teach machines to generate definition-like relation descriptions by letting them learn from defining entities. Specifically, we fine-tune Pre-trained Language Models (PLMs) to produce definitions conditioned on extracted entity pairs. To help PLMs reason between entities and provide additional relational knowledge to PLMs for open relation modeling, we incorporate reasoning paths in KGs and include a reasoning path selection mechanism. Experimental results show that our model can generate concise but informative relation descriptions that capture the representative characteristics of entities.

11:00-12:30 (Forum)
#51 To be or not to be an Integer? Encoding Variables for Mathematical Text
Deborah Ferreira, Mokanarangan Thayaparan, Marco Valentino, Julia Rozanova and Andre Freitas
The application of Natural Language Inference (NLI) methods over large textual corpora can facilitate scientific discovery, reducing the gap between current research and the available large-scale scientific knowledge. However, contemporary NLI models are still limited in interpreting mathematical knowledge written in Natural Language, even though mathematics is an integral part of scientific argumentation for many disciplines. One of the fundamental requirements towards mathematical language understanding, is the creation of models able to meaningfully represent variables. This problem is particularly challenging since the meaning of a variable should be assigned exclusively from its defining type, i.e., the representation of a variable should come from its context. Recent research has formalised the variable typing task, a benchmark for the understanding of abstract mathematical types and variables in a sentence. In this work, we propose VarSlot, a Variable Slot-based approach, which not only delivers state-of-the-art results in the task of variable typing, but is also able to create context-based representations for variables.

11:00-12:30 (Forum)
#52 Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models
Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer and Yinfei Yang
We provide the first exploration of sentence embeddings from text-to-text transformers (T5) including the effects of scaling up sentence encoders to 11B parameters. Sentence embeddings are broadly useful for language processing tasks. While T5 achieves impressive performance on language tasks, it is unclear how to produce sentence embeddings from encoder-decoder models. We investigate three methods to construct Sentence-T5 (ST5) models: two utilize only the T5 encoder and one using the full T5 encoder-decoder. We establish a new sentence representation transfer benchmark, SentGLUE, which extends the SentEval toolkit to nine tasks from the GLUE benchmark. Our encoder-only models outperform the previous best models on both SentEval and SentGLUE transfer tasks, including semantic textual similarity (STS). Scaling up ST5 from millions to billions of parameters shown to consistently improve performance. Finally, our encoder-decoder method achieves a new state-of-the-art on STS when using sentence embeddings.[4]

[4]Our models are released at https://tfhub.dev/google/collections/sentence-t5/1.

11:00-12:30 (Forum)
#53 Unsupervised Natural Language Inference Using PHL Triplet Generation
Neeraj Varshney, Pratyay Banerjee, Tejas Gokhale and Chitta Baral
Transformer-based models achieve impressive performance on numerous Natural Language Inference (NLI) benchmarks when trained on respective training datasets. However, in certain cases, training samples may not be available or collecting them could be time-consuming and resource-intensive. In this work, we address the above challenge and present an explorative study on unsupervised NLI, a paradigm in which no human-annotated training samples are available. We investigate it under three settings: PH, P, and NPH that differ in the extent of unlabeled data available for learning. As a solution, we propose a procedural data generation approach that leverages a set of sentence transformations to collect PHL (Premise, Hypothesis, Label) triplets for training NLI models, bypassing the need for human-annotated training data. Comprehensive experiments with several NLI datasets show that the proposed approach results in accuracies of up to 66.75

11:00-12:30 (Forum)
#54 VALUE: Understanding Dialect Disparity in NLU
Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson and Diyi Yang
English Natural Language Understanding (NLU) systems have achieved great performances and even outperformed humans on benchmarks like GLUE and SuperGLUE. However, these benchmarks contain only textbook Standard American English (SAE). Other dialects have been largely overlooked in the NLP community. This leads to biased and inequitable NLU systems that serve only a sub-population of speakers. To understand disparities in current models and to facilitate more dialect-competent NLU systems, we introduce the VernAcular Language Understanding Evaluation (VALUE) benchmark, a challenging variant of GLUE that we created with a set of lexical and morphosyntactic transformation rules. In this initial release (V.1), we construct rules for 11 features of African American Vernacular English (AAVE), and we recruit fluent AAVE speakers to validate each feature transformation via linguistic acceptability judgments in a participatory design manner. Experiments show that these new dialectal features can lead to a drop in model performance.

11:00-12:30 (Forum)
[DEMO] **SocioFillmore: A Tool for Discovering Perspectives**
Gosse Minnema, Sara Gemelli, Chiara Zanchi, Tommaso Caselli and Malvina Nissim
SOCIOFILLMORE is a multilingual tool which helps to bring to the fore the focus or the perspective that a text expresses in depicting an event. Our tool, whose rationale we also support through a large collection of human judgements, is theoretically grounded on frame semantics and cognitive linguistics, and implemented using the LOME frame semantic parser. We describe SOCIOFILLMORE's development and functionalities, show how non-NLP researchers can easily interact with the tool, and present some example case studies which are already incorporated in the system, together with the kind of analysis that can be visualised.

Poster Session 4: Phonology, Morphology and Word Segmentation

11:00-12:30 (Forum)

11:00-12:30 (Forum)
#55 Morphosyntactic Tagging with Pre-trained Language Models for Arabic and its Dialects
Go Inoue, Salam Khalifa and Nizar Habash
We present state-of-the-art results on morphosyntactic tagging across different varieties of Arabic using fine-tuned pre-trained transformer language models. Our models consistently outperform existing systems in Modern Standard Arabic and all the Arabic dialects we study, achieving 2.6

11:00-12:30 (Forum)
#56 Morphological Reinflection with Multiple Arguments: An Extended Annotation schema and a Georgian Case Study
David Guriel, Omer Goldman and Reut Tsarfaty
In recent years, a flurry of morphological datasets had emerged, most notably UniMorph, aa multi-lingual repository of inflection tables. However, the flat structure of the current morphological annotation makes the treatment of some languages quirky, if not impossible, specifically in cases of polypersonal agreement. In this paper we propose a general solution for such cases and expand the UniMorph annotation schema to naturally address this phenomenon, in which verbs agree with multiple arguments using true affixes. We apply this extended schema to one such language, Georgian, and provide a human-verified, accurate and balanced morphological dataset for Georgian verbs. The dataset has 4 times more tables and 6 times more verb forms compared to the existing UniMorph dataset, covering all possible variants of argument marking, demonstrating the adequacy of our proposed scheme. Experiments on a reinflection task show that generalization is easy when the data is split at the form level, but extremely hard when splitting along lemma lines. Expanding the other languages in UniMorph according to this schema is expected to improve both the coverage, consistency and interpretability of this benchmark.

11:00-12:30 (Forum)
#57 WLASL-LEX: a Dataset for Recognising Phonological Properties in American Sign Language
Federico Tavella, Viktor Schlegel, Marta Romeo, Aphrodite Galata and Angelo Cangelosi
Signed Language Processing (SLP) concerns the automated processing of signed languages, the main means of communication of Deaf and hearing impaired individuals. SLP features many different tasks, ranging from sign recognition to translation and production of signed speech, but has been overlooked by the NLP community thus far. In this paper, we bring to attention the task of modelling the phonology of sign languages. We leverage existing resources to construct a large-scale dataset of American Sign Language signs annotated with six different phonological properties. We then conduct an extensive empirical study to investigate whether data-driven end-to-end and feature-based approaches can be optimised to automatically recognise these properties. We find that, despite the inherent challenges of the task, graph-based neural networks that operate over skeleton features extracted from raw videos are able to succeed at the task to a varying degree. Most importantly, we show that this performance pertains even on signs unobserved during training.

11:00-12:30 (Forum)
#58 CaMEL: Case Marker Extraction without Labels
Leonie Weissweiler, Valentin Hofmann, Masoud Jalili Sabet and Hinrich Schuetze

We introduce **CaMEL** (**Ca**se **M**arker **E**xtraction without **L**abels), a novel and challenging task in computational morphology that is especially relevant for low-resource languages. We propose a first model for CaMEL that uses a massively multilingual corpus to extract case markers in 83 languages based only on a noun phrase chunker and an alignment system. To evaluate CaMEL, we automatically construct a silver standard from UniMorph. The case markers extracted by our model can be used to detect and visualise similarities and differences between the case systems of different languages as well as to annotate fine-grained deep cases in languages in which they are not overtly marked.

11:00-12:30 (Forum)
#59 (Un)solving Morphological Inflection: Lemma Overlap Artificially Inflates Models' Performance
Omer Goldman, David Guriel and Reut Tsarfaty
In the domain of Morphology, Inflection is a fundamental and important task that gained a lot of traction in recent years, mostly via SIGMORPHON's shared-tasks. With average accuracy above 0.9 over the scores of all languages, the task is considered mostly solved using relatively generic neural seq2seq models, even with little data provided. In this work, we propose to re-evaluate morphological inflection models by employing harder train-test splits that will challenge the generalization capacity of the models. In particular, as opposed to the naïve split-by-form, we propose a split-by-lemma method to challenge the performance on existing benchmarks. Our experiments with the three top-ranked systems on the SIGMORPHON's 2020 shared-task show that the lemma-split presents an average drop of 30 percentage points in macro-average for the 90 languages included. The effect is most significant for low-resourced languages with a drop as high as 95 points, but even high-resourced languages lose about 10 points on average. Our results clearly show that generalizing inflection to unseen lemmas is far from being solved, presenting a simple yet effective means to promote more sophisticated models.

11:00-12:30 (Forum)
#60 More Than Words: Collocation Retokenization for Latent Dirichlet Allocation Models
Jin Cheevaprawatdomrong, Alexandra Schofield and Attapol Rutherford
Traditionally, Latent Dirichlet Allocation (LDA) ingests words in a collection of documents to discover their latent topics using word-document co-occurrences. Previous studies show that representing bigrams collocations in the input can improve topic coherence in English. However, it is unclear how to achieve the best results for languages without marked word boundaries such as Chinese and Thai. Here, we explore the use of retokenization based on chi-squared measures, t-statistics, and raw frequency to merge frequent token ngrams into collocations when preparing input to the LDA model. Based on the goodness of fit and the coherence metric, we show that topics trained with merged tokens result in topic keys that are clearer, more coherent, and more effective at distinguishing topics than those of unmerged models.

11:00-12:30 (Forum)
#61 An Embarrassingly Simple Method to Mitigate Undesirable Properties of Pretrained Language Model Tokenizers
Valentin Hofmann, Hinrich Schuetze and Janet B. Pierrehumbert
We introduce FLOTA (Few Longest Token Approximation), a simple yet effective method to improve the tokenization of pretrained language models (PLMs). FLOTA uses the vocabulary of a standard tokenizer but tries to preserve the morphological structure of words during tokenization. We evaluate FLOTA on morphological gold segmentations as well as a text classification task, using BERT, GPT-2, and XLNet as example PLMs. FLOTA leads to performance gains, makes inference more efficient, and enhances the robustness of PLMs with respect to whitespace noise.

11:00-12:30 (Forum)
#62 A Functionalist Account of Vowel System Typology
Ryan D Cotterell and Jason Eisner
The typology of sound systems in spoken human languages should be explained in part by functional pressures on communication. Two competing pressures are per-phoneme transmission rate and ease of communication. A system with few phonemes transmits limited information per phoneme, but the phonemes can be easily pronounced and perceived. Adding more phonemes can increase the amount of information per phoneme—but that requires using "outlier" sounds, which are more difficult to pronounce or perceive, or else splitting phonemes, which requires more speaker and hearer effort to keep them distinct. We encode these two competing pressures into a proposed universal prior for a generative probability model. We find that a model of vowel token formants is more predictive of held-out data if it is trained with the help of this prior (that is, by MAP rather than ML).

Poster Session 4: Discourse and Pragmatics

11:00-12:30 (Forum)

11:00-12:30 (Forum)
#63 Graph Refinement for Coreference Resolution
Lesly Miculicich and James Henderson
The state-of-the-art models for coreference resolution are based on independent mention pair-wise decisions. We propose a modelling approach that learns coreference at the document-level and takes global decisions. For this purpose, we model coreference links in a graph structure where the nodes are tokens in the text, and the edges represent the relationship between them. Our model predicts the graph in a non-autoregressive manner, then iteratively refines it based on previous predictions, allowing global dependencies between decisions. The experimental results show improvements over various baselines, reinforcing the hypothesis that document-level information improves conference resolution.

11:00-12:30 (Forum)
[TACL] **#64 Out-of-Domain Discourse Dependency Parsing via Bootstrapping: An Empirical Analysis on Its Effectiveness and Limitation**
Noriki Nishida and Yuji Matsumoto

11:00-12:30 (Forum)
#65 How Do We Answer Complex Questions: Discourse Structure of Long-form Answers

Fangyuan Xu, Junyi Jessy Li and Eunsol Choi
Long-form answers, consisting of multiple sentences, can provide nuanced and comprehensive answers to a broader set of questions. To better understand this complex and understudied task, we study the functional structure of long-form answers collected from three datasets, ELI5, WebGPT and Natural Questions. Our main goal is to understand how humans organize information to craft complex answers. We develop an ontology of six sentence-level functional roles for long-form answers, and annotate 3.9k sentences in 640 answer paragraphs. Different answer collection methods manifest in different discourse structures. We further analyze model-generated answers – finding that annotators agree less with each other when annotating model-generated answers compared to annotating human-written answers. Our annotated data enables training a strong classifier that can be used for automatic analysis. We hope our work can inspire future research on discourse-level modeling and evaluation of long-form QA systems.

11:00-12:30 (Forum)
#66 Entity-based Neural Local Coherence Modeling
Sungho Jeon and Michael Strube
In this paper, we propose an entity-based neural local coherence model which is linguistically more sound than previously proposed neural coherence models. Recent neural coherence models encode the input document using large-scale pretrained language models. Hence their basis for computing local coherence are words and even sub-words. The analysis of their output shows that these models frequently compute coherence on the basis of connections between (sub-)words which, from a linguistic perspective, should not play a role. Still, these models achieve state-of-the-art performance in several end applications. In contrast to these models, we compute coherence on the basis of entities by constraining the input to noun phrases and proper names. This provides us with an explicit representation of the most important items in sentences leading to the notion of focus. This brings our model linguistically in line with pre-neural models of computing coherence. It also gives us better insight into the behaviour of the model thus leading to better explainability. Our approach is also in accord with a recent study (O'Connor and Andreas, 2021), which shows that most usable information is captured by nouns and verbs in transformer-based language models. We evaluate our model on three downstream tasks showing that it is not only linguistically more sound than previous models but also that it outperforms them in end applications.

11:00-12:30 (Forum)
#67 Rethinking Self-Supervision Objectives for Generalizable Coherence Modeling
Prathyusha Jwalapuram, Shafiq Joty and Xiang Lin
Given the claims of improved text generation quality across various pre-trained neural models, we consider the coherence evaluation of machine generated text to be one of the principal applications of coherence models that needs to be investigated. Prior work in neural coherence modeling has primarily focused on devising new architectures for solving the permuted document task. We instead use a basic model architecture and show significant improvements over state of the art within the same training regime. We then design a harder self-supervision objective by increasing the ratio of negative samples within a contrastive learning setup, and enhance the model further through automatic hard negative mining coupled with a large global negative queue encoded by a momentum encoder. We show empirically that increasing the density of negative samples improves the basic model, and using a global negative queue further improves and stabilizes the model while training with hard negative samples. We evaluate the coherence model on task-independent test sets that resemble real-world applications and show significant improvements in coherence evaluations of downstream tasks.

11:00-12:30 (Forum)
#68 What does it take to bake a cake? The RecipeRef corpus and anaphora resolution in procedural text
Biaoyan Fang, Timothy Baldwin and Karin Verspoor
Procedural text contains rich anaphoric phenomena, yet has not received much attention in NLP. To fill this gap, we investigate the textual properties of two types of procedural text, recipes and chemical patents, and generalize an anaphora annotation framework developed for the chemical domain for modeling anaphoric phenomena in recipes. We apply this framework to annotate the RecipeRef corpus with both bridging and coreference relations. Through comparison to chemical patents, we show the complexity of anaphora resolution in recipes. We demonstrate empirically that transfer learning from the chemical domain improves resolution of anaphora in recipes, suggesting transferability of general procedural knowledge.

11:00-12:30 (Forum)
#69 The Change that Matters in Discourse Parsing: Estimating the Impact of Domain Shift on Parser Error
Katherine Atwell, Anthony Sicilia, Seong Jae Hwang and Malihe Alikhani
Discourse analysis allows us to attain inferences of a text document that extend beyond the sentence-level. The current performance of discourse models is very low on texts outside of the training distribution's coverage, diminishing the practical utility of existing models. There is need for a measure that can inform us to what extent our model generalizes from the training to the test sample when these samples may be drawn from distinct distributions. While this can be estimated via distribution shift, we argue that this does not directly correlate with change in the observed error of a classifier (i.e. error-gap). Thus, we propose to use a statistic from the theoretical domain adaptation literature which can be directly tied to error-gap. We study the bias of this statistic as an estimator of error-gap both theoretically and through a large-scale empirical study of over 2400 experiments on 6 discourse datasets from domains including, but not limited to: news, biomedical texts, TED talks, Reddit posts, and fiction. Our results not only motivate our proposal and help us to understand its limitations, but also provide insight on the properties of discourse models and datasets which improve performance in domain adaptation. For instance, we find that non-news datasets are slightly easier to transfer to than news datasets when the training and test sets are very different. Our code and an associated Python package are available to allow practitioners to make more informed model and dataset choices.

Poster Session 4: Sentiment Analysis, Stylistic Analysis, and Argument Mining

11:00-12:30 (Forum)

11:00-12:30 (Forum)
#70 Fair and Argumentative Language Modeling for Computational Argumentation
Carolin Holtermann, Anne Lauscher and Simone Paolo Ponzetto
Although much work in NLP has focused on measuring and mitigating stereotypical bias in semantic spaces, research addressing bias in computational argumentation is still in its infancy. In this paper, we address this research gap and conduct a thorough investigation of bias in

argumentative language models. To this end, we introduce ABBA, a novel resource for bias measurement specifically tailored to argumentation. We employ our resource to assess the effect of argumentative fine-tuning and debiasing on the intrinsic bias found in transformer-based language models using a lightweight adapter-based approach that is more sustainable and parameter-efficient than full fine-tuning. Finally, we analyze the potential impact of language model debiasing on the performance in argument quality prediction, a downstream task of computational argumentation. Our results show that we are able to successfully and sustainably remove bias in general and argumentative language models while preserving (and sometimes improving) model performance in downstream tasks. We make all experimental code and data available at https://github.com/umanlp/FairArgumentativeLM.

11:00-12:30 (Forum)
[CL] **#71 Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis**
Saif M. Mohammad

11:00-12:30 (Forum)
#72 ECO v1: Towards Event-Centric Opinion Mining
Ruoxi Xu, Hongyu Lin, Meng Liao, Xianpei Han, Jin Xu, Wei Tan, Yingfei Sun and Le Sun
Events are considered as the fundamental building blocks of the world. Mining event-centric opinions can benefit decision making, people communication, and social good. Unfortunately, there is little literature addressing event-centric opinion mining, although which significantly diverges from the well-studied entity-centric opinion mining in connotation, structure, and expression. In this paper, we propose and formulate the task of event-centric opinion mining based on event-argument structure and expression categorizing theory. We also benchmark this task by constructing a pioneer corpus and designing a two-step benchmark framework. Experiment results show that event-centric opinion mining is feasible and challenging, and the proposed task, dataset, and baselines are beneficial for future studies.

11:00-12:30 (Forum)
[TACL] **#73 Comprehensive Investigation of Multi-task Argument Mining**
Gaku Morio, Hiroaki Ozaki, Terufumi Morishita and Kohsuke Yanai

11:00-12:30 (Forum)
#74 Identifying the Human Values behind Arguments
Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth and Benno Stein
This paper studies the (often implicit) human values behind natural language arguments, such as to have freedom of thought or to be broad-minded. Values are commonly accepted answers to why some option is desirable in the ethical sense and are thus essential both in real-world argumentation and theoretical argumentation frameworks. However, their large variety has been a major obstacle to modeling them in argument mining. To overcome this obstacle, we contribute an operationalization of human values, namely a multi-level taxonomy with 54 values that is in line with psychological research. Moreover, we provide a dataset of 5270 arguments from four geographical cultures, manually annotated for human values. First experiments with the automatic classification of human values are promising, with F_1-scores up to 0.81 and 0.25 on average.

11:00-12:30 (Forum)
#75 DoCoGen: Domain Counterfactual Generation for Low Resource Domain Adaptation
Nitay Calderon, Eyal Ben-David, Amir Feder and Roi Reichart
Natural language processing (NLP) algorithms have become very successful, but they still struggle when applied to out-of-distribution examples. In this paper we propose a controllable generation approach in order to deal with this domain adaptation (DA) challenge. Given an input text example, our DoCoGen algorithm generates a domain-counterfactual textual example (D-con) - that is similar to the original in all aspects, including the task label, but its domain is changed to a desired one. Importantly, DoCoGen is trained using only unlabeled examples from multiple domains - no NLP task labels or parallel pairs of textual examples and their domain-counterfactuals are required. We show that DoCoGen can generate coherent counterfactuals consisting of multiple sentences. We use the D-cons generated by DoCoGen to augment a sentiment classifier and a multi-label intent classifier in 20 and 78 DA setups, respectively, where source-domain labeled data is scarce. Our model outperforms strong baselines and improves the accuracy of a state-of-the-art unsupervised DA algorithm.

11:00-12:30 (Forum)
#76 Can Unsupervised Knowledge Transfer from Social Discussions Help Argument Mining?
Subhabrata Dutta, Jeevesh Juneja, Dipankar Das and Tanmoy Chakraborty
Identifying argument components from unstructured texts and predicting the relationships expressed among them are two primary steps of argument mining. The intrinsic complexity of these tasks demands powerful learning models. While pretrained Transformer-based Language Models (LM) have been shown to provide state-of-the-art results over different NLP tasks, the scarcity of manually annotated data and the highly domain-dependent nature of argumentation restrict the capabilities of such models. In this work, we propose a novel transfer learning strategy to overcome these challenges. We utilize argumentation-rich social discussions from the *ChangeMyView* subreddit as a source of unsupervised, argumentative discourse-aware knowledge by finetuning pretrained LMs on a selectively masked language modeling task. Furthermore, we introduce a novel prompt-based strategy for inter-component relation prediction that compliments our proposed finetuning method while leveraging on the discourse context. Exhaustive experiments show the generalization capability of our method on these two tasks over within-domain as well as out-of-domain datasets, outperforming several existing and employed strong baselines.

11:00-12:30 (Forum)
#77 The Moral Debater: A Study on the Computational Generation of Morally Framed Arguments
Milad Alshomary, Roxanne El Baff, Timon Gurcke and Henning Wachsmuth
An audience's prior beliefs and morals are strong indicators of how likely they will be affected by a given argument. Utilizing such knowledge can help focus on shared values to bring disagreeing parties towards agreement. In argumentation technology, however, this is barely exploited so far. This paper studies the feasibility of automatically generating morally framed arguments as well as their effect on different audiences. Following the moral foundation theory, we propose a system that effectively generates arguments focusing on different morals. In an in-depth user study, we ask liberals and conservatives to evaluate the impact of these arguments. Our results suggest that, particularly when prior beliefs are challenged, an audience becomes more affected by morally framed arguments.

11:00-12:30 (Forum)
#78 A Rationale-Centric Framework for Human-in-the-loop Machine Learning
Jinghui Lu, Linyi Yang, Brian Mac Namee and Yue Zhang
We present a novel rational-centric framework with human-in-the-loop – Rationales-centric Double-robustness Learning (RDL) – to boost model out-of-distribution performance in few-shot learning scenarios. By using static semi-factual generation and dynamic human-intervened correction, RDL, acting like a sensible "inductive bias", exploits rationales (i.e. phrases that cause the prediction), human interventions and semi-factual augmentations to decouple spurious associations and bias models towards generally applicable underlying distributions, which enables fast and accurate generalisation. Experimental results show that RDL leads to significant prediction benefits on both in-distribution and out-of-distribution tests, especially for few-shot learning scenarios, compared to many state-of-the-art benchmarks. We also perform extensive ablation studies to support in-depth analyses of each component in our framework.

11:00-12:30 (Forum)
#79 So Different Yet So Alike! Constrained Unsupervised Text Style Transfer
Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, Roger Zimmermann and Soujanya Poria
Automatic transfer of text between domains has become popular in recent times. One of its aims is to preserve the semantic content while adapting to the target domain. However, it does not explicitly maintain other attributes between the source and translated text: e.g., text length and descriptiveness. Maintaining constraints in transfer has several downstream applications, including data augmentation and debiasing. We introduce a method for such constrained unsupervised text style transfer by introducing two complementary losses to the generative adversarial network (GAN) family of models. Unlike the competing losses used in GANs, we introduce cooperative losses where the discriminator and the generator cooperate and reduce the same loss. The first is a contrastive loss and the second is a classification loss — aiming to regularize the latent space further and bring similar sentences closer together. We demonstrate that such training retains lexical, syntactic and domain-specific constraints between domains for multiple benchmark datasets, including ones where more than one attribute change. We show that the complementary cooperative losses improve text quality, according to both automated and human evaluation measures.

11:00-12:30 (Forum)
#80 Direct parsing to sentiment graphs
David Samuel, Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid and Erik Velldal
This paper demonstrates how a graph-based semantic parser can be applied to the task of structured sentiment analysis, directly predicting sentiment graphs from text. We advance the state of the art on 4 out of 5 standard benchmark sets. We release the source code, models and predictions.

11:00-12:30 (Forum)
#81 "You might think about slightly revising the title": Identifying Hedges in Peer-tutoring Interactions
Yann Raphalen, Chloé Clavel and Justine Cassell
Hedges have an important role in the management of rapport. In peer-tutoring, they are notably used by tutors in dyads experiencing low rapport to tone down the impact of instructions and negative feedback. Pursuing the objective of building a tutoring agent that manages rapport with teenagers in order to improve learning, we used a multimodal peer-tutoring dataset to construct a computational framework for identifying hedges. We compared approaches relying on pre-trained resources with others that integrate insights from the social science literature. Our best performance involved a hybrid approach that outperforms the existing baseline while being easier to interpret. We employ a model explainability tool to explore the features that characterize hedges in peer-tutoring conversations, and we identify some novel features, and the benefits of a such a hybrid model approach.

11:00-12:30 (Forum)
#82 Incorporating Stock Market Signals for Twitter Stance Detection
Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd and Nigel Collier
Research in stance detection has so far focused on models which leverage purely textual input. In this paper, we investigate the integration of textual and financial signals for stance detection in the financial domain. Specifically, we propose a robust multi-task neural architecture that combines textual input with high-frequency intra-day time series from stock market prices. Moreover, we extend wt–wt, an existing stance detection dataset which collects tweets discussing Mergers and Acquisitions operations, with the relevant financial signal. Importantly, the obtained dataset aligns with Stander, an existing news stance detection dataset, thus resulting in a unique multimodal, multi-genre stance detection resource. We show experimentally and through detailed result analysis that our stance detection system benefits from financial information, and achieves state-of-the-art results on the wt–wt dataset: this demonstrates that the combination of multiple input signals is effective for cross-target stance detection, and opens interesting research directions for future work.

Lunch Break
12:30-13:30 - **Auditorium** (Lunch is not served)

ACL Business Meeting + Panel on the Future of Reviewing in NLP
13:30-15:00 - **Auditorium** (Auditorium)

Mini Break
15:00-15:15 - **Auditorium** (Forum)

Session 5 - 15:15-16:15

Language Grounding, Speech and Multimodality 2

15:15-16:15 (Liffey Hall 1)

15:15-15:30 (Liffey Hall 1)
CARETS: A Consistency And Robustness Evaluative Test Suite for VQA
Carlos E Jimenez, Olga Russakovsky and Karthik R Narasimhan
We introduce CARETS, a systematic test suite to measure consistency and robustness of modern VQA models through a series of six fine-grained capability tests. In contrast to existing VQA test sets, CARETS features balanced question generation to create pairs of instances to test models, with each pair focusing on a specific capability such as rephrasing, logical symmetry or image obfuscation. We evaluate six modern VQA systems on CARETS and identify several actionable weaknesses in model comprehension, especially with concepts such as negation, disjunction, or hypernym invariance. Interestingly, even the most sophisticated models are sensitive to aspects such as swapping the order of terms in a conjunction or varying the number of answer choices mentioned in the question. We release CARETS to be used as an extensible tool for evaluating multi-modal model robustness.

15:30-15:45 (Liffey Hall 1)
VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena
Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto and Albert Gatt
We propose VALSE (Vision And Language Structured Evaluation), a novel benchmark designed for testing general-purpose pretrained vision and language (V&L) models for their visio-linguistic grounding capabilities on specific linguistic phenomena. VALSE offers a suite of six tests covering various linguistic constructs. Solving these requires models to ground linguistic phenomena in the visual modality, allowing more fine-grained evaluations than hitherto possible. We build VALSE using methods that support the construction of valid foils, and report results from evaluating five widely-used V&L models. Our experiments suggest that current models have considerable difficulty addressing most phenomena. Hence, we expect VALSE to serve as an important benchmark to measure future progress of pretrained V&L models from a linguistic perspective, complementing the canonical task-centred V&L evaluations.

15:45-16:00 (Liffey Hall 1)
Searching for fingerspelled content in American Sign Language
Bowen Shi, Diane Brentari, Greg Shakhnarovich and Karen Livescu
Natural language processing for sign language video—including tasks like recognition, translation, and search—is crucial for making artificial intelligence technologies accessible to deaf individuals, and is gaining research interest in recent years. In this paper, we address the problem of searching for fingerspelled keywords or key phrases in raw sign language videos. This is an important task since significant content in sign language is often conveyed via fingerspelling, and to our knowledge the task has not been studied before. We propose an end-to-end model for this task, FSS-Net, that jointly detects fingerspelling and matches it to a text sequence. Our experiments, done on a large public dataset of ASL fingerspelling in the wild, show the importance of fingerspelling detection as a component of a search and retrieval model. Our model significantly outperforms baseline methods adapted from prior work on related tasks.

16:00-16:15 (Liffey Hall 1)
Contrastive Visual Semantic Pretraining Magnifies the Semantics of Natural Language Representations
Robert Wolfe and Aylin Caliskan
We examine the effects of contrastive visual semantic pretraining by comparing the geometry and semantic properties of contextualized English language representations formed by GPT-2 and CLIP, a zero-shot multimodal image classifier which adapts the GPT-2 architecture to encode image captions. We find that contrastive visual semantic pretraining significantly mitigates the anisotropy found in contextualized word embeddings from GPT-2, such that the intra-layer self-similarity (mean pairwise cosine similarity) of CLIP word embeddings is under $.25$ in all layers, compared to greater than $.95$ in the top layer of GPT-2. CLIP word embeddings outperform GPT-2 on word-level semantic intrinsic evaluation tasks, and achieve a new corpus-based state of the art for the RG65 evaluation, at $.88$. CLIP also forms fine-grained semantic representations of sentences, and obtains Spearman's $\rho = .73$ on the SemEval-2017 Semantic Textual Similarity Benchmark with no fine-tuning, compared to no greater than $\rho = .45$ in any layer of GPT-2. Finally, intra-layer self-similarity of CLIP sentence embeddings decreases as the layer index increases, finishing at $.25$ in the top layer, while the self-similarity of GPT-2 sentence embeddings formed using the EOS token increases layer-over-layer and never falls below $.97$. Our results indicate that high anisotropy is not an inevitable consequence of contextualization, and that visual semantic pretraining is beneficial not only for ordering visual representations, but also for encoding useful semantic representations of language, both on the word level and the sentence level.

Machine Learning for NLP 4

15:15-16:15 (The Liffey B)

15:15-15:30 (The Liffey B)
E-LANG: Energy-Based Joint Inferencing of Super and Swift Language Models
Mohammad Akbari, Amin Banitalebi-Dehkordi and Yong Zhang
Building huge and highly capable language models has been a trend in the past years. Despite their great performance, they incur high computational cost. A common solution is to apply model compression or choose light-weight architectures, which often need a separate fixed-size model for each desirable computational budget, and may lose performance in case of heavy compression. This paper proposes an effective dynamic inference approach, called E-LANG, which distributes the inference between large accurate Super-models and light-weight Swift models. To this end, a decision making module routes the inputs to Super or Swift models based on the energy characteristics of the representations in the latent space. This method is easily adoptable and architecture agnostic. As such, it can be applied to black-box pre-trained models without a need for architectural manipulations, reassembling of modules, or re-training. Unlike existing methods that are only applicable to encoder-only backbones and classification tasks, our method also works for encoder-decoder structures and sequence-to-sequence tasks such as translation. The E-LANG performance is verified through a set of experiments with T5 and BERT backbones on GLUE, SuperGLUE, and WMT. In particular, we outperform T5-11B with an average computations speed-up of 3.3X on GLUE and 2.9X on SuperGLUE. We also achieve BERT-based SOTA on GLUE with 3.2X less computations. Code and demo are available in supplementary materials.

15:30-15:45 (The Liffey B)
Learning Disentangled Textual Representations via Statistical Measures of Similarity
Pierre Colombo, Guillaume Staerman, Nathan Noiry and Pablo Piantanida
When working with textual data, a natural application of disentangled representations is the fair classification where the goal is to make predictions without being biased (or influenced) by sensible attributes that may be present in the data (e.g., age, gender or race). Dominant approaches to disentangle a sensitive attribute from textual representations rely on learning simultaneously a penalization term that involves either an adversary loss (e.g., a discriminator) or an information measure (e.g., mutual information). However, these methods require the training of a deep neural network with several parameter updates for each update of the representation model. As a matter of fact, the resulting nested optimization loop is both times consuming, adding complexity to the optimization dynamic, and requires a fine hyperparameter selection (e.g., learning rates, architecture). In this work, we introduce a family of regularizers for learning disentangled representations that do not require training. These regularizers are based on statistical measures of similarity between the conditional probability distributions with respect to the sensible attributes. Our novel regularizers do not require additional training, are faster and do not involve additional tuning while achieving better results both when combined with pretrained and randomly initialized text encoders.

15:45-16:00 (The Liffey B)
Uncertainty Estimation of Transformer Predictions for Misclassification Detection
Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gennadjevich Gusev, Mikhail Burtsev, Manvel Avetisian and Leonid Zhukov
Uncertainty estimation (UE) of model predictions is a crucial step for a variety of tasks such as active learning, misclassification detection, adversarial attack detection, out-of-distribution detection, etc. Most of the works on modeling the uncertainty of deep neural networks evaluate these methods on image classification tasks. Little attention has been paid to UE in natural language processing. To fill this gap, we perform a vast empirical investigation of state-of-the-art UE methods for Transformer models on misclassification detection in named entity recognition and text classification tasks and propose two computationally efficient modifications, one of which approaches or even outperforms computationally intensive methods.

16:00-16:10 (The Liffey B)
When to Use Multi-Task Learning vs Intermediate Fine-Tuning for Pre-Trained Encoder Transfer Learning
Orion Weller, Kevin Seppi and Matt Gardner
Transfer learning (TL) in natural language processing (NLP) has seen a surge of interest in recent years, as pre-trained models have shown an impressive ability to transfer to novel tasks. Three main strategies have emerged for making use of multiple supervised datasets during fine-tuning: training on an intermediate task before training on the target task (STILTs), using multi-task learning (MTL) to train jointly on a supplementary task and the target task (pairwise MTL), or simply using MTL to train jointly on all available datasets (MTL-ALL). In this work, we compare all three TL methods in a comprehensive analysis on the GLUE dataset suite. We find that there is a simple heuristic for when to use one of these techniques over the other: pairwise MTL is better than STILTs when the target task has fewer instances than the supporting task and vice versa. We show that this holds true in more than 92

Machine Translation and Multilinguality 3

15:15-16:15 (The Liffey A)

15:15-15:30 (The Liffey A)
Composable Sparse Fine-Tuning for Cross-Lingual Transfer
Alan Ansell, Edoardo Ponti, Anna Korhonen and Ivan Vulić
Fine-tuning the entire set of parameters of a large pretrained model has become the mainstream approach for transfer learning. To increase its efficiency and prevent catastrophic forgetting and interference, techniques like adapters and sparse fine-tuning have been developed. Adapters are modular, as they can be combined to adapt a model towards different facets of knowledge (e.g., dedicated language and/or task adapters). Sparse fine-tuning is expressive, as it controls the behavior of all model components. In this work, we introduce a new fine-tuning method with both these desirable properties. In particular, we learn sparse, real-valued masks based on a simple variant of the Lottery Ticket Hypothesis. Task-specific masks are obtained from annotated data in a source language, and language-specific masks from masked language modeling in a target language. Both these masks can then be composed with the pretrained model. Unlike adapter-based fine-tuning, this method neither increases the number of parameters at inference time nor alters the original model architecture. Most importantly, it outperforms adapters in zero-shot cross-lingual transfer by a large margin in a series of multilingual benchmarks, including Universal Dependencies, MasakhaNER, and AmericasNLI. Based on an in-depth analysis, we additionally find that sparsity is crucial to prevent both 1) interference between the fine-tunings to be composed and 2) overfitting. We release the code and models at https://github.com/cambridgeltl/composable-sft.

15:30-15:45 (The Liffey A)
Match the Script, Adapt if Multilingual: Analyzing the Effect of Multilingual Pretraining on Cross-lingual Transferability
Yoshinari Fujinuma, Jordan Lee Boyd-Graber and Katharina Kann
Pretrained multilingual models enable zero-shot learning even for unseen languages, and that performance can be further improved via adaptation prior to finetuning. However, it is unclear how the number of pretraining languages influences a model's zero-shot learning for languages unseen during pretraining. To fill this gap, we ask the following research questions: (1) How does the number of pretraining languages influence zero-shot performance on unseen target languages? (2) Does the answer to that question change with model adaptation? (3) Do the findings for our first question change if the languages used for pretraining are all related? Our experiments on pretraining with related languages indicate that choosing a diverse set of languages is crucial. Without model adaptation, surprisingly, increasing the number of pretraining languages yields better results up to adding related languages, after which performance plateaus. In contrast, with model adaptation via continued pretraining, pretraining on a larger number of languages often gives further improvement, suggesting that model adaptation is crucial to exploit additional pretraining languages.

15:45-16:00 (The Liffey A)
Multilingual Generative Language Models for Zero-Shot Cross-Lingual Event Argument Extraction
Kuan-Hao Huang, I-Hung Hsu, Prem Natarajan, Kai-Wei Chang and Nanyun Peng
We present a study on leveraging multilingual pre-trained generative language models for zero-shot cross-lingual event argument extraction (EAE). By formulating EAE as a language generation task, our method effectively encodes event structures and captures the dependencies

between arguments. We design language-agnostic templates to represent the event argument structures, which are compatible with any language, hence facilitating the cross-lingual transfer. Our proposed model finetunes multilingual pre-trained generative language models to generate sentences that fill in the language-agnostic template with arguments extracted from the input passage. The model is trained on source languages and is then directly applied to target languages for event argument extraction. Experiments demonstrate that the proposed model outperforms the current state-of-the-art models on zero-shot cross-lingual EAE. Comprehensive studies and error analyses are presented to better understand the advantages and the current limitations of using generative language models for zero-shot cross-lingual transfer EAE.

16:00-16:15 (The Liffey A)
[CL] **Challenges of Neural Machine Translation for Short Texts**
Yu Wan, Baosong Yang, Derek Fai Wong, Lidia Sam Chao, Liang Yao, Haibo Zhang and Boxing Chen

Question Answering 2

15:15-16:15 (Wicklow Hall 1)

15:15-15:30 (Wicklow Hall 1)
Sequence-to-Sequence Knowledge Graph Completion and Question Answering
Apoorv Umang Saxena, Adrian Kochsiek and Rainer Gemulla
Knowledge graph embedding (KGE) models represent each entity and relation of a knowledge graph (KG) with low-dimensional embedding vectors. These methods have recently been applied to KG link prediction and question answering over incomplete KGs (KGQA). KGEs typically create an embedding for each entity in the graph, which results in large model sizes on real-world graphs with millions of entities. For downstream tasks these atomic entity representations often need to be integrated into a multi stage pipeline, limiting their utility. We show that an off-the-shelf encoder-decoder Transformer model can serve as a scalable and versatile KGE model obtaining state-of-the-art results for KG link prediction and incomplete KG question answering. We achieve this by posing KG link prediction as a sequence-to-sequence task and exchange the triple scoring approach taken by prior KGE methods with autoregressive decoding. Such a simple but powerful method reduces the model size up to 98

15:30-15:45 (Wicklow Hall 1)
KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering
Donghan Yu, Chenguang Zhu, Yuwei Fang, Wenhao Yu, Shuohang Wang, Yichong Xu, Xiang Ren, Yiming Yang and Michael Zeng
Current Open-Domain Question Answering (ODQA) models typically include a retrieving module and a reading module, where the retriever selects potentially relevant passages from open-source documents for a given question, and the reader produces an answer based on the retrieved passages. The recently proposed Fusion-in-Decoder (FiD) framework is a representative example, which is built on top of a dense passage retriever and a generative reader, achieving the state-of-the-art performance. In this paper we further improve the FiD approach by introducing a knowledge-enhanced version, namely KG-FiD. Our new model uses a knowledge graph to establish the structural relationship among the retrieved passages, and a graph neural network (GNN) to re-rank the passages and select only a top few for further processing. Our experiments on common ODQA benchmark datasets (Natural Questions and TriviaQA) demonstrate that KG-FiD can achieve comparable or better performance in answer prediction than FiD, with less than 40

15:45-16:00 (Wicklow Hall 1)
RNG-KBQA: Generation Augmented Iterative Ranking for Knowledge Base Question Answering
Xi Ye, Semih Yavuz, Kazuma Hashimoto, Yingbo Zhou and Caiming Xiong
Existing KBQA approaches, despite achieving strong performance on i.i.d. test data, often struggle in generalizing to questions involving unseen KB schema items. Prior ranking-based approaches have shown some success in generalization, but suffer from the coverage issue. We present RnG-KBQA, a Rank-and-Generate approach for KBQA, which remedies the coverage issue with a generation model while preserving a strong generalization capability. Our approach first uses a contrastive ranker to rank a set of candidate logical forms obtained by searching over the knowledge graph. It then introduces a tailored generation model conditioned on the question and the top-ranked candidates to compose the final logical form. We achieve new state-of-the-art results on GrailQA and WebQSP datasets. In particular, our method surpasses the prior state-of-the-art by a large margin on the GrailQA leaderboard. In addition, RnG-KBQA outperforms all prior approaches on the popular WebQSP benchmark, even including the ones that use the oracle entity linking. The experimental results demonstrate the effectiveness of the interplay between ranking and generation, which leads to the superior performance of our proposed approach across all settings with especially strong improvements in zero-shot generalization.

16:00-16:15 (Wicklow Hall 1)
[TACL] **Time-Aware Language Models as Temporal Knowledge Bases**
Bhuwan Dhingra, Jeremy Cole, Julian Eisenschlos, Daniel Gillick, Jacob Eisenstein and William Cohen

Resources and Evaluation 3

15:15-16:15 (Liffey Hall 2)

15:15-15:30 (Liffey Hall 2)
[TACL] **Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics**
Paula Czarnowska, Yogarshi Vyas and Kashif Shah

15:30-15:45 (Liffey Hall 2)
[TACL] **Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations**
Aida Mostafazadeh Davani, Mark Díaz and Vinodkumar Prabhakaran

15:45-16:00 (Liffey Hall 2)
[CL] **Annotation Curricula to Implicitly Train Non-Expert Annotators**
Ji-Ung Lee, Jan-Christoph Klie, Iryna Gurevych

16:00-16:10 (Liffey Hall 2)
k-Rater Reliability: The Correct Unit of Reliability for Aggregated Human Annotations
Ka Wong and Praveen Paritosh
Since the inception of crowdsourcing, aggregation has been a common strategy for dealing with unreliable data. Aggregate ratings are more reliable than individual ones. However, many Natural Language Processing (NLP) applications that rely on aggregate ratings only report the reliability of individual ratings, which is the incorrect unit of analysis. In these instances, the data reliability is under-reported, and a proposed *k*-rater reliability (kRR) should be used as the correct data reliability for aggregated datasets. It is a multi-rater generalization of inter-rater reliability (IRR). We conducted two replications of the WordSim-353 benchmark, and present empirical, analytical, and bootstrap-based methods for computing kRR on WordSim-353. These methods produce very similar results. We hope this discussion will nudge researchers to report kRR in addition to IRR.

Semantics 2

15:15-16:15 (Wicklow Hall 2a)

15:15-15:30 (Wicklow Hall 2a)
Fully-Semantic Parsing and Generation: the BabelNet Meaning Representation
Abelardo Carlos Martinez Lorenzo, Marco Maru and Roberto Navigli
A language-independent representation of meaning is one of the most coveted dreams in Natural Language Understanding. With this goal in mind, several formalisms have been proposed as frameworks for meaning representation in Semantic Parsing. And yet, the dependencies these formalisms share with respect to language-specific repositories of knowledge make the objective of closing the gap between high- and low-resourced languages hard to accomplish. In this paper, we present the BabelNet Meaning Representation (BMR), an interlingual formalism that abstracts away from language-specific constraints by taking advantage of the multilingual semantic resources of BabelNet and VerbAtlas. We describe the rationale behind the creation of BMR and put forward BMR 1.0, a dataset labeled entirely according to the new formalism. Moreover, we show how BMR is able to outperform previous formalisms thanks to its fully-semantic framing, which enables top-notch multilingual parsing and generation. We release the code at https://github.com/SapienzaNLP/bmr.

15:30-15:45 (Wicklow Hall 2a)
Probing for Predicate Argument Structures in Pretrained Language Models
Simone Conia and Roberto Navigli
Thanks to the effectiveness and wide availability of modern pretrained language models (PLMs), recently proposed approaches have achieved remarkable results in dependency- and span-based, multilingual and cross-lingual Semantic Role Labeling (SRL). These results have prompted researchers to investigate the inner workings of modern PLMs with the aim of understanding how, where, and to what extent they encode information about SRL. In this paper, we follow this line of research and probe for predicate argument structures in PLMs. Our study shows that PLMs do encode semantic structures directly into the contextualized representation of a predicate, and also provides insights into the correlation between predicate senses and their structures, the degree of transferability between nominal and verbal structures, and how such structures are encoded across languages. Finally, we look at the practical implications of such insights and demonstrate the benefits of embedding predicate argument structure information into an SRL model.

15:45-16:00 (Wicklow Hall 2a)
ExtEnD: Extractive Entity Disambiguation
Edoardo Barba, Luigi Procopio and Roberto Navigli
Local models for Entity Disambiguation (ED) have today become extremely powerful, in most part thanks to the advent of large pre-trained language models. However, despite their significant performance achievements, most of these approaches frame ED through classification formulations that have intrinsic limitations, both computationally and from a modeling perspective. In contrast with this trend, here we propose ExtEnD, a novel local formulation for ED where we frame this task as a text extraction problem, and present two Transformer-based architectures that implement it. Based on experiments in and out of domain, and training over two different data regimes, we find our approach surpasses all its competitors in terms of both data efficiency and raw performance. ExtEnD outperforms its alternatives by as few as 6 F1 points on the more constrained of the two data regimes and, when moving to the other higher-resourced regime, sets a new state of the art on 4 out of 4 benchmarks under consideration, with average improvements of 0.7 F1 points overall and 1.1 F1 points out of domain. In addition, to gain better insights from our results, we also perform a fine-grained evaluation of our performances on different classes of label frequency, along with an ablation study of our architectural choices and an error analysis. We release our code and models for research purposes at https://github.com/SapienzaNLP/extend.

16:00-16:15 (Wicklow Hall 2a)
[TACL] **Weisfeiler-Leman in the BAMBOO: Novel AMR Graph Metrics and a Benchmark for AMR Graph Similarity**
Juri Opitz, Anette Frank and Angel Daza

Information Extraction 2

15:15-16:10 (Wicklow Hall 2b)

15:15-15:30 (Wicklow Hall 2b)
MILIE: Modular & Iterative Multilingual Open Information Extraction

Bhushan Kotnis, Kiril Gashteovski, Daniel Onoro Rubio, Ammar Shaker, Vanesa Rodriguez-Tembras, Makoto Takamoto, Mathias Niepert and Carolin Lawrence
Open Information Extraction (OpenIE) is the task of extracting (subject, predicate, object) triples from natural language sentences. Current OpenIE systems extract all triple slots independently. In contrast, we explore the hypothesis that it may be beneficial to extract triple slots iteratively: first extract easy slots, followed by the difficult ones by conditioning on the easy slots, and therefore achieve a better overall extraction. Based on this hypothesis, we propose a neural OpenIE system, MILIE, that operates in an iterative fashion. Due to the iterative nature, the system is also modularit is possible to seamlessly integrate rule based extraction systems with a neural end-to-end system, thereby allowing rule based systems to supply extraction slots which MILIE can leverage for extracting the remaining slots. We confirm our hypothesis empirically: MILIE outperforms SOTA systems on multiple languages ranging from Chinese to Arabic. Additionally, we are the first to provide an OpenIE test dataset for Arabic and Galician.

15:30-15:45 (Wicklow Hall 2b)
[TACL] **Multilingual Autoregressive Entity Linking**
Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel and Fabio Petroni

15:45-16:00 (Wicklow Hall 2b)
Prix-LM: Pretraining for Multilingual Knowledge Base Construction
Wenxuan Zhou, Fangyu Liu, Ivan Vulić, Nigel Collier and Muhao Chen
Knowledge bases (KBs) contain plenty of structured world and commonsense knowledge. As such, they often complement distributional text-based information and facilitate various downstream tasks. Since their manual construction is resource- and time-intensive, recent efforts have tried leveraging large pretrained language models (PLMs) to generate additional monolingual knowledge facts for KBs. However, such methods have not been attempted for building and enriching multilingual KBs. Besides wider application, such multilingual KBs can provide richer combined knowledge than monolingual (e.g., English) KBs. Knowledge expressed in different languages may be complementary and unequally distributed: this implies that the knowledge available in high-resource languages can be transferred to low-resource ones. To achieve this, it is crucial to represent multilingual knowledge in a shared/unified space. To this end, we propose a unified representation model, Prix-LM, for multilingual KB construction and completion. We leverage two types of knowledge, monolingual triples and cross-lingual links, extracted from existing multilingual KBs, and tune a multilingual language encoder XLM-R via a causal language modeling objective. Prix-LM integrates useful multilingual and KB-based factual knowledge into a single model. Experiments on standard entity-related tasks, such as link prediction in multiple languages, cross-lingual entity linking and bilingual lexicon induction, demonstrate its effectiveness, with gains reported over strong task-specialised baselines.

16:00-16:10 (Wicklow Hall 2b)
Towards Consistent Document-level Entity Linking: Joint Models for Entity Linking and Coreference Resolution
Klim Zaporojets, Johannes Deleu, Yiwei Jiang, Thomas Demeester and Chris Develder
We consider the task of document-level entity linking (EL), where it is important to make consistent decisions for entity mentions over the full document jointly. We aim to leverage explicit "connections" among mentions within the document itself: we propose to join EL and coreference resolution (coref) in a single structured prediction task over directed trees and use a globally normalized model to solve it. This contrasts with related works where two separate models are trained for each of the tasks and additional logic is required to merge the outputs. Experimental results on two datasets show a boost of up to +5

Poster Session 5: Dialogue and Interactive Systems

15:15-16:15 (Forum)

12:00-12:15 (Forum)
#1 Think Before You Speak: Explicitly Generating Implicit Commonsense Knowledge for Response Generation
Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu and Dilek Hakkani-Tur
Implicit knowledge, such as common sense, is key to fluid human conversations. Current neural response generation (RG) models are trained to generate responses directly, omitting unstated implicit knowledge. In this paper, we present Think-Before-Speaking (TBS), a generative approach to first externalize implicit commonsense knowledge ($think$) and use this knowledge to generate responses ($speak$). We argue that externalizing implicit knowledge allows more efficient learning, produces more informative responses, and enables more explainable models. We analyze different choices to collect knowledge-aligned dialogues, represent implicit knowledge, and transition between knowledge and dialogues. Empirical results show TBS models outperform end-to-end and knowledge-augmented RG baselines on most automatic metrics and generate more informative, specific, and commonsense-following responses, as evaluated by human annotators. TBS also generates $knowledge$ that makes sense and is relevant to the dialogue around 85

15:15-16:15 (Forum)
#2 Dynamic Schema Graph Fusion Network for Multi-Domain Dialogue State Tracking
Yue Feng, Aldo Lipani, Fanghua Ye, Qiang Zhang and Emine Yilmaz
Dialogue State Tracking (DST) aims to keep track of users' intentions during the course of a conversation. In DST, modelling the relations among domains and slots is still an under-studied problem. Existing approaches that have considered such relations generally fall short in: (1) fusing prior slot-domain membership relations and dialogue-aware dynamic slot relations explicitly, and (2) generalizing to unseen domains. To address these issues, we propose a novel **D**ynamic **S**chema **G**raph **F**usion **Net**work (**DSGFNet**), which generates a dynamic schema graph to explicitly fuse the prior slot-domain membership relations and dialogue-aware dynamic slot relations. It also uses the schemata to facilitate knowledge transfer to new domains. DSGFNet consists of a dialogue utterance encoder, a schema graph encoder, a dialogue-aware schema graph evolving network, and a schema graph enhanced dialogue state decoder. Empirical results on benchmark datasets (i.e., SGD, MultiWOZ2.1, and MultiWOZ2.2), show that DSGFNet outperforms existing methods.

15:15-16:15 (Forum)
#3 Internet-Augmented Dialogue Generation
Mojtaba Komeili, Kurt Shuster and Jason E Weston
The largest store of continually updating knowledge on our planet can be accessed via internet search. In this work we study giving access

to this information to conversational agents. Large language models, even though they store an impressive amount of knowledge within their weights, are known to hallucinate facts when generating dialogue (Shuster et al., 2021); moreover, those facts are frozen in time at the point of model training. In contrast, we propose an approach that learns to generate an internet search query based on the context, and then conditions on the search results to finally generate a response, a method that can employ up-to-the-minute relevant information. We train and evaluate such models on a newly collected dataset of human-human conversations whereby one of the speakers is given access to internet search during knowledgedriven discussions in order to ground their responses. We find that search-query based access of the internet in conversation provides superior performance compared to existing approaches that either use no augmentation or FAISS-based retrieval (Lewis et al., 2020b).

15:15-16:15 (Forum)
#4 SaFeRDialogues: Taking Feedback Gracefully after Conversational Safety Failures
Megan Ung, Jing Xu and Y-Lan Boureau
Current open-domain conversational models can easily be made to talk in inadequate ways. Online learning from conversational feedback given by the conversation partner is a promising avenue for a model to improve and adapt, so as to generate fewer of these safety failures. However, current state-of-the-art models tend to react to feedback with defensive or oblivious responses. This makes for an unpleasant experience and may discourage conversation partners from giving feedback in the future. This work proposes SaFeRDialogues, a task and dataset of graceful responses to conversational feedback about safety failures. We collect a dataset of 8k dialogues demonstrating safety failures, feedback signaling them, and a response acknowledging the feedback. We show how fine-tuning on this dataset results in conversations that human raters deem considerably more likely to lead to a civil conversation, without sacrificing engagingness or general conversational ability.

15:15-16:15 (Forum)
#5 DEAM: Dialogue Coherence Evaluation using AMR-based Semantic Manipulations
Sarik Ghazarian, Nuan Wen, Aram Galstyan and Nanyun Peng
Automatic evaluation metrics are essential for the rapid development of open-domain dialogue systems as they facilitate hyper-parameter tuning and comparison between models. Although recently proposed trainable conversation-level metrics have shown encouraging results, the quality of the metrics is strongly dependent on the quality of training data. Prior works mainly resort to heuristic text-level manipulations (e.g. utterances shuffling) to bootstrap incoherent conversations (negative examples) from coherent dialogues (positive examples). Such approaches are insufficient to appropriately reflect the incoherence that occurs in interactions between advanced dialogue models and humans. To tackle this problem, we propose DEAM, a Dialogue coherence Evaluation metric that relies on Abstract Meaning Representation (AMR) to apply semantic-level Manipulations for incoherent (negative) data generation. AMRs naturally facilitate the injection of various types of incoherence sources, such as coreference inconsistency, irrelevancy, contradictions, and decrease engagement, at the semantic level, thus resulting in more natural incoherent samples. Our experiments show that DEAM achieves higher correlations with human judgments compared to baseline methods on several dialog datasets by significant margins. We also show that DEAM can distinguish between coherent and incoherent dialogues generated by baseline manipulations, whereas those baseline models cannot detect incoherent examples generated by DEAM. Our results demonstrate the potential of AMR-based semantic manipulations for natural negative example generation.

15:15-16:15 (Forum)
#6 Towards Large-Scale Interpretable Knowledge Graph Reasoning for Dialogue Systems
Yi-Lin Tuan, Sajjad Beygi, Maryam Fazel-Zarandi, Qiaozi Gao, Alessandra Cervone and William Yang Wang
Users interacting with voice assistants today need to phrase their requests in a very specific manner to elicit an appropriate response. This limits the user experience, and is partly due to the lack of reasoning capabilities of dialogue platforms and the hand-crafted rules that require extensive labor. One possible solution to improve user experience and relieve the manual efforts of designers is to build an end-to-end dialogue system that can do reasoning itself while perceiving user's utterances. In this work, we propose a novel method to incorporate the knowledge reasoning capability into dialog systems in a more scalable and generalizable manner. Our proposed method allows a single transformer model to directly walk on a large-scale knowledge graph to generate responses. To the best of our knowledge, this is the first work to have transformer models generate responses by reasoning over differentiable knowledge graphs. We investigate the reasoning abilities of the proposed method on both task-oriented and domain-specific chit-chat dialogues. Empirical results show that this method can effectively and efficiently incorporate a knowledge graph into a dialogue system with fully-interpretable reasoning paths.

15:15-16:15 (Forum)
#7 DS-TOD: Efficient Domain Specialization for Task-Oriented Dialog
Chia-Chien Hung, Anne Lauscher, Simone Paolo Ponzetto and Goran Glavaš
Recent work has shown that self-supervised dialog-specific pretraining on large conversational datasets yields substantial gains over traditional language modeling (LM) pretraining in downstream task-oriented dialog (TOD). These approaches, however, exploit general dialogic corpora (e.g., Reddit) and thus presumably fail to reliably embed domain-specific knowledge useful for concrete downstream TOD domains. In this work, we investigate the effects of domain specialization of pretrained language models (PLMs) for TOD. Within our DS-TOD framework, we first automatically extract salient domain-specific terms, and then use them to construct DomainCC and DomainReddit – resources that we leverage for domain-specific pretraining, based on (i) masked language modeling (MLM) and (ii) response selection (RS) objectives, respectively. We further propose a resource-efficient and modular domain specialization by means of domain adapters – additional parameter-light layers in which we encode the domain knowledge. Our experiments with prominent TOD tasks – dialog state tracking (DST) and response retrieval (RR) – encompassing five domains from the MultiWOZ benchmark demonstrate the effectiveness of DS-TOD. Moreover, we show that the light-weight adapter-based specialization (1) performs comparably to full fine-tuning in single domain setups and (2) is particularly suitable for multi-domain specialization, where besides advantageous computational footprint, it can offer better TOD performance.

15:15-16:15 (Forum)
#8 ASSIST: Towards Label Noise-Robust Dialogue State Tracking
Fanghua Ye, Yue Feng and Emine Yilmaz
The MultiWOZ 2.0 dataset has greatly boosted the research on dialogue state tracking (DST). However, substantial noise has been discovered in its state annotations. Such noise brings about huge challenges for training DST models robustly. Although several refined versions, including MultiWOZ 2.1-2.4, have been published recently, there are still lots of noisy labels, especially in the training set. Besides, it is costly to rectify all the problematic annotations. In this paper, instead of improving the annotation quality further, we propose a general framework, named ASSIST (lAbel noiSe-robuSt dIalogue State Tracking), to train DST models robustly from noisy labels. ASSIST first generates pseudo labels for each sample in the training set by using an auxiliary model trained on a small clean dataset, then puts the generated pseudo labels and vanilla noisy labels together to train the primary model. We show the validity of ASSIST theoretically. Experimental results also demonstrate that ASSIST improves the joint goal accuracy of DST by up to 28.16% on MultiWOZ 2.0 and 8.41% on MultiWOZ 2.4, compared

to using only the vanilla noisy labels.

15:15-16:15 (Forum)
[TACL] **#9 TopiOCQA: Open-domain Conversational Question Answering with Topic Switching**
Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm Vries and Siva Reddy

15:15-16:15 (Forum)
[TACL] **#10 Designing an Automatic Agent for Repeated Language based Persuasion Games**
Roi Reichart, Maya Raifer, Guy Rotman, Reut Apel and Moshe Tennenholtz

15:15-16:15 (Forum)
#11 Addressing Resource and Privacy Constraints in Semantic Parsing Through Data Augmentation
Kevin Yang, Olivia Deng, Charles Chen, Richard Shin, Subhro Roy and Benjamin Van Durme
We introduce a novel setup for low-resource task-oriented semantic parsing which incorporates several constraints that may arise in real-world scenarios: (1) lack of similar datasets/models from a related domain, (2) inability to sample useful logical forms directly from a grammar, and (3) privacy requirements for unlabeled natural utterances. Our goal is to improve a low-resource semantic parser using utterances collected through user interactions. In this highly challenging but realistic setting, we investigate data augmentation approaches involving generating a set of structured canonical utterances corresponding to logical forms, before simulating corresponding natural language and filtering the resulting pairs. We find that such approaches are effective despite our restrictive setup: in a low-resource setting on the complex SMCalFlow calendaring dataset (Andreas et al. 2020), we observe 33

15:15-16:15 (Forum)
#12 DialFact: A Benchmark for Fact-Checking in Dialogue
Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu and Caiming Xiong
Fact-checking is an essential tool to mitigate the spread of misinformation and disinformation. We introduce the task of fact-checking in dialogue, which is a relatively unexplored area. We construct DialFact, a testing benchmark dataset of 22,245 annotated conversational claims, paired with pieces of evidence from Wikipedia. There are three sub-tasks in DialFact: 1) Verifiable claim detection task distinguishes whether a response carries verifiable factual information; 2) Evidence retrieval task retrieves the most relevant Wikipedia snippets as evidence; 3) Claim verification task predicts a dialogue response to be supported, refuted, or not enough information. We found that existing fact-checking models trained on non-dialogue data like FEVER fail to perform well on our task, and thus, we propose a simple yet data-efficient solution to effectively improve fact-checking performance in dialogue. We point out unique challenges in DialFact such as handling the colloquialisms, coreferences, and retrieval ambiguities in the error analysis to shed light on future research in this direction.

15:15-16:15 (Forum)
#13 Achieving Conversational Goals with Unsupervised Post-hoc Knowledge Injection
Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick and Julian McAuley
A limitation of current neural dialog models is that they tend to suffer from a lack of specificity and informativeness in generated responses, primarily due to dependence on training data that covers a limited variety of scenarios and conveys limited knowledge. One way to alleviate this issue is to extract relevant knowledge from external sources at decoding time and incorporate it into the dialog response. In this paper, we propose a post-hoc knowledge-injection technique where we first retrieve a diverse set of relevant knowledge snippets conditioned on both the dialog history and an initial response from an existing dialog model. We construct multiple candidate responses, individually injecting each retrieved snippet into the initial response using a gradient-based decoding method, and then select the final response with an unsupervised ranking step. Our experiments in goal-oriented and knowledge-grounded dialog settings demonstrate that human annotators judge the outputs from the proposed method to be more engaging and informative compared to responses from prior dialog systems. We further show that knowledge-augmentation promotes success in achieving conversational goals in both experimental settings.

15:15-16:15 (Forum)
#14 CICERO: A Dataset for Contextualized Commonsense Inference in Dialogues
Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea and Soujanya Poria
This paper addresses the problem of dialogue reasoning with contextualized commonsense inference. We curate CICERO, a dataset of dyadic conversations with five types of utterance-level reasoning-based inferences: cause, subsequent event, prerequisite, motivation, and emotional reaction. The dataset contains 53,105 of such inferences from 5,672 dialogues. We use this dataset to solve relevant generative and discriminative tasks: generation of cause and subsequent event; generation of prerequisite, motivation, and listener's emotional reaction; and selection of plausible alternatives. Our results ascertain the value of such dialogue-centric commonsense knowledge datasets. It is our hope that CICERO will open new research avenues into commonsense-based dialogue reasoning.

15:15-16:15 (Forum)
#15 When did you become so smart, oh wise one?! Sarcasm Explanation in Multi-modal Multi-party Dialogues
Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar and Tanmoy Chakraborty
Indirect speech such as sarcasm achieves a constellation of discourse goals in human communication. While the indirectness of figurative language warrants speakers to achieve certain pragmatic goals, it is challenging for AI agents to comprehend such idiosyncrasies of human communication. Though sarcasm identification has been a well-explored topic in dialogue analysis, for conversational systems to truly grasp a conversation's innate meaning and generate appropriate responses, simply detecting sarcasm is not enough; it is vital to explain its underlying sarcastic connotation to capture its true essence. In this work, we study the discourse structure of sarcastic conversations and propose a novel task – Sarcasm Explanation in Dialogue (SED). Set in a multimodal and code-mixed setting, the task aims to generate natural language explanations of satirical conversations. To this end, we curate WITS, a new dataset to support our task. We propose MAF (Modality Aware Fusion), a multimodal context-aware attention and global information fusion module to capture multimodality and use it to benchmark WITS. The proposed attention module surpasses the traditional multimodal fusion baselines and reports the best performance on almost all metrics. Lastly, we carry out detailed analysis both quantitatively and qualitatively.

15:15-16:15 (Forum)
#16 Can Visual Dialogue Models Do Scorekeeping? Exploring How Dialogue Representations Incrementally Encode Shared Knowledge
Brielen Madureira and David Schlangen
Cognitively plausible visual dialogue models should keep a mental scoreboard of shared established facts in the dialogue context. We propose a theory-based evaluation method for investigating to what degree models pretrained on the VisDial dataset incrementally build representations that appropriately do scorekeeping. Our conclusion is that the ability to make the distinction between shared and privately known statements along the dialogue is moderately present in the analysed models, but not always incrementally consistent, which may partially be due to the limited need for grounding interactions in the original task.

15:15-16:15 (Forum)
#17 Situated Dialogue Learning through Procedural Environment Generation
Prithviraj Ammanabrolu, Renee Jia and Mark Riedl
We teach goal-driven agents to interactively act and speak in situated environments by training on generated curriculums. Our agents operate in LIGHT (Urbanek et al. 2019)—a large-scale crowd-sourced fantasy text adventure game wherein an agent perceives and interacts with the world through textual natural language. Goals in this environment take the form of character-based quests, consisting of personas and motivations. We augment LIGHT by learning to procedurally generate additional novel textual worlds and quests to create a curriculum of steadily increasing difficulty for training agents to achieve such goals. In particular, we measure curriculum difficulty in terms of the rarity of the quest in the original training distribution—an easier environment is one that is more likely to have been found in the unaugmented dataset. An ablation study shows that this method of learning from the tail of a distribution results in significantly higher generalization abilities as measured by zero-shot performance on never-before-seen quests.

15:15-16:15 (Forum)
#18 Mismatch between Multi-turn Dialogue and its Evaluation Metric in Dialogue State Tracking
Takyoung Kim, Hoonsang Yoon, Yukyung Lee, Pilsung Kang and Misuk Kim
Dialogue state tracking (DST) aims to extract essential information from multi-turn dialog situations and take appropriate actions. A belief state, one of the core pieces of information, refers to the subject and its specific content, and appears in the form of `domain-slot-value`. The trained model predicts "accumulated" belief states in every turn, and joint goal accuracy and slot accuracy are mainly used to evaluate the prediction; however, we specify that the current evaluation metrics have a critical limitation when evaluating belief states accumulated as the dialogue proceeds, especially in the most used MultiWOZ dataset. Additionally, we propose **relative slot accuracy** to complement existing metrics. Relative slot accuracy does not depend on the number of predefined slots, and allows intuitive evaluation by assigning relative scores according to the turn of each dialog. This study also encourages not solely the reporting of joint goal accuracy, but also various complementary metrics in DST tasks for the sake of a realistic evaluation.

15:15-16:15 (Forum)
#19 Where to Go for the Holidays: Towards Mixed-Type Dialogs for Clarification of User Goals
Zeming Liu, Jun Xu, Zeyang Lei, Haifeng Wang, Zheng-Yu Niu and Hua Wu
Most dialog systems posit that users have figured out clear and specific goals before starting an interaction. For example, users have determined the departure, the destination, and the travel time for booking a flight. However, in many scenarios, limited by experience and knowledge, users may know what they need, but still struggle to figure out clear and specific goals by determining all the necessary slots.
In this paper, we identify this challenge, and make a step forward by collecting a new human-to-human mixed-type dialog corpus. It contains 5k dialog sessions and 168k utterances for 4 dialog types and 5 domains. Within each session, an agent first provides user-goal-related knowledge to help figure out clear and specific goals, and then help achieve them.
Furthermore, we propose a mixed-type dialog model with a novel Prompt-based continual learning mechanism. Specifically, the mechanism enables the model to continually strengthen its ability on any specific type by utilizing existing dialog corpora effectively.

15:15-16:15 (Forum)
#20 GlobalWoZ: Globalizing MultiWoZ to Develop Multilingual Task-Oriented Dialogue Systems
Bosheng Ding, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si and Chunyan Miao
Over the last few years, there has been a move towards data curation for multilingual task-oriented dialogue (ToD) systems that can serve people speaking different languages. However, existing multilingual ToD datasets either have a limited coverage of languages due to the high cost of data curation, or ignore the fact that dialogue entities barely exist in countries speaking these languages. To tackle these limitations, we introduce a novel data curation method that generates GlobalWoZ — a large-scale multilingual ToD dataset globalized from an English ToD dataset for three unexplored use cases of multilingual ToD systems. Our method is based on translating dialogue templates and filling them with local entities in the target-language countries. Besides, we extend the coverage of target languages to 20 languages. We will release our dataset and a set of strong baselines to encourage research on multilingual ToD systems for real use cases.

15:15-16:15 (Forum)
#21 Multi-Task Pre-Training for Plug-and-Play Task-Oriented Dialogue System
Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai and Yi Zhang
Pre-trained language models have been recently shown to benefit task-oriented dialogue (TOD) systems. Despite their success, existing methods often formulate this task as a cascaded generation problem which can lead to error accumulation across different sub-tasks and greater data annotation overhead. In this study, we present PPTOD, a unified plug-and-play model for task-oriented dialogue. In addition, we introduce a new dialogue multi-task pre-training strategy that allows the model to learn the primary TOD task completion skills from heterogeneous dialog corpora. We extensively test our model on three benchmark TOD tasks, including end-to-end dialogue modelling, dialogue state tracking, and intent classification. Experimental results show that PPTOD achieves new state of the art on all evaluated tasks in both high-resource and low-resource scenarios. Furthermore, comparisons against previous SOTA methods show that the responses generated by PPTOD are more factually correct and semantically coherent as judged by human annotators.

15:15-16:15 (Forum)
#22 Achieving Reliable Human Assessment of Open-Domain Dialogue Systems
Tianbo Ji, Yvette Graham, Gareth J. F. Jones, Chenyang Lyu and Qun Liu
Evaluation of open-domain dialogue systems is highly challenging and development of better techniques is highlighted time and again as

desperately needed. Despite substantial efforts to carry out reliable live evaluation of systems in recent competitions, annotations have been abandoned and reported as too unreliable to yield sensible results. This is a serious problem since automatic metrics are not known to provide a good indication of what may or may not be a high-quality conversation. Answering the distress call of competitions that have emphasized the urgent need for better evaluation techniques in dialogue, we present the successful development of human evaluation that is highly reliable while still remaining feasible and low cost. Self-replication experiments reveal almost perfectly repeatable results with a correlation of $r = 0.969$. Furthermore, due to the lack of appropriate methods of statistical significance testing, the likelihood of potential improvements to systems occurring due to chance is rarely taken into account in dialogue evaluation, and the evaluation we propose facilitates application of standard tests. Since we have developed a highly reliable evaluation method, new insights into system performance can be revealed. We therefore include a comparison of state-of-the-art models (i) with and without personas, to measure the contribution of personas to conversation quality, as well as (ii) prescribed versus freely chosen topics. Interestingly with respect to personas, results indicate that personas do not positively contribute to conversation quality as expected.

15:15-16:15 (Forum)
#23 N-Shot Learning for Augmenting Task-Oriented Dialogue State Tracking
Ibrahim Taha Aksu, Zhengyuan Liu, Min-Yen Kan and Nancy F. Chen
Augmentation of task-oriented dialogues has followed standard methods used for plain-text such as back-translation, word-level manipulation, and paraphrasing despite its richly annotated structure. In this work, we introduce an augmentation framework that utilizes belief state annotations to match turns from various dialogues and form new synthetic dialogues in a bottom-up manner. Unlike other augmentation strategies, it operates with as few as five examples. Our augmentation strategy yields significant improvements when both adapting a DST model to a new domain, and when adapting a language model to the DST task, on evaluations with TRADE and TOD-BERT models. Further analysis shows that our model performs better on seen values during training, and it is also more robust to unseen values. We conclude that exploiting belief state annotations enhances dialogue augmentation and results in improved models in n-shot training scenarios.

15:15-16:15 (Forum)
#24 Towards Transparent Interactive Semantic Parsing via Step-by-Step Correction
Lingbo Mo, Ashley Lewis, Huan Sun and Michael White
Existing studies on semantic parsing focus on mapping a natural-language utterance to a logical form (LF) in one turn. However, because natural language may contain ambiguity and variability, this is a difficult challenge. In this work, we investigate an interactive semantic parsing framework that explains the predicted LF step by step in natural language and enables the user to make corrections through natural-language feedback for individual steps. We focus on question answering over knowledge bases (KBQA) as an instantiation of our framework, aiming to increase the transparency of the parsing process and help the user trust the final answer. We construct INSPIRED, a crowdsourced dialogue dataset derived from the ComplexWebQuestions dataset. Our experiments show that this framework has the potential to greatly improve overall parse accuracy. Furthermore, we develop a pipeline for dialogue simulation to evaluate our framework w.r.t. a variety of state-of-the-art KBQA models without further crowdsourcing effort. The results demonstrate that our framework promises to be effective across such models.

15:15-16:15 (Forum)
#25 VISITRON: Visual Semantics-Aligned Interactively Trained Object-Navigator
Ayush Shrivastava, Karthik Gopalakrishnan, Yang Liu, Robinson Piramuthu, Gokhan Tur, Devi Parikh and Dilek Hakkani-Tur
Interactive robots navigating photo-realistic environments need to be trained to effectively leverage and handle the dynamic nature of dialogue in addition to the challenges underlying vision-and-language navigation (VLN). In this paper, we present VISITRON, a multi-modal Transformer-based navigator better suited to the interactive regime inherent to Cooperative Vision-and-Dialog Navigation (CVDN). VISITRON is trained to: i) identify and associate object-level concepts and semantics between the environment and dialogue history, ii) identify when to interact vs. navigate via imitation learning of a binary classification head. We perform extensive pre-training and fine-tuning ablations with VISITRON to gain empirical insights and improve performance on CVDN. VISITRON's ability to identify when to interact leads to a natural generalization of the game-play mode introduced by Roman et al. (2020) for enabling the use of such models in different environments. VISITRON is competitive with models on the static CVDN leaderboard and attains state-of-the-art performance on the Success weighted by Path Length (SPL) metric.

15:15-16:15 (Forum)
#26 One Agent To Rule Them All: Towards Multi-agent Conversational AI
Christopher Clarke, Joseph J Peper, Karthik Krishnamurthy, Walter Talamonti, Kevin Leach, Walter Lasecki, Yiping Kang, Lingjia Tang and Jason Mars
The increasing volume of commercially available conversational agents (CAs) on the market has resulted in users being burdened with learning and adopting multiple agents to accomplish their tasks. Though prior work has explored supporting a multitude of domains within the design of a single agent, the interaction experience suffers due to the large action space of desired capabilities. To address these problems, we introduce a new task BBAI: Black-Box Agent Integration, focusing on combining the capabilities of multiple black-box CAs at scale. We explore two techniques: question agent pairing and question response pairing aimed at resolving this task. Leveraging these techniques, we design One For All (OFA), a scalable system that provides a unified interface to interact with multiple CAs. Additionally, we introduce MARS: Multi-Agent Response Selection, a new encoder model for question response pairing that jointly encodes user question and agent response pairs. We demonstrate that OFA is able to automatically and accurately integrate an ensemble of commercially available CAs spanning disparate domains. Specifically, using the MARS encoder we achieve the highest accuracy on our BBAI task, outperforming strong baselines.

15:15-16:15 (Forum)
#27 Data Augmentation and Learned Layer Aggregation for Improved Multilingual Language Understanding in Dialogue
Evgeniia Razumovskaia, Ivan Vulić and Anna Korhonen
Scaling dialogue systems to a multitude of domains, tasks and languages relies on costly and time-consuming data annotation for different domain-task-language configurations. The annotation efforts might be substantially reduced by the methods that generalise well in zero- and few-shot scenarios, and also effectively leverage external unannotated data sources (e.g., Web-scale corpora). We propose two methods to this aim, offering improved dialogue natural language understanding (NLU) across multiple languages: 1) Multi-SentAugment, and 2) LayerAgg. Multi-SentAugment is a self-training method which augments available (typically few-shot) training data with similar (automatically labelled) in-domain sentences from large monolingual Web-scale corpora. LayerAgg learns to select and combine useful semantic information scattered across different layers of a Transformer model (e.g., mBERT); it is especially suited for zero-shot scenarios as semantically richer representations should strengthen the model's cross-lingual capabilities. Applying the two methods with state-of-the-art NLU models obtains consistent improvements across two standard multilingual NLU datasets covering 16 diverse languages. The gains are observed in

zero-shot, few-shot, and even in full-data scenarios. The results also suggest that the two methods achieve a synergistic effect: the best overall performance in few-shot setups is attained when the methods are used together.

15:15-16:15 (Forum)
#28 Dialogue Summaries as Dialogue States (DS2), Template-Guided Summarization for Few-shot Dialogue State Tracking
Jamin Shin, Hangyeol Yu, Hyeongdon Moon, Andrea Madotto and Juneyoung Park
Annotating task-oriented dialogues is notorious for the expensive and difficult data collection process. Few-shot dialogue state tracking (DST) is a realistic solution to this problem. In this paper, we hypothesize that dialogue summaries are essentially unstructured dialogue states; hence, we propose to reformulate dialogue state tracking as a dialogue summarization problem. To elaborate, we train a text-to-text language model with synthetic template-based dialogue summaries, generated by a set of rules from the dialogue states. Then, the dialogue states can be recovered by inversely applying the summary generation rules. We empirically show that our method DS2 outperforms previous works on few-shot DST in MultiWoZ 2.0 and 2.1, in both cross-domain and multi-domain settings. Our method also exhibits vast speedup during both training and inference as it can generate all states at once. Finally, based on our analysis, we discover that the naturalness of the summary templates plays a key role for successful training.

15:15-16:15 (Forum)
#29 What is wrong with you?: Leveraging User Sentiment for Automatic Dialog Evaluation
Sarik Ghazarian, Behnam Hedayatnia, Alexandros Papangelis, Yang Liu and Dilek Hakkani-Tur
Accurate automatic evaluation metrics for open-domain dialogs are in high demand. Existing model-based metrics for system response evaluation are trained on human annotated data, which is cumbersome to collect. In this work, we propose to use information that can be automatically extracted from the next user utterance, such as its sentiment or whether the user explicitly ends the conversation, as a proxy to measure the quality of the previous system response. This allows us to train on a massive set of dialogs with weak supervision, without requiring manual system turn quality annotations. Experiments show that our model is comparable to models trained on human annotated data. Furthermore, our model generalizes across both spoken and written open-domain dialog corpora collected from real and paid users.

15:15-16:15 (Forum)
#30 The Moral Integrity Corpus: A Benchmark for Ethical Dialogue Systems
Caleb Ziems, Jane A. Yu, Yi-Chia Wang, Alon Y. Halevy and Diyi Yang
Conversational agents have come increasingly closer to human competence in open-domain dialogue settings; however, such models can reflect insensitive, hurtful, or entirely incoherent viewpoints that erode a user's trust in the moral integrity of the system. Moral deviations are difficult to mitigate because moral judgments are not universal, and there may be multiple competing judgments that apply to a situation simultaneously. In this work, we introduce a new resource, not to authoritatively resolve moral ambiguities, but instead to facilitate systematic understanding of the intuitions, values and moral judgments reflected in the utterances of dialogue systems. The Moral Integrity Corpus, MIC, is such a resource, which captures the moral assumptions of 38k prompt-reply pairs, using 99k distinct Rules of Thumb (RoTs). Each RoT reflects a particular moral conviction that can explain why a chatbot's reply may appear acceptable or problematic. We further organize RoTs with a set of 9 moral and social attributes and benchmark performance for attribute classification. Most importantly, we show that current neural language models can automatically generate new RoTs that reasonably describe previously unseen interactions, but they still struggle with certain scenarios. Our findings suggest that MIC will be a useful resource for understanding and language models' implicit moral assumptions and flexibly benchmarking the integrity of conversational agents. To download the data, see https://github.com/GT-SALT/mic

15:15-16:15 (Forum)
#31 HybriDialogue: An Information-Seeking Dialogue Dataset Grounded on Tabular and Textual Data
Kai Nakamura, Sharon Levy, Yi-Lin Tuan, Wenhu Chen and William Yang Wang
A pressing challenge in current dialogue systems is to successfully converse with users on topics with information distributed across different modalities. Previous work in multiturn dialogue systems has primarily focused on either text or table information. In more realistic scenarios, having a joint understanding of both is critical as knowledge is typically distributed over both unstructured and structured forms. We present a new dialogue dataset, HybriDialogue, which consists of crowdsourced natural conversations grounded on both Wikipedia text and tables. The conversations are created through the decomposition of complex multihop questions into simple, realistic multiturn dialogue interactions. We propose retrieval, system state tracking, and dialogue response generation tasks for our dataset and conduct baseline experiments for each. Our results show that there is still ample opportunity for improvement, demonstrating the importance of building stronger dialogue systems that can reason over the complex setting of informationseeking dialogue grounded on tables and text.

15:15-16:15 (Forum)
#32 Probing the Robustness of Trained Metrics for Conversational Dialogue Systems
Jan Milan Deriu, Don Tuggener, Pius Von Däniken and Mark Cieliebak
This paper introduces an adversarial method to stress-test trained metrics for the evaluation of conversational dialogue systems. The method leverages Reinforcement Learning to find response strategies that elicit optimal scores from the trained metrics. We apply our method to test recently proposed trained metrics. We find that they all are susceptible to giving high scores to responses generated by rather simple and obviously flawed strategies that our method converges on. For instance, simply copying parts of the conversation context to form a response yields competitive scores or even outperforms responses written by humans.

15:15-16:15 (Forum)
[DEMO] **Guided K-best Selection for Semantic Parsing Annotation**
Anton Belyy, Chieh-yang Huang, Jacob Andreas, Emmanouil Antonios Platanios, Sam Thomson, Richard Shin, Subhro Roy, Aleksandr Nisnevich, Charles Chen and Benjamin Van Durme
Collecting data for conversational semantic parsing is a time-consuming and demanding process. In this paper we consider, given an incomplete dataset with only a small amount of data, how to build an AI-powered human-in-the-loop process to enable efficient data collection. A guided K-best selection process is proposed, which (i) generates a set of possible valid candidates; (ii) allows users to quickly traverse the set and filter incorrect parses; and (iii) asks users to select the correct parse, with minimal modification when necessary. We investigate how to best support users in efficiently traversing the candidate set and locating the correct parse, in terms of speed and accuracy. In our user study, consisting of five annotators labeling 300 instances each, we find that combining keyword searching, where keywords can be used to query

relevant candidates, and keyword suggestion, where representative keywords are automatically generated, enables fast and accurate annotation.

Poster Session 5: Summarization

15:15-16:15 (Forum)

15:15-16:15 (Forum)
#33 ASPECTNEWS: Aspect-Oriented Summarization of News Documents
Ojas Ahuja, Jiacheng Xu, Akshay Kumar Gupta, Kevin Horecka and Greg Durrett
Generic summaries try to cover an entire document and query-based summaries try to answer document-specific questions. But real users' needs often fall in between these extremes and correspond to aspects, high-level topics discussed among similar types of documents. In this paper, we collect a dataset of realistic aspect-oriented summaries, AspectNews, which covers different subtopics about articles in news sub-domains. We annotate data across two domains of articles, earthquakes and fraud investigations, where each article is annotated with two distinct summaries focusing on different aspects for each domain. A system producing a single generic summary cannot concisely satisfy both aspects. Our focus in evaluation is how well existing techniques can generalize to these domains without seeing in-domain training data, so we turn to techniques to construct synthetic training data that have been used in query-focused summarization work. We compare several training schemes that differ in how strongly keywords are used and how oracle summaries are extracted. Our evaluation shows that our final approach yields (a) focused summaries, better than those from a generic summarization system or from keyword matching; (b) a system sensitive to the choice of keywords.

15:15-16:15 (Forum)
#34 A Well-Composed Text is Half Done! Composition Sampling for Diverse Conditional Generation
Shashi Narayan, Gonçalo Simões, Yao Zhao, Joshua Maynez, Dipanjan Das, Michael Collins and Mirella Lapata
We propose Composition Sampling, a simple but effective method to generate diverse outputs for conditional generation of higher quality compared to previous stochastic decoding strategies. It builds on recently proposed plan-based neural generation models (FROST, Narayan et al, 2021) that are trained to first create a composition of the output and then generate by conditioning on it and the input. Our approach avoids text degeneration by first sampling a composition in the form of an entity chain and then using beam search to generate the best possible text grounded to this entity chain. Experiments on summarization (CNN/DailyMail and XSum) and question generation (SQuAD), using existing and newly proposed automatic metrics together with human-based evaluation, demonstrate that Composition Sampling is currently the best available decoding strategy for generating diverse meaningful outputs.

15:15-16:15 (Forum)
#35 Training Dynamics for Text Summarization Models
Tanya Goyal, Jiacheng Xu, Junyi Jessy Li and Greg Durrett
Pre-trained language models (e.g. BART) have shown impressive results when fine-tuned on large summarization datasets. However, little is understood about this fine-tuning process, including what knowledge is retained from pre-training time or how content selection and generation strategies are learnt across iterations. In this work, we analyze the training dynamics for generation models, focusing on summarization. Across different datasets (CNN/DM, XSum, MediaSum) and summary properties, such as abstractiveness and hallucination, we study what the model learns at different stages of its fine-tuning process. We find that a propensity to copy the input is learned early in the training process consistently across all datasets studied. On the other hand, factual errors, such as hallucination of unsupported facts, are learnt in the later stages, though this behavior is more varied across domains. Based on these observations, we explore complementary approaches for modifying training: first, disregarding high-loss tokens that are challenging to learn and second, disregarding low-loss tokens that are learnt very quickly in the latter stages of the training process. We show that these simple training modifications allow us to configure our model to achieve different goals, such as improving factuality or improving abstractiveness.

15:15-16:15 (Forum)
[TACL] **#36 Planning with Learned Entity Prompts for Abstractive Summarization**
Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Ryan McDonald and Vitaly Nikolaev

15:15-16:15 (Forum)
[TACL] **#37 Document Summarization with Latent Queries**
Yumo Xu and Mirella Lapata

15:15-16:15 (Forum)
#38 Efficient Unsupervised Sentence Compression by Fine-tuning Transformers with Reinforcement Learning
Demian Gholipour Ghalandari, Chris Hokamp and Georgiana Ifrim
Sentence compression reduces the length of text by removing non-essential content while preserving important facts and grammaticality. Unsupervised objective driven methods for sentence compression can be used to create customized models without the need for ground-truth training data, while allowing flexibility in the objective function(s) that are used for learning and inference. Recent unsupervised sentence compression approaches use custom objectives to guide discrete search; however, guided search is expensive at inference time. In this work, we explore the use of reinforcement learning to train effective sentence compression models that are also fast when generating predictions. In particular, we cast the task as binary sequence labelling and fine-tune a pre-trained transformer using a simple policy gradient approach. Our approach outperforms other unsupervised models while also being more efficient at inference time.

15:15-16:15 (Forum)
#39 EntSUM: A Data Set for Entity-Centric Extractive Summarization
Mounica Maddela, Mayank Kulkarni and Daniel Preotiuc-Pietro
Controllable summarization aims to provide summaries that take into account user-specified aspects and preferences to better assist them with their information need, as opposed to the standard summarization setup which build a single generic summary of a document. We introduce a human-annotated data set EntSUM for controllable summarization with a focus on named entities as the aspects to control. We

conduct an extensive quantitative analysis to motivate the task of entity-centric summarization and show that existing methods for controllable summarization fail to generate entity-centric summaries. We propose extensions to state-of-the-art summarization approaches that achieve substantially better results on our data set. Our analysis and results show the challenging nature of this task and of the proposed data set.

15:15-16:15 (Forum)
#40 A Multi-Document Coverage Reward for RELAXed Multi-Document Summarization
Jacob Parnell, Inigo Jauregi Unanue and Massimo Piccardi
Multi-document summarization (MDS) has made significant progress in recent years, in part facilitated by the availability of new, dedicated datasets and capacious language models. However, a standing limitation of these models is that they are trained against limited references and with plain maximum-likelihood objectives. As for many other generative tasks, reinforcement learning (RL) offers the potential to improve the training of MDS models; yet, it requires a carefully-designed reward that can ensure appropriate leverage of both the reference summaries and the input documents. For this reason, in this paper we propose fine-tuning an MDS baseline with a reward that balances a reference-based metric such as ROUGE with coverage of the input documents. To implement the approach, we utilize RELAX (Grathwohl et al., 2018), a contemporary gradient estimator which is both low-variance and unbiased, and we fine-tune the baseline in a few-shot style for both stability and computational efficiency. Experimental results over the Multi-News and WCEP MDS datasets show significant improvements of up to +0.95 pp average ROUGE score and +3.17 pp METEOR score over the baseline, and competitive results with the literature. In addition, they show that the coverage of the input documents is increased, and evenly across all documents.

15:15-16:15 (Forum)
#41 Learning Non-Autoregressive Models from Search for Unsupervised Sentence Summarization
Puyuan Liu, Chenyang Huang and Lili Mou
Text summarization aims to generate a short summary for an input text. In this work, we propose a Non-Autoregressive Unsupervised Summarization (NAUS) approach, which does not require parallel data for training. Our NAUS first performs edit-based search towards a heuristically defined score, and generates a summary as pseudo-groundtruth. Then, we train an encoder-only non-autoregressive Transformer based on the search result. We also propose a dynamic programming approach for length-control decoding, which is important for the summarization task. Experiments on two datasets show that NAUS achieves state-of-the-art performance for unsupervised summarization, yet largely improving inference efficiency. Further, our algorithm is able to perform explicit length-transfer summary generation.

15:15-16:15 (Forum)
#42 SummaReranker: A Multi-Task Mixture-of-Experts Re-ranking Framework for Abstractive Summarization
Mathieu Ravaut, Shafiq Joty and Nancy F. Chen
Sequence-to-sequence neural networks have recently achieved great success in abstractive summarization, especially through fine-tuning large pre-trained language models on the downstream dataset. These models are typically decoded with beam search to generate a unique summary. However, the search space is very large, and with the exposure bias, such decoding is not optimal. In this paper, we show that it is possible to directly train a second-stage model performing re-ranking on a set of summary candidates. Our mixture-of-experts SummaReranker learns to select a better candidate and consistently improves the performance of the base model. With a base PEGASUS, we push ROUGE scores by 5.44

15:15-16:15 (Forum)
#43 NEWTS: A Corpus for News Topic-Focused Summarization
Seyed Ali Bahrainian, Sheridan Feucht and Carsten Eickhoff
Text summarization models are approaching human levels of fidelity. Existing benchmarking corpora provide concordant pairs of full and abridged versions of Web, news or professional content. To date, all summarization datasets operate under a one-size-fits-all paradigm that may not reflect the full range of organic summarization needs. Several recently proposed models (e.g., plug and play language models) have the capacity to condition the generated summaries on a desired range of themes. These capacities remain largely unused and unevaluated as there is no dedicated dataset that would support the task of topic-focused summarization.
This paper introduces the first topical summarization corpus NEWTS, based on the well-known CNN/Dailymail dataset, and annotated via online crowd-sourcing. Each source article is paired with two reference summaries, each focusing on a different theme of the source document. We evaluate a representative range of existing techniques and analyze the effectiveness of different prompting methods.

15:15-16:15 (Forum)
#44 Should We Trust This Summary? Bayesian Abstractive Summarization to The Rescue
Alexios Gidiotis and Grigorios Tsoumakas
We explore the notion of uncertainty in the context of modern abstractive summarization models, using the tools of Bayesian Deep Learning. Our approach approximates Bayesian inference by first extending state-of-the-art summarization models with Monte Carlo dropout and then using them to perform multiple stochastic forward passes. Based on Bayesian inference we are able to effectively quantify uncertainty at prediction time. Having a reliable uncertainty measure, we can improve the experience of the end user by filtering out generated summaries of high uncertainty. Furthermore, uncertainty estimation could be used as a criterion for selecting samples for annotation, and can be paired nicely with active learning and human-in-the-loop approaches. Finally, Bayesian inference enables us to find a Bayesian summary which performs better than a deterministic one and is more robust to uncertainty. In practice, we show that our Variational Bayesian equivalents of BART and PEGASUS can outperform their deterministic counterparts on multiple benchmark datasets.

15:15-16:15 (Forum)
#45 Revisiting Automatic Evaluation of Extractive Summarization Task: Can We Do Better than ROUGE?
Mousumi Akter, Naman Bansal and Shubhra Kanti Karmaker
It has been the norm for a long time to evaluate automated summarization tasks using the popular ROUGE metric. Although several studies in the past have highlighted the limitations of ROUGE, researchers have struggled to reach a consensus on a better alternative until today. One major limitation of the traditional ROUGE metric is the lack of semantic understanding (relies on direct overlap of n-grams). In this paper, we exclusively focus on the extractive summarization task and propose a semantic-aware nCG (normalized cumulative gain)-based evaluation metric (called Sem-nCG) for evaluating this task. One fundamental contribution of the paper is that it demonstrates how we can generate more reliable semantic-aware ground truths for evaluating extractive summarization tasks without any additional human intervention. To the best of our knowledge, this work is the first of its kind. We have conducted extensive experiments with this new metric using the widely used CNN/DailyMail dataset. Experimental results show that the new Sem-nCG metric is indeed semantic-aware, shows higher correlation with human judgement (more reliable) and yields a large number of disagreements with the original ROUGE metric (suggesting that ROUGE often

leads to inaccurate conclusions also verified by humans).

15:15-16:15 (Forum)
[TACL] **#46 SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization**
Philippe Laban, Tobias Schnabel, Paul Bennett and Marti Hearst

15:15-16:15 (Forum)
#47 BRIO: Bringing Order to Abstractive Summarization
Yixin Liu, Pengfei Liu, Dragomir Radev and Graham Neubig
Abstractive summarization models are commonly trained using maximum likelihood estimation, which assumes a deterministic (one-point) target distribution in which an ideal model will assign all the probability mass to the reference summary. This assumption may lead to performance degradation during inference, where the model needs to compare several system-generated (candidate) summaries that have deviated from the reference summary. To address this problem, we propose a novel training paradigm which assumes a non-deterministic distribution so that different candidate summaries are assigned probability mass according to their quality. Our method achieves a new state-of-the-art result on the CNN/DailyMail (47.78 ROUGE-1) and XSum (49.07 ROUGE-1) datasets. Further analysis also shows that our model can estimate probabilities of candidate summaries that are more correlated with their level of quality.

Poster Session 5: Generation

15:15-16:15 (Forum)

15:15-16:15 (Forum)
#48 Hierarchical Recurrent Aggregative Generation for Few-Shot NLG
Giulio Zhou, Gerasimos Lampouras and Ignacio Iacobacci
Large pretrained models enable transfer learning to low-resource domains for language generation tasks. However, previous end-to-end approaches do not account for the fact that some generation sub-tasks, specifically aggregation and lexicalisation, can benefit from transfer learning in different extents. To exploit these varying potentials for transfer learning, we propose a new hierarchical approach for few-shot and zero-shot generation. Our approach consists of a three-moduled jointly trained architecture: the first module independently lexicalises the distinct units of information in the input as sentence sub-units (e.g. phrases), the second module recurrently aggregates these sub-units to generate a unified intermediate output, while the third module subsequently post-edits it to generate a coherent and fluent final text. We perform extensive empirical analysis and ablation studies on few-shot and zero-shot settings across 4 datasets. Automatic and human evaluation shows that the proposed hierarchical approach is consistently capable of achieving state-of-the-art results when compared to previous work.

15:15-16:15 (Forum)
#49 GRS: Combining Generation and Revision in Unsupervised Sentence Simplification
Mohammad Dehghan, Dhruv Kumar and Lukasz Golab
We propose GRS: an unsupervised approach to sentence simplification that combines text generation and text revision. We start with an iterative framework in which an input sentence is revised using explicit edit operations, and add paraphrasing as a new edit operation. This allows us to combine the advantages of generative and revision-based approaches: paraphrasing captures complex edit operations, and the use of explicit edit operations in an iterative manner provides controllability and interpretability. We demonstrate these advantages of GRS compared to existing methods on the Newsela and ASSET datasets.

15:15-16:15 (Forum)
#50 Fine-Grained Controllable Text Generation Using Non-Residual Prompting
Fredrik Carlsson, Joey Öhman, Fangyu Liu, Severine Verlinden, Joakim Nivre and Magnus Sahlgren
The introduction of immensely large Causal Language Models (CLMs) has rejuvenated the interest in open-ended text generation. However, controlling the generative process for these Transformer-based models is at large an unsolved problem. Earlier work has explored either plug-and-play decoding strategies, or more powerful but blunt approaches such as prompting. There hence currently exists a trade-off between fine-grained control, and the capability for more expressive high-level instructions. To alleviate this trade-off, we propose an encoder-decoder architecture that enables intermediate text prompts at arbitrary time steps. We propose a resource-efficient method for converting a pre-trained CLM into this architecture, and demonstrate its potential on various experiments, including the novel task of contextualized word inclusion. Our method provides strong results on multiple experimental settings, proving itself to be both expressive and versatile.

15:15-16:15 (Forum)
#51 Hierarchical Sketch Induction for Paraphrase Generation
Tom Hosking, Hao Tang and Mirella Lapata
We propose a generative model of paraphrase generation, that encourages syntactic diversity by conditioning on an explicit syntactic sketch. We introduce Hierarchical Refinement Quantized Variational Autoencoders (HRQ-VAE), a method for learning decompositions of dense encodings as a sequence of discrete latent variables that make iterative refinements of increasing granularity. This hierarchy of codes is learned through end-to-end training, and represents fine-to-coarse grained information about the input. We use HRQ-VAE to encode the syntactic form of an input sentence as a path through the hierarchy, allowing us to more easily predict syntactic sketches at test time. Extensive experiments, including a human evaluation, confirm that HRQ-VAE learns a hierarchical representation of the input space, and generates paraphrases of higher quality than previous systems.

15:15-16:15 (Forum)
#52 Rewarding Semantic Similarity under Optimized Alignments for AMR-to-Text Generation
Lisa Jin and Daniel Gildea
A common way to combat exposure bias is by applying scores from evaluation metrics as rewards in reinforcement learning (RL). Metrics leveraging contextualized embeddings appear more flexible than their n-gram matching counterparts and thus ideal as training rewards. However, metrics such as BERTScore greedily align candidate and reference tokens, which can allow system outputs to receive excess credit

relative to a reference. Furthermore, past approaches featuring semantic similarity rewards suffer from repetitive outputs and overfitting. We address these issues by proposing metrics that replace the greedy alignments in BERTScore with optimized ones. We compute them on a model's trained token embeddings to prevent domain mismatch. Our model optimizing discrete alignment metrics consistently outperforms cross-entropy and BLEU reward baselines on AMR-to-text generation. In addition, we find that this approach enjoys stable training compared to a non-RL setting.

15:15-16:15 (Forum)
#53 uFACT: Unfaithful Alien-Corpora Training for Semantically Consistent Data-to-Text Generation
Tisha Anders, Alexandru Coca and Bill Byrne
We propose uFACT (Un-Faithful Alien Corpora Training), a training corpus construction method for data-to-text (d2t) generation models. We show that d2t models trained on uFACT datasets generate utterances which represent the semantic content of the data sources more accurately compared to models trained on the target corpus alone. Our approach is to augment the training set of a given target corpus with alien corpora which have different semantic representations. We show that while it is important to have faithful data from the target corpus, the faithfulness of additional corpora only plays a minor role. Consequently, uFACT datasets can be constructed with large quantities of unfaithful data. We show how uFACT can be leveraged to obtain state-of-the-art results on the WebNLG benchmark using METEOR as our performance metric. Furthermore, we investigate the sensitivity of the generation faithfulness to the training corpus structure using the PARENT metric, and provide a baseline for this metric on the WebNLG (Gardent et al., 2017) benchmark to facilitate comparisons with future work.

15:15-16:15 (Forum)
#54 High probability or low information? The probability–quality paradox in language generation
Clara Isabel Meister, Gian Wiher, Tiago Pimentel and Ryan D Cotterell
When generating natural language from neural probabilistic models, high probability does not always coincide with high quality. Rather, mode-seeking decoding methods can lead to incredibly unnatural language, while stochastic methods produce text perceived as much more human-like. In this note, we offer an explanation for this phenomenon by analyzing language as a means of communication in the information-theoretic sense. We posit that human-like language usually contains an expected amount of information—quantified as negative log-probability—and that language with substantially more (or less) information is undesirable. We provide preliminary empirical evidence for this hypothesis using quality ratings for both human and machine-generated text, covering multiple tasks and common decoding schemes.

15:15-16:15 (Forum)
#55 Quality Controlled Paraphrase Generation
Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shnayderman, Noam Slonim and Liat Ein-Dor
Paraphrase generation has been widely used in various downstream tasks. Most tasks benefit mainly from high quality paraphrases, namely those that are semantically similar to, yet linguistically diverse from, the original sentence. Generating high-quality paraphrases is challenging as it becomes increasingly hard to preserve meaning as linguistic diversity increases. Recent works achieve nice results by controlling specific aspects of the paraphrase, such as its syntactic tree. However, they do not allow to directly control the quality of the generated paraphrase, and suffer from low flexibility and scalability. Here we propose QCPG, a quality-guided controlled paraphrase generation model, that allows directly controlling the quality dimensions. Furthermore, we suggest a method that given a sentence, identifies points in the quality control space that are expected to yield optimal generated paraphrases. We show that our method is able to generate paraphrases which maintain the original meaning while achieving higher diversity than the uncontrolled baseline. The models, the code, and the data can be found in https://github.com/IBM/quality-controlled-paraphrase-generation.

15:15-16:15 (Forum)
#56 Tailor: Generating and Perturbing Text with Semantic Controls
Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E Peters and Matt Gardner
Controlled text perturbation is useful for evaluating and improving model generalizability. However, current techniques rely on training a model for every target perturbation, which is expensive and hard to generalize. We present Tailor, a semantically-controlled text generation system. Tailor builds on a pretrained seq2seq model and produces textual outputs conditioned on control codes derived from semantic representations. We craft a set of operations to modify the control codes, which in turn steer generation towards targeted attributes. These operations can be further composed into higher-level ones, allowing for flexible perturbation strategies. We demonstrate the effectiveness of these perturbations in multiple applications. First, we use Tailor to automatically create high-quality contrast sets for four distinct natural language processing (NLP) tasks. These contrast sets contain fewer spurious artifacts and are complementary to manually annotated ones in their lexical diversity. Second, we show that Tailor perturbations can improve model generalization through data augmentation. Perturbing just $\sim$2% of training data leads to a 5.8-point gain on an NLI challenge set measuring reliance on syntactic heuristics.

15:15-16:15 (Forum)
#57 Improving Compositional Generalization with Self-Training for Data-to-Text Generation
Sanket Vaibhav Mehta, Jinfeng Rao, Yi Tay, Mihir Kale, Ankur P Parikh and Emma Strubell
Data-to-text generation focuses on generating fluent natural language responses from structured meaning representations (MRs). Such representations are compositional and it is costly to collect responses for all possible combinations of atomic meaning schemata, thereby necessitating few-shot generalization to novel MRs. In this work, we systematically study the compositional generalization of the state-of-the-art T5 models in few-shot data-to-text tasks. We show that T5 models fail to generalize to unseen MRs, and we propose a template-based input representation that considerably improves the model's generalization capability. To further improve the model's performance, we propose an approach based on self-training using fine-tuned BLEURT for pseudo-response selection. On the commonly-used SGD and Weather benchmarks, the proposed self-training approach improves tree accuracy by $46\%+$ and reduces the slot error rates by $73\%+$ over the strong T5 baselines in few-shot settings.

15:15-16:15 (Forum)
#58 Evaluating Factuality in Text Simplification
Ashwin Devaraj, William Berkeley Sheffield, Byron C Wallace and Junyi Jessy Li
Automated simplification models aim to make input texts more readable. Such methods have the potential to make complex information accessible to a wider audience, e.g., providing access to recent medical literature which might otherwise be impenetrable for a lay reader. However, such models risk introducing errors into automatically simplified texts, for instance by inserting statements unsupported by the corresponding original text, or by omitting key information. Providing more readable but inaccurate versions of texts may in many cases be

worse than providing no such access at all. The problem of factual accuracy (and the lack thereof) has received heightened attention in the context of summarization models, but the factuality of automatically simplified texts has not been investigated. We introduce a taxonomy of errors that we use to analyze both references drawn from standard simplification datasets and state-of-the-art model outputs. We find that errors often appear in both that are not captured by existing evaluation metrics, motivating a need for research into ensuring the factual accuracy of automated simplification models.

15:15-16:15 (Forum)
#59 An Imitation Learning Curriculum for Text Editing with Non-Autoregressive Models
Sweta Agrawal and Marine Carpuat
We propose a framework for training non-autoregressive sequence-to-sequence models for editing tasks, where the original input sequence is iteratively edited to produce the output. We show that the imitation learning algorithms designed to train such models for machine translation introduces mismatches between training and inference that lead to undertraining and poor generalization in editing scenarios. We address this issue with two complementary strategies: 1) a roll-in policy that exposes the model to intermediate training sequences that it is more likely to encounter during inference, 2) a curriculum that presents easy-to-learn edit operations first, gradually increasing the difficulty of training samples as the model becomes competent. We show the efficacy of these strategies on two challenging English editing tasks: controllable text simplification and abstractive summarization. Our approach significantly improves output quality on both tasks and controls output complexity better on the simplification task.

15:15-16:15 (Forum)
#60 Updated Headline Generation: Creating Updated Summaries for Evolving News Stories
Sheena Panthaplackel, Adrian Benton and Mark Dredze
We propose the task of updated headline generation, in which a system generates a headline for an updated article, considering both the previous article and headline. The system must identify the novel information in the article update, and modify the existing headline accordingly. We create data for this task using the NewsEdits corpus by automatically identifying contiguous article versions that are likely to require a substantive headline update. We find that models conditioned on the prior headline and body revisions produce headlines judged by humans to be as factual as gold headlines while making fewer unnecessary edits compared to a standard headline generation model. Our experiments establish benchmarks for this new contextual summarization task.

15:15-16:15 (Forum)
#61 Neural Pipeline for Zero-Shot Data-to-Text Generation
Zdeněk Kasner and Ondrej Dusek
In data-to-text (D2T) generation, training on in-domain data leads to overfitting to the data representation and repeating training data noise. We examine how to avoid finetuning pretrained language models (PLMs) on D2T generation datasets while still taking advantage of surface realization capabilities of PLMs. Inspired by pipeline approaches, we propose to generate text by transforming single-item descriptions with a sequence of modules trained on general-domain text-based operations: ordering, aggregation, and paragraph compression. We train PLMs for performing these operations on a synthetic corpus WikiFluent which we build from English Wikipedia. Our experiments on two major triple-to-text datasets—WebNLG and E2E—show that our approach enables D2T generation from RDF triples in zero-shot settings.

15:15-16:15 (Forum)
#62 Few-shot Controllable Style Transfer for Low-Resource Multilingual Settings
Kalpesh Krishna, Deepak Nathani, Xavier Garcia, Bidisha Samanta and Partha Talukdar
Style transfer is the task of rewriting a sentence into a target style while approximately preserving content. While most prior literature assumes access to a large style-labelled corpus, recent work (Riley et al. 2021) has attempted "few-shot" style transfer using only 3-10 sentences at inference for style extraction. In this work we study a relevant low-resource setting: style transfer for languages where no style-labelled corpora are available. We notice that existing few-shot methods perform this task poorly, often copying inputs verbatim. We push the state-of-the-art for few-shot style transfer with a new method modeling the stylistic difference between paraphrases. When compared to prior work, our model achieves 2-3x better performance in formality transfer and code-mixing addition across seven languages. Moreover, our method is better at controlling the style transfer magnitude using an input scalar knob. We report promising qualitative results for several attribute transfer tasks (sentiment transfer, simplification, gender neutralization, text anonymization) all without retraining the model. Finally, we find model evaluation to be difficult due to the lack of datasets and metrics for many languages. To facilitate future research we crowdsource formality annotations for 4000 sentence pairs in four Indic languages, and use this data to design our automatic evaluations.

15:15-16:15 (Forum)
#63 Multilingual Pre-training with Language and Task Adaptation for Multilingual Text Style Transfer
Huiyuan Lai, Antonio Toral and Malvina Nissim
We exploit the pre-trained seq2seq model mBART for multilingual text style transfer. Using machine translated data as well as gold aligned English sentences yields state-of-the-art results in the three target languages we consider. Besides, in view of the general scarcity of parallel data, we propose a modular approach for multilingual formality transfer, which consists of two training strategies that target adaptation to both language and task. Our approach achieves competitive performance without monolingual task-specific parallel data and can be applied to other style transfer tasks as well as to other languages.

15:15-16:15 (Forum)
#64 Generating Scientific Claims for Zero-Shot Scientific Fact Checking
Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein and Lucy Lu Wang
Automated scientific fact checking is difficult due to the complexity of scientific language and a lack of significant amounts of training data, as annotation requires domain expertise. To address this challenge, we propose scientific claim generation, the task of generating one or more atomic and verifiable claims from scientific sentences, and demonstrate its usefulness in zero-shot fact checking for biomedical claims. We propose CLAIMGEN-BART, a new supervised method for generating claims supported by the literature, as well as KBIN, a novel method for generating claim negations. Additionally, we adapt an existing unsupervised entity-centric method of claim generation to biomedical claims, which we call CLAIMGEN-ENTITY. Experiments on zero-shot fact checking demonstrate that both CLAIMGEN-ENTITY and CLAIMGEN-BART, coupled with KBIN, achieve up to 90

15:15-16:15 (Forum)
#65 A Feasibility Study of Answer-Unaware Question Generation for Education
Liam Dugan, Eleni Miltsakaki, Shriyash Kaustubh Upadhyay, Etan Jacob Ginsberg, Hannah Gonzalez, DaHyeon Choi, Chuning Yuan and Chris Callison-Burch
We conduct a feasibility study into the applicability of answer-unaware question generation models to textbook passages. We show that a significant portion of errors in such systems arise from asking irrelevant or un-interpretable questions and that such errors can be ameliorated by providing summarized input. We find that giving these models human-written summaries instead of the original text results in a significant increase in acceptability of generated questions (33

15:15-16:15 (Forum)
#66 Probing Factually Grounded Content Transfer with Factual Ablation
Peter West, Chris Quirk, Michel Galley and Yejin Choi
Despite recent success, large neural models often generate factually incorrect text. Compounding this is the lack of a standard automatic evaluation for factuality–it cannot be meaningfully improved if it cannot be measured. Grounded generation promises a path to solving both of these problems: models draw on a reliable external document (grounding) for factual information, simplifying the challenge of factuality. Measuring factuality is also simplified–to factual consistency, testing whether the generation agrees with the grounding, rather than all facts. Yet, without a standard automatic metric for factual consistency, factually grounded generation remains an open problem.
We study this problem for content transfer, in which generations extend a prompt, using information from factual grounding. Particularly, this domain allows us to introduce the notion of factual ablation for automatically measuring factual consistency: this captures the intuition that the model should be less likely to produce an output given a less relevant grounding document. In practice, we measure this by presenting a model with two grounding documents, and the model should prefer to use the more factually relevant one. We contribute two evaluation sets to measure this. Applying our new evaluation, we propose multiple novel methods improving over strong baselines.

15:15-16:15 (Forum)
#67 Hybrid Semantics for Goal-Directed Natural Language Generation
Connor Baumler and Soumya Ray
We consider the problem of generating natural language given a communicative goal and a world description. We ask the question: is it possible to combine complementary meaning representations to scale a goal-directed NLG system without losing expressiveness? In particular, we consider using two meaning representations, one based on logical semantics and the other based on distributional semantics. We build upon an existing goal-directed generation system, S-STRUCT, which models sentence generation as planning in a Markov decision process. We develop a hybrid approach, which uses distributional semantics to quickly and imprecisely add the main elements of the sentence and then uses first-order logic based semantics to more slowly add the precise details. We find that our hybrid method allows S-STRUCT's generation to scale significantly better in early phases of generation and that the hybrid can often generate sentences with the same quality as S-STRUCT in substantially less time. However, we also observe and give insight into cases where the imprecision in distributional semantics leads to generation that is not as good as using pure logical semantics.

15:15-16:15 (Forum)
#68 A Recipe for Arbitrary Text Style Transfer with Large Language Models
Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch and Jason Wei
In this paper, we leverage large language models (LLMs) to perform zero-shot text style transfer. We present a prompting method that we call augmented zero-shot learning, which frames style transfer as a sentence rewriting task and requires only a natural language instruction, without model fine-tuning or exemplars in the target style. Augmented zero-shot learning is simple and demonstrates promising results not just on standard style transfer tasks such as sentiment, but also on arbitrary transformations such as 'make this melodramatic' or 'insert a metaphor.'

15:15-16:15 (Forum)
#69 Non-neural Models Matter: a Re-evaluation of Neural Referring Expression Generation Systems
Fahime Same, Guanyi Chen and Kees Van Deemter
In recent years, neural models have often outperformed rule-based and classic Machine Learning approaches in NLG. These classic approaches are now often disregarded, for example when new neural models are evaluated. We argue that they should not be overlooked, since, for some tasks, well-designed non-neural approaches achieve better performance than neural ones. In this paper, the task of generating referring expressions in linguistic context is used as an example. We examined two very different English datasets (WEBNLG and WSJ), and evaluated each algorithm using both automatic and human evaluations. Overall, the results of these evaluations suggest that rule-based systems with simple rule sets achieve on-par or better performance on both datasets compared to state-of-the-art neural REG systems. In the case of the more realistic dataset, WSJ, a machine learning-based system with well-designed linguistic features performed best. We hope that our work can encourage researchers to consider non-neural models in future.

15:15-16:15 (Forum)
#70 Cross-Task Generalization via Natural Language Crowdsourcing Instructions
Swaroop Mishra, Daniel Khashabi, Chitta Baral and Hannaneh Hajishirzi
Humans (e.g., crowdworkers) have a remarkable ability in solving different tasks, by simply reading textual instructions that define them and looking at a few examples. Despite the success of the conventional supervised learning on individual datasets, such models often struggle with generalization across tasks (e.g., a question-answering system cannot solve classification tasks). A long-standing challenge in AI is to build a model that learns a new task by understanding the human-readable instructions that define it. To study this, we introduce NATURAL INSTRUCTIONS, a dataset of 61 distinct tasks, their human-authored instructions, and 193k task instances (input-output pairs). The instructions are obtained from crowdsourcing instructions used to create existing NLP datasets and mapped to a unified schema. Using this meta-dataset, we measure cross-task generalization by training models on seen tasks and measuring generalization to the remaining unseen ones. We adopt generative pre-trained language models to encode task-specific instructions along with input and generate task output. Our results indicate that models benefit from instructions when evaluated in terms of generalization to unseen tasks (19

Coffee Break
16:15-16:45 - **Auditorium** (Forum)

Keynote 2: Fire-Side Chat with Barbara Grosz and Yejin Choi: "The

Trajectory of ACL and the Next 60 Years”
16:45-18:30 - **Auditorium** (Auditorium)

Virtual Poster Session 2 - 19:00-20:00

VPS2: Computational Social Science and Cultural Analytics

19:00-20:00 (GatherTown)

Findings: Dynamically Refined Regularization for Improving Cross-corpora Hate Speech Detection
Speaker: Tulika Bose

Findings: Human Language Modeling
Speaker: Nikita Soni

Long: Misinfo Reaction Frames: Reasoning about Readers' Reactions to News Headlines
Speaker: Saadia Gabriel

Findings: EnCBP: A New Benchmark Dataset for Finer-Grained Cultural Background Prediction in English
Speaker: Weicheng Ma

Findings: Good Night at 4 pm?! Time Expressions in Different Cultures
Speaker: Vered Shwartz

Short: A Risk-Averse Mechanism for Suicidality Assessment on Social Media
Speaker: Ramit Sawhney

Findings: From Stance to Concern: Adaptation of Propositional Analysis to New Tasks and Domains
Speaker: Brodie Mather

Findings: Modular Domain Adaptation
Speaker: Junshen K Chen

VPS2: Dialogue and Interactive Systems

19:00-20:00 (GatherTown)

Long: New Intent Discovery with Pre-training and Contrastive Learning
Speaker: Yuwei Zhang

Long: Multi-Party Empathetic Dialogue Generation: A New Task for Dialog Systems
Speaker: LingYu Zhu

Long: ProphetChat: Enhancing Dialogue Generation with Simulation of Future Conversation
Speaker: Chang Liu

Long: Where to Go for the Holidays: Towards Mixed-Type Dialogs for Clarification of User Goals
Speaker: Zeming Liu

Long: Think Before You Speak: Explicitly Generating Implicit Commonsense Knowledge for Response Generation
Speaker: Pei Zhou

Short: Learning-by-Narrating: Narrative Pre-Training for Zero-Shot Dialogue Comprehension
Speaker: Chao Zhao

Outstanding Paper: Online Semantic Parsing for Latency Reduction in Task-Oriented Dialogue
Speaker: Jiawei Zhou

Findings: C^3KG: A Chinese Commonsense Conversation Knowledge Graph
Speaker: Dawei Li

Long: DialFact: A Benchmark for Fact-Checking in Dialogue
Speaker: Prakhar Gupta

Findings: Long Time No See! Open-Domain Conversation with Long-Term Persona Memory
Speaker: Xinchao Xu

Long: An Interpretable Neuro-Symbolic Reasoning Framework for Task-Oriented Dialogue Generation
Speaker: Shiquan Yang

Long: SaFeRDialogues: Taking Feedback Gracefully after Conversational Safety Failures
Speaker: Megan Ung

Short: Can Visual Dialogue Models Do Scorekeeping? Exploring How Dialogue Representations Incrementally Encode Shared Knowledge
Speaker: Brielen Madureira

Findings: DARER: Dual-task Temporal Relational Recurrent Reasoning Network for Joint Dialog Sentiment Classification and Act Recognition
Speaker: Bowen Xing

Findings: Addressing Resource and Privacy Constraints in Semantic Parsing Through Data Augmentation
Speaker: Kevin Yang

Findings: Rethinking Offensive Text Detection as a Multi-Hop Reasoning Problem
Speaker: Qiang Zhang

Long: Lexical Knowledge Internalization for Neural Dialog Generation
Speaker: Zhiyong Wu

Long: A Model-agnostic Data Manipulation Method for Persona-based Dialogue Generation
Speaker: Yu Cao

Findings: A Few-Shot Semantic Parser for Wizard-of-Oz Dialogues with the Precise ThingTalk Representation
Speaker: Giovanni Campagna

Long: Internet-Augmented Dialogue Generation
Speaker: Mojtaba Komeili

TACL: Designing an Automatic Agent for Repeated Language based Persuasion Games
Speaker: Roi Reichart

SRW: Sketching a Linguistically-Driven Reasoning Dialog Model for Social Talk
Speaker: Alex Luu

VPS2: Discourse and Pragmatics & Ethics in NLP

19:00-20:00 (GatherTown)

Long: Constrained Multi-Task Learning for Bridging Resolution
Speaker: Hideo Kobayashi, Hideo Kobayashi

Long: Rethinking Self-Supervision Objectives for Generalizable Coherence Modeling
Speaker: Prathyusha Jwalapuram

VPS2: Ethics in NLP

19:00-20:00 (GatherTown)

Findings: Mitigating Gender Bias in Distilled Language Models via Counterfactual Role Reversal
Speaker: Umang Gupta

Findings: Learning Bias-reduced Word Embeddings Using Dictionary Definitions
Speaker: Haozhe An

Long: ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection
Speaker: Thomas Hartvigsen

Long: Improving Multi-label Malevolence Detection in Dialogues through Multi-faceted Label Correlation Enhancement
Speaker: Yangjun Zhang

Findings: Your fairness may vary: Pretrained language model fairness in toxic text classification
Speaker: Ioana Baldini

Findings: Assessing Multilingual Fairness in Pre-trained Multimodal Representations
Speaker: Jialu Wang

Short: On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations
Speaker: Yang Cao

Long: Reinforcement Guided Multi-Task Learning Framework for Low-Resource Stereotype Detection
Speaker: Rajkumar Pujari

Long: The Dangers of Underclaiming: Reasons for Caution When Reporting How NLP Systems Fail
Speaker: Samuel Bowman

SRW: Darkness can not drive out darkness: Investigating Bias in Hate SpeechDetection Models
Speaker: Fatma Elsafoury

VPS2: Generation

19:00-20:00 (GatherTown)

Long: Mix and Match: Learning-free Controllable Text Generationusing Energy Language Models
Speaker: Fatemehsadat Mireshghalla

Long: Explanation Graph Generation via Pre-trained Language Models: An Empirical Study with Contrastive Learning
Speaker: Swarnadeep Saha

Long: Spurious Correlations in Reference-Free Evaluation of Text Generation
Speaker: Esin Durmus

Findings: Diversifying Content Generation for Commonsense Reasoning with Mixture of Knowledge Graph Experts
Speaker: Wenhao Yu

Findings: CaM-Gen: Causally Aware Metric-Guided Text Generation
Speaker: Navita Goyal

Long: Neural Pipeline for Zero-Shot Data-to-Text Generation
Speaker: Zdenek Kasner

Long: Keywords and Instances: A Hierarchical Contrastive Learning Framework Unifying Hybrid Granularities for Text Generation
Speaker: Mingzhe Li

Findings: Controllable Natural Language Generation with Contrastive Prefixes
Speaker: Jing Qian

Findings: Using Pre-Trained Language Models for Producing Counter Narratives Against Hate Speech: a Comparative Study
Speaker: Serra Sinem Tekiroglu

Findings: Controlling the Focus of Pretrained Language Generation Models
Speaker: Jiabao Ji

Outstanding Paper: Evaluating Factuality in Text Simplification
Speaker: Ashwin Devaraj

Findings: Probing Factually Grounded Content Transfer with Factual Ablation
Speaker: Peter West

Long: An Imitation Learning Curriculum for Text Editing with Non-Autoregressive Models
Speaker: Sweta Agrawal

Long: How Do Seq2Seq Models Perform on End-to-End Data-to-Text Generation?
Speaker: Xunjian Yin

Long: $extitlatent$-GLAT: Glancing at Latent Variables for Parallel Text Generation
Speaker: Yu Bao

Long: Generating Biographies on Wikipedia: The Impact of Gender Bias on the Retrieval-Based Generation of Women Biographies
Speaker: Angela Fan

VPS2: Information Extraction

19:00-20:00 (GatherTown)

Long: Domain Adaptation in Multilingual and Multi-Domain Monolingual Settings for Complex Word Identification
Speaker: George-Eduard Zaharia

Findings: Query and Extract: Refining Event Extraction as Type-oriented Binary Decoding
Speaker: Sijia Wang

Long: An Unsupervised Multiple-Task and Multiple-Teacher Model for Cross-lingual Named Entity Recognition
Speaker: Zhuoran Li

Findings: Eider: Empowering Document-level Relation Extraction with Efficient Evidence Extraction and Inference-stage Fusion
Speaker: Yiqing Xie

Long: Multilingual Knowledge Graph Completion with Self-Supervised Adaptive Graph Alignment
Speaker: Zijie Huang

Short: Simple and Effective Knowledge-Driven Query Expansion for QA-Based Product Attribute Extraction
Speaker: Keiji Shinzato

Long: Learning from Sibling Mentions with Scalable Graph Inference in Fine-Grained Entity Typing
Speaker: Yi Chen

Long: Text-to-Table: A New Way of Information Extraction
Speaker: Xueqing Wu

Long: A Meta-framework for Spatiotemporal Quantity Extraction from Text
Speaker: Qiang Ning

Long: Continual Few-shot Relation Learning via Embedding Space Regularization and Data Augmentation
Speaker: Chengwei Qin

Findings: Leveraging Expert Guided Adversarial Augmentation For Improving Generalization in Named Entity Recognition
Speaker: Aaron Reich

Findings: Detection, Disambiguation, Re-ranking: Autoregressive Entity Linking as a Multi-Task Problem
Speaker: Khalil Mrini

Long: Automatic Error Analysis for Document-level Information Extraction
Speaker: Aliva Das

Findings: A Graph Enhanced BERT Model for Event Prediction
Speaker: LI DU

Long: Dynamic Global Memory for Document-level Argument Extraction
Speaker: Xinya Du

Long: MINER: Improving Out-of-Vocabulary Named Entity Recognition from an Information Theoretic Perspective
Speaker: Xiao Wang

Long: Pre-training and Fine-tuning Neural Topic Model: A Simple yet Effective Approach to Incorporating External Knowledge
Speaker: Linhai Zhang

Long: MarkupLM: Pre-training of Text and Markup Language for Visually Rich Document Understanding
Speaker: Junlong Li

Long: Does Recommend-Revise Produce Reliable Annotations? An Analysis on Missing Instances in DocRED
Speaker: Quzhe Huang

Long: Rethinking Negative Sampling for Handling Missing Entity Annotations
Speaker: Yangming Li

TACL: VILA: Improving Structured Content Extraction from Scientific PDFs Using Visual Layout Groups
Speaker: Zejiang Shen

SRW: What Do You Mean by Relation Extraction? A Survey on Datasets and Study on Scientific Relation Classification
Speaker: Elisa Bassignana, Elisa Bassignana

SRW: MEKER: Memory Efficient Knowledge Embedding Representation for Link Prediction and Question Answering
Speaker: Viktoriia Chekalina

VPS2: Information Retrieval and Text Mining

19:00-20:00 (GatherTown)

Findings: Domain Representative Keywords Selection: A Probabilistic Approach
Speaker: Pritom Saha Akash

Findings: Compressing Sentence Representation for Semantic Retrieval via Homomorphic Projective Distillation
Speaker: Xuandong Zhao

Long: An Effective and Efficient Entity Alignment Decoding Algorithm via Third-Order Tensor Isomorphism
Speaker: Xin Mao

Long: Hyperlink-induced Pre-training for Passage Retrieval in Open-domain Question Answering
Speaker: Jiawei Zhou

Findings: ED2LM: Encoder-Decoder to Language Model for Faster Document Re-ranking Inference
Speaker: Kai Hui

Findings: Zero-Shot Dense Retrieval with Momentum Adversarial Domain Invariant Representations
Speaker: Ji Xin

Findings: Zero-Shot Dense Retrieval with Momentum Adversarial Domain Invariant Representations
Speaker: Ji Xin

VPS2: Interpretability and Analysis of Models for NLP

19:00-20:00 (GatherTown)

Short: Are Shortest Rationales the Best Explanations for Human Understanding?
Speaker: Hua Shen

Findings: Extracting Latent Steering Vectors from Pretrained Language Models
Speaker: Nishant Subramani

Short: Counterfactual Explanations for Natural Language Interfaces
Speaker: George Tolkachev

Short: How does the pre-training objective affect what large language models learn about linguistic properties?
Speaker: Ahmed Alajrami

Long: Understanding Gender Bias in Knowledge Base Embeddings
Speaker: Yupei Du

Short: Data Contamination: From Memorization to Exploitation
Speaker: Inbal Magar

Findings: Coloring the Blank Slate: Pre-training Imparts a Hierarchical Inductive Bias to Sequence-to-sequence Models
Speaker: Aaron Mueller

Long: Robust Lottery Tickets for Pre-trained Language Models
Speaker: rui zheng

Long: ProtoTEx: Explaining Model Decisions with Prototype Tensors
Speaker: Anubrata Das

Findings: Combining Feature and Instance Attribution to Detect Artifacts
Speaker: Pouya Pezeshkpour

Short: Problems with Cosine as a Measure of Embedding Similarity for High Frequency Words
Speaker: Kaitlyn Zhou

Findings: Richer Countries and Richer Representations
Speaker: Kaitlyn Zhou

Long: Is Attention Explanation? An Introduction to the Debate
Speaker: Adrien Bibal

Findings: Training Text-to-Text Transformers with Privacy Guarantees
Speaker: Natalia Ponomareva

Findings: A Novel Perspective to Look At Attention: Bi-level Attention-based Explainable Topic Modeling for News Classification
Speaker: Dairui Liu

Findings: Systematicity, Compositionality and Transitivity of Deep NLP Models: a Metamorphic Testing Perspective
Speaker: Edoardo Manino

Long: Rewire-then-Probe: A Contrastive Recipe for Probing Biomedical Knowledge of Pre-trained Language Models
Speaker: Zaiqiao Meng

Findings: extitGeneralized but not Robust? Comparing the Effects of Data Modification Methods on Out-of-Domain Generalization and Adversarial Robustness
Speaker: Tejas Gokhale

Long: Low-Rank Softmax Can Have Unargmaxable Classes in Theory but Rarely in Practice
Speaker: Andreas Grivas

Short: Efficient Classification of Long Documents Using Transformers
Speaker: Hyunji Hayley Park

Long: Overcoming a Theoretical Limitation of Self-Attention
Speaker: David Chiang

Findings: On the data requirements of probing
Speaker: Zining Zhu

TACL: Word Acquisition in Neural Language Models
Speaker: Tyler Chang

VPS2: Language Groundings, Speech and Multimodality

19:00-20:00 (GatherTown)

Long: Improving Personalized Explanation Generation through Visualization
Speaker: Shijie Geng

Long: There's a Time and Place for Reasoning Beyond the Image
Speaker: Xingyu Fu

Long: Unified Speech-Text Pre-training for Speech Translation and Recognition
Speaker: Yun Tang

Findings: End-to-End Speech Translation for Code Switched Speech
Speaker: Orion Weller

Findings: Semantically Distributed Robust Optimization for Vision-and-Language Inference
Speaker: Tejas Gokhale

Long: Leveraging Unimodal Self-Supervised Learning for Multimodal Audio-Visual Speech Recognition
Speaker: Xichen Pan

Findings: CRAFT: A Benchmark for Causal Reasoning About Forces and inTeractions
Speaker: Tayfun Ates

Findings: Modeling Intensification for Sign Language Generation: A Computational Approach
Speaker: Mert Inan

Findings: Comprehensive Multi-Modal Interactions for Referring Image Segmentation
Speaker: Kanishk Jain

Long: Vision-and-Language Navigation: A Survey of Tasks, Methods, and Future Directions
Speaker: Jing Gu

Short: Understanding Game-Playing Agents with Natural Language Annotations
Speaker: Nicholas Tomlin

Long: Self-supervised Semantic-driven Phoneme Discovery for Zero-resource Speech Recognition
Speaker: Liming Wang

Long: Inferring Rewards from Language in Context
Speaker: Jessy Lin

SRW: On the Locality of Attention in Direct Speech Translation
Speaker: Belen Alastruey

VPS2: Linguistic Theories, Cognitive Modeling and Psycholinguistics

19:00-20:00 (GatherTown)

Short: Estimating the Entropy of Linguistic Distributions
Speaker: Aryaman Arora, Aryaman Arora

Long: GPT-D: Inducing Dementia-related Linguistic Anomalies by Deliberate Degradation of Artificial Neural Language Models
Speaker: Changye Li

Long: Metaphors in Pre-Trained Language Models: Probing and Generalization Across Datasets and Languages
Speaker: Ehsan Aghazadeh

Long: Characterizing Idioms: Conventionality and Contingency
Speaker: Michaela Socolof

Short: Developmental Negation Processing in Transformer Language Models
Speaker: Antonio Laverghetta Jr.

Long: Probing Simile Knowledge from Pre-trained Language Models
Speaker: Weijie Chen

Long: Transformers in the loop: Polarity in neural models of language
Speaker: Lisa Bylinina

Long: Neural reality of argument structure constructions
Speaker: Bai Li

Long: Flexible Generation from Fragmentary Linguistic Input
Speaker: Peng Qian

CL: Assessing corpus evidence for formal and psycholinguistic constraints on nonprojectivity
Speaker: Himanshu Yadav

TACL: Neuro-symbolic Natural Logic with Introspective Revision for Natural Language Inference
Speaker: Yufei Feng, Yufei Feng

VPS2: Machine Learning for NLP

19:00-20:00 (GatherTown)

Short: BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models
Speaker: Elad Ben Zaken

Long: Sparse Progressive Distillation: Resolving Overfitting under Pretrain-and-Finetune Paradigm
Speaker: Shaoyi Huang

Long: Meta-learning via Language Model In-context Tuning
Speaker: Yanda Chen

Long: RoCBert: Robust Chinese Bert with Multimodal Contrastive Pretraining
Speaker: Hui Su

Long: Multi-Granularity Structural Knowledge Distillation for Language Model Compression
Speaker: Chang Liu

Long: Better Language Model with Hypernym Class Prediction
Speaker: He Bai

Findings: Learning Adaptive Axis Attentions in Fine-tuning: Beyond Fixed Sparse Attention Patterns
Speaker: Zihan Wang

Findings: Distributed NLI: Learning to Predict Human Opinion Distributions for Language Reasoning
Speaker: Xiang Zhou

Short: Kronecker Decomposition for GPT Compression
Speaker: Ali Edalati

Findings: When Chosen Wisely, More Data Is What You Need: A Universal Sample-Efficient Strategy For Data Augmentation
Speaker: Ehsan Kamalloo

Long: Enhancing Chinese Pre-trained Language Model via Heterogeneous Linguistics Graph
Speaker: Yanzeng Li

Long: Knowledgeable Prompt-tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification
Speaker: Shengding Hu

Long: ConTinTin: Continual Learning from Task Instructions
Speaker: Wenpeng Yin

Long: Making Transformers Solve Compositional Tasks
Speaker: Santiago Ontanon

Short: SCD: Self-Contrastive Decorrelation of Sentence Embeddings
Speaker: Tassilo Klein

Long: Prompt-free and Efficient Few-shot Learning with Language Models
Speaker: Rabeeh Karimi Mahabadi

Long: Continual Sequence Generation with Adaptive Compositional Modules
Speaker: Yanzhe Zhang

Long: Token Dropping for Efficient BERT Pretraining
Speaker: Le Hou

Long: The Trade-offs of Domain Adaptation for Neural Language Models
Speaker: David Grangier

Short: Revisiting the Compositional Generalization Abilities of Neural Sequence Models
Speaker: Arkil Patel

Long: Bag-of-Words vs. Graph vs. Sequence in Text Classification: Questioning the Necessity of Text-Graphs and the Surprising Strength of a Wide MLP
Speaker: Lukas Galke

Long: Generative Pretraining for Paraphrase Evaluation
Speaker: Jack Weston

Long: Disentangled Sequence to Sequence Learning for Compositional Generalization
Speaker: Hao Zheng

Long: Noisy Channel Language Model Prompting for Few-Shot Text Classification
Speaker: Sewon Min

Long: On the Calibration of Pre-trained Language Models using Mixup Guided by Area Under the Margin and Saliency
Speaker: Seo Yeon Park, Seo Yeon Park

Long: infinity-former: Infinite Memory Transformer
Speaker: Pedro Henrique Martins

Long: UniPELT: A Unified Framework for Parameter-Efficient Language Model Tuning
Speaker: Yuning Mao

Findings: CUE Vectors: Modular Training of Language Models Conditioned on Diverse Contextual Signals
Speaker: Sreeparna Mukherjee

Findings: Aligned Weight Regularizers for Pruning Pretrained Neural Networks
Speaker: James O' Neill

Short: Contrastive Learning-Enhanced Nearest Neighbor Mechanism for Multi-Label Text Classification
Speaker: Xi'ao Su

Findings: EICO: Improving Few-Shot Text Classification via Explicit and Implicit Consistency Regularization
Speaker: Lei Zhao

Long: CAMERO: Consistency Regularized Ensemble of Perturbed Language Models with Weight Sharing
Speaker: Chen Liang

Short: NoisyTune: A Little Noise Can Help You Finetune Pretrained Language Models Better
Speaker: Chuhan Wu

Long: Sharpness-Aware Minimization Improves Language Model Generalization
Speaker: Dara Bahri

Long: Adapting Coreference Resolution Models through Active Learning
Speaker: Michelle Yuan

Short: Unsupervised multiple-choice question generation for out-of-domain Q&A fine-tuning
Speaker: Guillaume Le Berre

Long: Softmax Bottleneck Makes Language Models Unable to Represent Multi-mode Word Distributions
Speaker: Haw-Shiuan Chang

Long: A Comparison of Strategies for Source-Free Domain Adaptation
Speaker: Xin Su

Long: PPT: Pre-trained Prompt Tuning for Few-shot Learning
Speaker: Yuxian Gu

Short: On the Importance of Effectively Adapting Pretrained Language Models for Active Learning
Speaker: Katerina Margatina

Short: Text Smoothing: Enhance Various Data Augmentation Methods on Text Classification Tasks
Speaker: Xing Wu

Long: Pyramid-BERT: Reducing Complexity via Successive Core-set based Token Selection
Speaker: Xin Huang

TACL: Compressing Large-Scale Transformer-Based Models: A Case Study on BERT
Speaker: Prakhar Ganesh

TACL: PADA: Example-based Prompt Learning for on-the-fly Adaptation to Unseen Domains
Speaker: Roi Reichart

VPS2: Machine Translation and Multilinguality

19:00-20:00 (GatherTown)

Long: CipherDAug: Ciphertext based Data Augmentation for Neural Machine Translation
Speaker: Nishant Kambhatla

Long: Overlap-based Vocabulary Generation Improves Cross-lingual Transfer Among Related Languages
Speaker: Vaidehi Patil

Findings: Meta-X$_{NLG}$: A Meta-Learning Approach Based on Language Clustering for Zero-Shot Cross-Lingual Transfer and Generation
Speaker: Kaushal Kumar Maurya

Short: On Efficiently Acquiring Annotations for Multilingual Models
Speaker: Joel Ruben Antony Moniz

Findings: Automatic Song Translation for Tonal Languages
Speaker: Fenfei Guo

Long: Bias Mitigation in Machine Translation Quality Estimation
Speaker: Hanna Behnke

Long: DEEP: DEnoising Entity Pre-training for Neural Machine Translation
Speaker: Junjie Hu

Long: Efficient Cluster-Based k-Nearest-Neighbor Machine Translation
Speaker: Dexin Wang

Short: Translate-Train Embracing Translationese Artifacts
Speaker: Sicheng Yu

Long: Investigating Failures of Automatic Translation in the Case of Unambiguous Gender
Speaker: Adi Renduchintala

Findings: Combining Static and Contextualised Multilingual Embeddings
Speaker: Katharina Hämmerl

Long: Multilingual Mix: Example Interpolation Improves Multilingual Neural Machine Translation
Speaker: Yong Cheng

Long: Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder Translation Models
Speaker: Lorenzo Lupo

Long: Multilingual Generative Language Models for Zero-Shot Cross-Lingual Event Argument Extraction
Speaker: Kuan-Hao Huang

Long: Alternative Input Signals Ease Transfer in Multilingual Machine Translation
Speaker: Simeng Sun

Long: Multi Task Learning For Zero Shot Performance Prediction of Multilingual Models
Speaker: Kabir Ahuja

Findings: A Natural Diet: Towards Improving Naturalness of Machine Translation Output
Speaker: Markus Freitag

Long: From Simultaneous to Streaming Machine Translation by Leveraging Streaming History
Speaker: Javier Iranzo-Sánchez

Short: Focus on the Target's Vocabulary: Masked Label Smoothing for Machine Translation
Speaker: Liang Chen

Findings: Rethinking Document-level Neural Machine Translation
Speaker: Zewei Sun

Findings: First the Worst: Finding Better Gender Translations During Beam Search
Speaker: Danielle Saunders

Outstanding Paper: Learning to Generalize to More: Continuous Semantic Augmentation for Neural Machine Translation
Speaker: Xiangpeng Wei

Findings: CrossAligner & Co: Zero-Shot Transfer Methods for Task-Oriented Cross-lingual Natural Language Understanding
Speaker: Milan Gritta

SRW: Using Neural Machine Translation Methods for Sign Language Translation
Speaker: Galina Angelova

VPS2: NLP Applications

19:00-20:00 (GatherTown)

Long: Modeling U.S. State-Level Policies by Extracting Winners and Losers from Legislative Texts
Speaker: Maryam Davoodi

Long: Towards Comprehensive Patent Approval Predictions:Beyond Traditional Document Classification
Speaker: Xiaochen Gao

Long: TableFormer: Robust Transformer Modeling for Table-Text Encoding
Speaker: Jingfeng Yang

Long: It is AI's Turn to Ask Humans a Question: Question-Answer Pair Generation for Children's Story Books
Speaker: Dakuo Wang

Long: CAKE: A Scalable Commonsense-Aware Framework For Multi-View Knowledge Graph Completion
Speaker: Guanglin Niu

Long: Knowledge Enhanced Reflection Generation for Counseling Dialogues
Speaker: Siqi Shen

Long: Multilingual Molecular Representation Learning via Contrastive Pre-training
Speaker: Zhihui Guo

Long: From the Detection of Toxic Spans in Online Discussions to the Analysis of Toxic-to-Civil Transfer
Speaker: John Pavlopoulos

Findings: Question Generation for Reading Comprehension Assessment by Modeling How and What to Ask
Speaker: Bilal Ghanem

Long: Ensembling and Knowledge Distilling of Large Sequence Taggers for Grammatical Error Correction
Speaker: Maksym Tarnavskyi

Long: PromDA: Prompt-based Data Augmentation for Low-Resource NLU Tasks
Speaker: Yufei Wang

Long: FiNER: Financial Numeric Entity Recognition for XBRL Tagging
Speaker: Lefteris Loukas

Short: Canary Extraction in Natural Language Understanding Models
Speaker: Rahil Parikh

Long: Towards Learning (Dis)-Similarity of Source Code from Program Contrasts
Speaker: Yangruibo Ding

Long: Multilingual Detection of Personal Employment Status on Twitter
Speaker: Manuel Tonneau

Long: Leveraging Task Transferability to Meta-learning for Clinical Section Classification with Limited Data
Speaker: Zhuohao Chen

Findings: Incremental Intent Detection for Medical Domain with Contrast Replay Networks
Speaker: Guirong Bai, Guirong Bai

TACL: A Survey on Automated Fact-Checking
Speaker: Zhijiang Guo

SRW-164
Speaker: nan

SRW: Automatic Generation of Distractors for Fill-in-the-Blank Exercises with Round-Trip Neural Machine Translation
Speaker: Subhadarshi Panda

VPS2: Phonology, Morphology and Word Segmentation

19:00-20:00 (GatherTown)

Long: TopWORDS-Seg: Simultaneous Text Segmentation and Word Discovery for Open-Domain Chinese Texts via Bayesian Inference
Speaker: Changzai Pan

Short: Detecting Annotation Errors in Morphological Data with the Transformer
Speaker: Ling Liu, Ling Liu

Short: Morphological Reinflection with Multiple Arguments: An Extended Annotation schema and a Georgian Case Study
Speaker: David Guriel

Best Linguistic Insight: KinyaBERT: a Morphology-aware Kinyarwanda Language Model
Speaker: Antoine Nzeyimana, Antoine Nzeyimana

Long: CaMEL: Case Marker Extraction without Labels
Speaker: Leonie Weissweiler

Short: (Un)solving Morphological Inflection: Lemma Overlap Artificially Inflates Models' Performance
Speaker: Omer Goldman

VPS2: Question Answering

19:00-20:00 (GatherTown)

Long: Answer-level Calibration for Free-form Multiple Choice Question Answering
Speaker: Sawan Kumar

Findings: Question Answering Infused Pre-training of General-Purpose Contextualized Representations
Speaker: Robin Jia

Long: Synthetic Question Value Estimation for Domain Adaptation of Question Answering
Speaker: Xiang Yue

Findings: Using Interactive Feedback to Improve the Accuracy and Explainability of Question Answering Systems Post-Deployment
Speaker: Zichao Li

Findings: $extrmDuReader_extrmvis$: A Chinese Dataset for Open-domain Document Visual Question Answering
Speaker: Le Qi

Short: Leveraging Explicit Lexico-logical Alignments in Text-to-SQL Parsing
Speaker: Runxin Sun

Findings: Two-Step Question Retrieval for Open-Domain QA
Speaker: yeon seonwoo

Findings: Logic-Driven Context Extension and Data Augmentation for Logical Reasoning of Text
Speaker: Siyuan Wang

Findings: Hey AI, Can You Solve Complex Tasks by Talking to Agents?
Speaker: Tushar Khot

Short: C-MORE: Pretraining to Answer Open-Domain Questions by Consulting Millions of References
Speaker: Xiang Yue

Findings: Relevant CommonSense Subgraphs for "What if..." Procedural Reasoning
Speaker: Chen Zheng

Long: Simulating Bandit Learning from User Feedback for Extractive Question Answering
Speaker: Ge Gao

Long: Turning Tables: Generating Examples from Semi-structured Tables for Endowing Language Models with Reasoning Skills
Speaker: Ori Yoran

Long: Lite Unified Modeling for Discriminative Reading Comprehension
Speaker: Yilin Zhao

TACL: FeTaQA: Free-form Table Question Answering
Speaker: Linyong Nan

VPS2: Resources and Evaluation

19:00-20:00 (GatherTown)

Long: AlephBERT: Language Model Pre-training and Evaluation from Sub-Word to Sentence Level
Speaker: Amit Seker

Findings: Analyzing Dynamic Adversarial Training Data in the Limit
Speaker: Eric Wallace

Long: QuoteR: A Benchmark of Quote Recommendation for Writing
Speaker: Fanchao Qi

Long: e-CARE: a New Dataset for Exploring Explainable Causal Reasoning
Speaker: LI DU

Findings: Mukayese: Turkish NLP Strikes Back
Speaker: Ali Safaya

Long: FrugalScore: Learning Cheaper, Lighter and Faster Evaluation Metrics for Automatic Text Generation
Speaker: Moussa KAMAL EDDINE

Long: Down and Across: Introducing Crossword-Solving as a New NLP Benchmark
Speaker: Saurabh Kulshreshtha

Short: To Find Waldo You Need Contextual Cues: Debiasing Who's Waldo
Speaker: Yiran Luo

Long: TruthfulQA: Measuring How Models Mimic Human Falsehoods
Speaker: Stephanie Lin

Long: Understanding Iterative Revision from Human-Written Text
Speaker: Wanyu Du

Long: Detecting Unassimilated Borrowings in Spanish: An Annotated Corpus and Approaches to Modeling
Speaker: Elena Alvarez-Mellado

Long: Chart-to-Text: A Large-Scale Benchmark for Chart Summarization
Speaker: Shankar Kantharaj

Long: SRL4E – Semantic Role Labeling for Emotions: A Unified Evaluation Framework
Speaker: Cesare Campagnano

Short: Region-dependent temperature scaling for certainty calibration and application to class-imbalanced token classification
Speaker: Hillary Dawkins

Long: CLUES: A Benchmark for Learning Classifiers using Natural Language Explanations
Speaker: Rakesh Radhakrishnan Menon

Long: Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text
Speaker: Yao Dou

Long: CBLUE: A Chinese Biomedical Language Understanding Evaluation Benchmark
Speaker: Ningyu Zhang

Findings: E-KAR: A Benchmark for Rationalizing Natural Language Analogical Reasoning
Speaker: Jiangjie Chen

Short: CoDA21: Evaluating Language Understanding Capabilities of NLP Models With Context-Definition Alignment
Speaker: Lütfi Kerem Şenel

CL: Annotation Curricula to Implicitly Train Non-Expert Annotators
Speaker: Ji-Ung Lee

TACL: Decomposing and Recomposing Event Structure
Speaker: William Gantt

SRW: Evaluating zero-shot transfers and multilingual models for dependency parsing and POS tagging within the low-resource language family Tupían
Speaker: Frederic Blum

VPS2: Semantics

19:00-20:00 (GatherTown)

Long: LexSubCon: Integrating Knowledge from Lexical Resources into Contextual Embeddings for Lexical Substitution
Speaker: Georgios Michalopoulos

Long: WatClaimCheck: A new Dataset for Claim Entailment and Inference
Speaker: Kashif Khan

Long: On The Ingredients of an Effective Zero-shot Semantic Parser
Speaker: Pengcheng Yin

Long: Fully-Semantic Parsing and Generation: the BabelNet Meaning Representation
Speaker: Abelardo Carlos Martinez Lorenzo

Short: Hierarchical Curriculum Learning for AMR Parsing
Speaker: Peiyi Wang

Long: Variational Graph Autoencoding as Cheap Supervision for AMR Coreference Resolution
Speaker: Irene Li, Irene Li

Long: Probing for Predicate Argument Structures in Pretrained Language Models
Speaker: Simone Conia

Findings: Lacking the Embedding of a Word? Look it up into a Traditional Dictionary
Speaker: Elena Sofia Ruzzetti

Long: IMPLI: Investigating NLI Models' Performance on Figurative Language
Speaker: Kevin Stowe

Long: Bridging the Generalization Gap in Text-to-SQL Parsing with Schema Expansion
Speaker: Chen Zhao

Short: Sequence-to-sequence AMR Parsing with Ancestor Information
Speaker: Chen Yu

Long: SciNLI: A Corpus for Natural Language Inference on Scientific Text
Speaker: Mobashir Sadat

Short: An Analysis of Negation in Natural Language Understanding Corpora
Speaker: Md Mosharaf Hossain

Long: Few-Shot Learning with Siamese Networks and Label Tuning
Speaker: Thomas Müller

SRW: Exploring Cross-lingual Text Detoxification with Large Multilingual Language Models.
Speaker: Daniil Moskovskiy

SRW: Deep Neural Representations for Multiword Expressions Detection
Speaker: Kamil Kanclerz

SRW: Using dependency parsing for few-shot learning in distributional semantics
Speaker: Stefania Preda

VPS2: Sentiment Analysis, Stylistic Analysis, and Argument Mining

19:00-20:00 (GatherTown)

Findings: Efficient Argument Structure Extraction with Transfer Learning and Active Learning
Speaker: Xinyu Hua

Short: Pixie: Preference in Implicit and Explicit Comparisons
Speaker: Amanul Haque

Findings: Sentiment Word Aware Multimodal Refinement for Multimodal Sentiment Analysis with ASR Errors
Speaker: Yang Wu

Long: Adversarial Soft Prompt Tuning for Cross-Domain Sentiment Analysis
Speaker: Hui Wu

Long: Enhanced Multi-Channel Graph Convolutional Network for Aspect Sentiment Triplet Extraction
Speaker: Hao Chen

Long: Incorporating Stock Market Signals for Twitter Stance Detection
Speaker: Costanza Conforti

Short: Direct parsing to sentiment graphs
Speaker: David Samuel

Long: Identifying the Human Values behind Arguments
Speaker: Johannes Kiesel

CL: Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis
Speaker: Saif Mohammad

SRW: A Dataset and BERT-based Models for Targeted Sentiment Analysis on Turkish Texts
Speaker: Mustafa Melih Mutlu

VPS2: Special Theme on Language Diversity: From Low Resource to Endangered

19:00-20:00 (GatherTown)

Findings: Pre-Trained Multilingual Sequence-to-Sequence Models: A Hope for Low-Resource Language Translation?
Speaker: En-Shiun Lee

Long: Expanding Pretrained Models to Thousands More Languages via Lexicon-based Adaptation
Speaker: Xinyi Wang

Findings: Toward More Meaningful Resources for Lower-resourced Languages
Speaker: Constantine Lignos

Long: Computational Historical Linguistics and Language Diversity in South Asia
Speaker: Aryaman Arora, Aryaman Arora

Findings: Zero-shot Learning for Grapheme to Phoneme Conversion with Language Ensemble
Speaker: Xinjian Li

Long: Towards Afrocentric NLP for African Languages: Where We Are and Where We Can Go
Speaker: Ife Adebara

Long: Not always about you: Prioritizing community needs when developing endangered language technology
Speaker: Zoey Liu

Findings: Automatic Speech Recognition and Query By Example for Creole Languages Documentation
Speaker: Cécile Macaire

Long: Learning From Failure: Data Capture in an Australian Aboriginal Community
Speaker: Eric Le Ferrand, Eric Le Ferrand

Long: Multilingual unsupervised sequence segmentation transfers to extremely low-resource languages
Speaker: Agatha Downey

Long: Cree Corpus: A Collection of nêhiyawêwin Resources
Speaker: Daniela Teodorescu

Long: One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia
Speaker: Alham Aji

Long: Weakly Supervised Word Segmentation for Computational Language Documentation
Speaker: Shu Okabe

Short: Can a Transformer Pass the Wug Test? Tuning Copying Bias in Neural Morphological Inflection Models
Speaker: Ling Liu, Ling Liu

VPS2: Summarization

19:00-20:00 (GatherTown)

Long: HIBRIDS: Attention with Hierarchical Biases for Structure-aware Long Document Summarization
Speaker: Shuyang Cao

Findings: NEWTS: A Corpus for News Topic-Focused Summarization
Speaker: Seyed Ali Bahrainian

Long: Faithful or Extractive? On Mitigating the Faithfulness-Abstractiveness Trade-off in Abstractive Summarization
Speaker: Faisal Ladhak

Long: Summ^N: A Multi-Stage Summarization Framework for Long Input Dialogues and Documents
Speaker: Yusen Zhang

Long: BRIO: Bringing Order to Abstractive Summarization
Speaker: Yixin Liu

Long: Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization
Speaker: Meng Cao

Long: Towards Abstractive Grounded Summarization of Podcast Transcripts
Speaker: Kaiqiang Song

Long: A Multi-Document Coverage Reward for RELAXed Multi-Document Summarization
Speaker: Jacob Parnell

Findings: Comparative Opinion Summarization via Collaborative Decoding
Speaker: Hayate Iso

Long: SummScreen: A Dataset for Abstractive Screenplay Summarization
Speaker: Mingda Chen

VPS2: Syntax: Tagging, Chunking and Parsing

19:00-20:00 (GatherTown)

Long: Few-Shot Class-Incremental Learning for Named Entity Recognition
Speaker: Rui Wang

Findings: Co-training an Unsupervised Constituency Parser with Weak Supervision
Speaker: Nickil Maveli

Best Paper: Learned Incremental Representations for Parsing
Speaker: Nikita Kitaev

Findings: Revisiting the Effects of Leakage on Dependency Parsing
Speaker: Nathaniel Krasner

Long: Substructure Distribution Projection for Zero-Shot Cross-Lingual Dependency Parsing
Speaker: Freda Shi

Findings: Towards Few-shot Entity Recognition in Document Images: A Label-aware Sequence-to-Sequence Framework
Speaker: Zilong Wang

CL: Improved N-Best Extraction with an Evaluation on Language Data
Speaker: Johanna Björklund

Main Conference: Wednesday, May 25, 2022

Virtual Poster Session 3 - 07:30-08:30

VPS3: Computational Social Science and Cultural Analytics

07:30-08:30 (GatherTown)

Findings: Classification without (Proper) Representation: Political Heterogeneity in Social Media and Its Implications for Classification and Behavioral Analysis
Speaker: Kenan Alkiek

Long: Doctor Recommendation in Online Health Forums via Expertise Learning
Speaker: Xiaoxin Lu

Findings: Measuring the Language of Self-Disclosure across Corpora
Speaker: Ann-Katrin Reuel

Findings: Listening to Affected Communities to Define Extreme Speech: Dataset and Experiments
Speaker: Antonis Maronikolakis

Long: Misinfo Reaction Frames: Reasoning about Readers' Reactions to News Headlines
Speaker: Saadia Gabriel

Findings: Improving Neural Political Statement Classification with Class Hierarchical Information
Speaker: Erenay Dayanik

Long: Zoom Out and Observe: News Environment Perception for Fake News Detection
Speaker: Qiang Sheng

Findings: Suum Cuique: Studying Bias in Taboo Detection with a Community Perspective
Speaker: Osama Khalid

VPS3: Dialogue and Interactive Systems

07:30-08:30 (GatherTown)

Findings: Towards Transparent Interactive Semantic Parsing via Step-by-Step Correction
Speaker: Lingbo Mo

Findings: KSAM: Infusing Multi-Source Knowledge into Dialogue Generation via Knowledge Source Aware Multi-Head Decoding
Speaker: Sixing Wu

Long: Structural Characterization for Dialogue Disentanglement
Speaker: Xinbei Ma

Findings: Towards Large-Scale Interpretable Knowledge Graph Reasoning for Dialogue Systems
Speaker: Yi-Lin Tuan

Long: Perceiving the World: Question-guided Reinforcement Learning for Text-based Games
Speaker: Yunqiu Xu

Short: Disentangled Knowledge Transfer for OOD Intent Discovery with Unified Contrastive Learning
Speaker: Yutao Mou

Long: DEAM: Dialogue Coherence Evaluation using AMR-based Semantic Manipulations
Speaker: Sarik Ghazarian

Findings: Hierarchical Inductive Transfer for Continual Dialogue Learning
Speaker: Shaoxiong Feng

Long: Where to Go for the Holidays: Towards Mixed-Type Dialogs for Clarification of User Goals
Speaker: Zeming Liu

Long: Continual Prompt Tuning for Dialog State Tracking
Speaker: Qi Zhu

Findings: DS-TOD: Efficient Domain Specialization for Task-Oriented Dialog
Speaker: Chia-Chien Hung

Findings: Multi-Stage Prompting for Knowledgeable Dialogue Generation
Speaker: Zihan Liu

Long: Beyond the Granularity: Multi-Perspective Dialogue Collaborative Selection for Dialogue State Tracking
Speaker: Jinyu Guo

Findings: TegTok: Augmenting Text Generation via Task-specific and Open-world Knowledge
Speaker: Chao-Hong Tan

Findings: N-Shot Learning for Augmenting Task-Oriented Dialogue State Tracking
Speaker: Ibrahim Aksu

Long: GL-CLeF: A Global–Local Contrastive Learning Framework for Cross-lingual Spoken Language Understanding
Speaker: Libo Qin

Findings: VISITRON: Visual Semantics-Aligned Interactively Trained Object-Navigator
Speaker: Ayush Shrivastava

Findings: Data Augmentation and Learned Layer Aggregation for Improved Multilingual Language Understanding in Dialogue
Speaker: Evgeniia Razumovskaia

Long: Contextual Fine-to-Coarse Distillation for Coarse-grained Response Selection in Open-Domain Conversations
Speaker: Wei Chen

Findings: ASSIST: Towards Label Noise-Robust Dialogue State Tracking
Speaker: Fanghua Ye

Long: HeterMPC: A Heterogeneous Graph Neural Network for Response Generation in Multi-Party Conversations
Speaker: Jia-Chen Gu

Findings: Selecting Stickers in Open-Domain Dialogue through Multitask Learning
Speaker: Zhexin Zhang

Findings: One Agent To Rule Them All: Towards Multi-agent Conversational AI
Speaker: Christopher Clarke

Findings: Addressing Resource and Privacy Constraints in Semantic Parsing Through Data Augmentation
Speaker: Kevin Yang

Long: ChatMatch: Evaluating Chatbots by Autonomous Chat Tournaments
Speaker: Ruolan Yang

Short: Probing the Robustness of Trained Metrics for Conversational Dialogue Systems
Speaker: Jan Deriu

Findings: Dialogue Summaries as Dialogue States (DS2), Template-Guided Summarization for Few-shot Dialogue State Tracking
Speaker: Jamin Shin

VPS3: Discourse and Pragmatics & Ethics in NLP

07:30-08:30 (GatherTown)

Findings: The Change that Matters in Discourse Parsing: Estimating the Impact of Domain Shift on Parser Error
Speaker: Katherine Atwell

Findings: Graph Refinement for Coreference Resolution
Speaker: Lesly Miculicich, Lesly Miculicich

Long: Rethinking Self-Supervision Objectives for Generalizable Coherence Modeling
Speaker: Prathyusha Jwalapuram

Findings: Encoding and Fusing Semantic Connection and Linguistic Evidence for Implicit Discourse Relation Recognition
Speaker: Wei Xiang

Findings: What does it take to bake a cake? The RecipeRef corpus and anaphora resolution in procedural text
Speaker: Biaoyan Fang

VPS3: Ethics in NLP

07:30-08:30 (GatherTown)

Findings: Entropy-based Attention Regularization Frees Unintended Bias Mitigation from Lists
Speaker: Giuseppe Attanasio

Long: Sentence-level Privacy for Document Embeddings
Speaker: Casey Meehan

Long: Upstream Mitigation Is extitNot All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models
Speaker: Ryan Steed

Findings: Square One Bias in NLP: Towards a Multi-Dimensional Exploration of the Research Manifold
Speaker: Sebastian Ruder

Long: Reinforcement Guided Multi-Task Learning Framework for Low-Resource Stereotype Detection
Speaker: Rajkumar Pujari

Findings: Using NLP to quantify the environmental cost and diversity benefits of in-person NLP conferences
Speaker: Piotr Przybyła, Piotr Przybyła

VPS3: Generation

07:30-08:30 (GatherTown)

Findings: Compilable Neural Code Generation with Compiler Feedback
Speaker: Xin Wang

Long: Rare Tokens Degenerate All Tokens: Improving Neural Text Generation via Adaptive Gradient Gating for Rare Token Embeddings
Speaker: Sangwon Yu

Findings: Multi-Scale Distribution Deep Variational Autoencoder for Explanation Generation
Speaker: ZeFeng Cai

Findings: Controlled Text Generation Using Dictionary Prior in Variational Autoencoders
Speaker: Fang Xianghong

Findings: GRS: Combining Generation and Revision in Unsupervised Sentence Simplification
Speaker: Mohammad Dehghan

Findings: Multi-task Learning for Paraphrase Generation With Keyword and Part-of-Speech Reconstruction
Speaker: Xuhang Xie

Findings: Effective Unsupervised Constrained Text Generation based on Perturbed Masking
Speaker: Yingwen Fu

Long: PLANET: Dynamic Content Planning in Autoregressive Transformers for Long-form Text Generation
Speaker: Zhe Hu

Long: CTRLEval: An Unsupervised Reference-Free Metric for Evaluating Controlled Text Generation
Speaker: Pei Ke

Findings: A Feasibility Study of Answer-Unaware Question Generation for Education
Speaker: Liam Dugan

Findings: Hierarchical Recurrent Aggregative Generation for Few-Shot NLG
Speaker: Giulio Zhou

Findings: MReD: A Meta-Review Dataset for Structure-Controllable Text Generation
Speaker: Chenhui Shen

Long: Semi-Supervised Formality Style Transfer with Consistency Training
Speaker: Ao Liu

Findings: uFACT: Unfaithful Alien-Corpora Training for Semantically Consistent Data-to-Text Generation
Speaker: Tisha Anders

Findings: Synchronous Refinement for Neural Machine Translation
Speaker: Kehai Chen

Findings: Using Pre-Trained Language Models for Producing Counter Narratives Against Hate Speech: a Comparative Study
Speaker: Serra Sinem Tekiroglu

Long: Updated Headline Generation: Creating Updated Summaries for Evolving News Stories
Speaker: Sheena Panthaplackel

Findings: Event Transition Planning for Open-ended Text Generation
Speaker: Qintong Li

Long: CQG: A Simple and Effective Controlled Generation Framework for Multi-hop Question Generation
Speaker: zichu fei

Findings: Improving Controllable Text Generation with Position-Aware Weighted Decoding
Speaker: Yuxuan Gu

Long: Few-shot Controllable Style Transfer for Low-Resource Multilingual Settings
Speaker: Kalpesh Krishna

Long: Dependency-based Mixture Language Models
Speaker: Zhixian Yang

Findings: GCPG: A General Framework for Controllable Paraphrase Generation
Speaker: Kexin Yang

TACL: Relational Memory Augmented Language Models
Speaker: Qi Liu

VPS3: Information Extraction

07:30-08:30 (GatherTown)

Findings: RelationPrompt: Leveraging Prompts to Generate Synthetic Data for Zero-Shot Relation Triplet Extraction
Speaker: Yew Ken Chia

Findings: Extract-Select: A Span Selection Framework for Nested Named Entity Recognition with Generative Adversarial Training
Speaker: Peixin Huang, Peixin Huang

Findings: LEVEN: A Large-Scale Chinese Legal Event Detection Dataset
Speaker: Feng Yao

Findings: MDERank: A Masked Document Embedding Rank Approach for Unsupervised Keyphrase Extraction
Speaker: Linhan Zhang

Long: Nested Named Entity Recognition with Span-level Graphs
Speaker: Juncheng Wan

Long: Parallel Instance Query Network for Named Entity Recognition
Speaker: Yongliang Shen

Findings: A Simple yet Effective Relation Information Guided Approach for Few-Shot Relation Extraction
Speaker: Yang Liu

Findings: DeepStruct: Pre-Training of Language Models for Structure Prediction
Speaker: Chenguang Wang

Short: Event-Event Relation Extraction using Probabilistic Box Embedding
Speaker: EunJeong Hwang

Findings: Knowledge Graph Embedding by Adaptive Limit Scoring Loss Using Dynamic Weighting Strategy
Speaker: Jinfa Yang

Long: Cross-Lingual Contrastive Learning for Fine-Grained Entity Typing for Low-Resource Languages
Speaker: Xu Han

Short: Complex Evolutional Pattern Learning for Temporal Knowledge Graph Reasoning
Speaker: Zixuan Li

Long: Alignment-Augmented Consistent Translation for Multilingual Open Information Extraction
Speaker: Keshav Kolluru

Findings: Document-Level Relation Extraction with Adaptive Focal Loss and Knowledge Distillation
Speaker: Qingyu Tan

Short: PARE: A Simple and Strong Baseline for Monolingual and Multilingual Distantly Supervised Relation Extraction
Speaker: Vipul Rathore

Findings: Improving Relation Extraction through Syntax-induced Pre-training with Dependency Masking
Speaker: Yuanhe Tian, Yuanhe Tian

Long: Show Me More Details: Discovering Hierarchies of Procedures from Semi-structured Web Data
Speaker: Shuyan Zhou

Findings: Label Semantics for Few Shot Named Entity Recognition
Speaker: Jie Ma

Findings: Learn and Review: Enhancing Continual Named Entity Recognition via Reviewing Synthetic Samples
Speaker: Yu Xia

Findings: An Accurate Unsupervised Method for Joint Entity Alignment and Dangling Entity Detection
Speaker: Shengxuan Luo

Long: SimKGC: Simple Contrastive Knowledge Graph Completion with Pre-trained Language Models
Speaker: Liang Wang

Short: A Simple but Effective Pluggable Entity Lookup Table for Pre-trained Language Models
Speaker: Deming Ye

Long: Packed Levitated Marker for Entity and Relation Extraction
Speaker: Deming Ye

Long: Dynamic Prefix-Tuning for Generative Template-based Event Extraction
Speaker: Xiao Liu

Long: Unified Structure Generation for Universal Information Extraction
Speaker: Yaojie Lu

Findings: Fusing Heterogeneous Factors with Triaffine Mechanism for Nested Named Entity Recognition
Speaker: Zheng Yuan

Long: OIE@OIA: an Adaptable and Efficient Open Information Extraction Framework
Speaker: Xin Wang

Long: CONTaiNER: Few-Shot Named Entity Recognition via Contrastive Learning
Speaker: Sarkar Snigdha Sarathi Das

Findings: Consistent Representation Learning for Continual Relation Extraction
Speaker: Kang Zhao

Long: Boundary Smoothing for Named Entity Recognition
Speaker: Enwei Zhu

Findings: Do Pre-trained Models Benefit Knowledge Graph Completion? A Reliable Evaluation and a Reasonable Approach
Speaker: Xin Lv

Findings: Improving Candidate Retrieval with Entity Profile Generation for Wikidata Entity Linking
Speaker: Tuan Lai, Tuan Lai

SRW: Pretrained Knowledge Base Embeddings for improved Sentential Relation Extraction
Speaker: andrea papaluca

VPS3: Information Retrieval and Text Mining

07:30-08:30 (GatherTown)

Findings: MINER: Multi-Interest Matching Network for News Recommendation
Speaker: Jian Li

Long: Sentence-aware Contrastive Learning for Open-Domain Passage Retrieval
Speaker: Bohong Wu

Findings: TABi: Type-Aware Bi-Encoders for Open-Domain Entity Retrieval
Speaker: Megan Leszczynski

Findings: MTRec: Multi-Task Learning over BERT for News Recommendation
Speaker: Qiwei Bi

Long: Multi-View Document Representation Learning for Open-Domain Dense Retrieval
Speaker: Shunyu Zhang

Long: UCTopic: Unsupervised Contrastive Learning for Phrase Representations and Topic Mining
Speaker: Jiacheng Li

Findings: Two Birds with One Stone: Unified Model Learning for Both Recall and Ranking in News Recommendation
Speaker: Chuhan Wu

Findings: LaPraDoR: Unsupervised Pretrained Dense Retriever for Zero-Shot Text Retrieval
Speaker: Canwen Xu

VPS3: Interpretability and Analysis of Models for NLP

07:30-08:30 (GatherTown)

Findings: On the Importance of Data Size in Probing Fine-tuned Models
Speaker: Houman Mehrafarin

Short: Are Shortest Rationales the Best Explanations for Human Understanding?
Speaker: Hua Shen

Findings: Reframing Instructional Prompts to GPTk's Language
Speaker: Daniel Khashabi

Findings: MoEfication: Transformer Feed-forward Layers are Mixtures of Experts
Speaker: Zhengyan Zhang

Findings: Distinguishing Non-natural from Natural Adversarial Samples for More Robust Pre-trained Language Model
Speaker: Jiayi Wang

Short: Data Contamination: From Memorization to Exploitation
Speaker: Inbal Magar

Findings: Explaining Classes through Stable Word Attributions
Speaker: Samuel Rönnqvist

Findings: What to Learn, and How: Toward Effective Learning from Rationales
Speaker: Samuel Carton

Long: Probing Structured Pruning on Multilingual Pre-trained Models: Settings, Algorithms, and Efficiency
Speaker: Yanyang Li, Yanyang Li

Findings: Finding the Dominant Winning Ticket in Pre-Trained Language Models
Speaker: Zhuocheng Gong

Long: Are Prompt-based Models Clueless?
Speaker: Pride Kavumba

Findings: How Pre-trained Language Models Capture Factual Knowledge? A Causal-Inspired Analysis
Speaker: Shaobo Li

Long: ProtoTEx: Explaining Model Decisions with Prototype Tensors
Speaker: Anubrata Das

Findings: Does BERT really agree ? Fine-grained Analysis of Lexical Dependence on a Syntactic Task
Speaker: Karim Lasri

Long: A Comparative Study of Faithfulness Metrics for Model Interpretability Methods
Speaker: Chun Sik Chan

Long: Pass off Fish Eyes for Pearls: Attacking Model Selection of Pre-trained Models
Speaker: Biru Zhu

Findings: Factual Consistency of Multilingual Pretrained Language Models
Speaker: constanza fierro

Findings: Exploring the Impact of Negative Samples of Contrastive Learning: A Case Study of Sentence Embedding
Speaker: Rui Cao

Findings: IsoScore: Measuring the Uniformity of Embedding Space Utilization
Speaker: William Rudman

Long: Low-Rank Softmax Can Have Unargmaxable Classes in Theory but Rarely in Practice
Speaker: Andreas Grivas

Long: Signal in Noise: Exploring Meaning Encoded in Random Character Sequences with Character-Aware Language Models
Speaker: Mark Chu

Findings: Detection of Adversarial Examples in Text Classification: Benchmark and Baseline via Robust Density Estimation
Speaker: KiYoon Yoo

Long: Pretraining with Synthetic Language: Studying Transferable Knowledge in Language Models
Speaker: Ryokan Ri

Findings: Local Structure Matters Most: Perturbation Study in NLU
Speaker: Louis Clouatre-Latraverse

Findings: Discontinuous Constituency and BERT: A Case Study of Dutch
Speaker: Konstantinos Kogkalidis

Findings: Probing Multilingual Cognate Prediction Models
Speaker: Clémentine Fourrier

Findings: Interpretable Research Replication Prediction via Variational Contextual Consistency Sentence Masking
Speaker: Tianyi Luo

Long: Can Pre-trained Language Models Interpret Similes as Smart as Human?
Speaker: Qianyu He

Findings: Interpreting the Robustness of Neural NLP Models to Textual Perturbations
Speaker: Yunxiang Zhang

Findings: Exploring the Capacity of a Large-scale Masked Language Model to Recognize Grammatical Errors
Speaker: Ryo Nagata

Findings: On Length Divergence Bias in Textual Matching Models
Speaker: Lan Jiang

VPS3: Language Groundings, Speech and Multimodality

07:30-08:30 (GatherTown)

Long: Cross-Utterance Conditioned VAE for Non-Autoregressive Text-to-Speech
Speaker: Yang Li

Findings: Debiasing Event Understanding for Visual Commonsense Tasks
Speaker: Minji Seo

Long: Skill Induction and Planning with Latent Language
Speaker: Pratyusha Sharma

Long: Multi-Modal Sarcasm Detection via Cross-Modal Graph Convolutional Network
Speaker: Bin Liang, Bin Liang

Long: OpenHands: Making Sign Language Recognition Accessible with Pose-based Pretrained Models across Languages
Speaker: Prem Selvaraj

Long: Things not Written in Text: Exploring Spatial Commonsense from Visual Signals
Speaker: Xiao Liu

Long: Leveraging Visual Knowledge in Language Tasks: An Empirical Study on Intermediate Pre-training for Cross-Modal Knowledge Transfer
Speaker: Woojeong Jin

Long: A Good Prompt Is Worth Millions of Parameters: Low-resource Prompt-based Learning for Vision-Language Models
Speaker: Woojeong Jin

Findings: Modality-specific Learning Rates for Effective Multimodal Additive Late-fusion
Speaker: Yiqun Yao

Findings: Co-VQA : Answering by Interactive Sub Question Sequence
Speaker: Ruonan Wang

Short: XDBERT: Distilling Visual Information to BERT from Cross-Modal Systems to Improve Language Understanding
Speaker: Chan-Jan Hsu

Long: Visual-Language Navigation Pretraining via Prompt-based Environmental Self-exploration
Speaker: Xiwen Liang

Findings: Modular and Parameter-Efficient Multimodal Fusion with Prompting
Speaker: Sheng Liang, Sheng Liang

Long: SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing
Speaker: Junyi Ao

Findings: UNIMO-2: End-to-End Unified Vision-Language Grounded Learning
Speaker: Wei Li

Long: CLIP Models are Few-Shot Learners: Empirical Studies on VQA and Visual Entailment
Speaker: Haoyu Song

Long: On Vision Features in Multimodal Machine Translation
Speaker: Bei Li

Findings: Prior Knowledge and Memory Enriched Transformer for Sign Language Translation
Speaker: Tao Jin

Long: LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding
Speaker: Jiapeng Wang

Long: Learning the Beauty in Songs: Neural Singing Voice Beautifier
Speaker: Jinglin Liu

Long: SUPERB-SG: Enhanced Speech processing Universal PERformance Benchmark for Semantic and Generative Capabilities
Speaker: Hsiang-Sheng Tsai

Findings: Attention as Grounding: Exploring Textual and Cross-Modal Attention on Entities and Relations in Language-and-Vision Transformer
Speaker: Nikolai Ilinykh

Long: Text-Free Prosody-Aware Generative Spoken Language Modeling
Speaker: Eugene Kharitonov

Long: End-to-End Modeling via Information Tree for One-Shot Natural Language Spatial Video Grounding
Speaker: Mengze Li

SRW: Flexible Visual Grounding
Speaker: Yongmin Kim

SRW: Scene-Text Aware Image and Text Retrieval with Dual-Encoder
Speaker: Shumpei Miyawaki

VPS3: Linguistic Theories, Cognitive Modeling and Psycholinguistics

07:30-08:30 (GatherTown)

Long: CogTaskonomy: Cognitively Inspired Task Taxonomy Is Beneficial to Transfer Learning in NLP
Speaker: Yifei Luo

Long: Slangvolution: A Causal Analysis of Semantic Change and Frequency Dynamics in Slang
Speaker: Daphna Keidar

Long: Measuring the Impact of (Psycho-)Linguistic and Readability Features and Their Spill Over Effects on the Prediction of Eye Movement Patterns
Speaker: Daniel Wiechmann

VPS3: Machine Learning for NLP

07:30-08:30 (GatherTown)

Findings: Dual Context-Guided Continuous Prompt Tuning for Few-Shot Learning
Speaker: Jie Zhou

Long: Long-range Sequence Modeling with Predictable Sparse Attention
Speaker: Yimeng Zhuang

Long: GLM: General Language Model Pretraining with Autoregressive Blank Infilling
Speaker: Zhengxiao Du

Long: Improving Meta-learning for Low-resource Text Classification and Generation via Memory Imitation
Speaker: Yingxiu Zhao

Long: Prompt-Based Rule Discovery and Boosting for Interactive Weakly-Supervised Learning
Speaker: Rongzhi Zhang

Long: An Information-theoretic Approach to Prompt Engineering Without Ground Truth Labels
Speaker: Taylor Sorensen

Long: An Information-theoretic Approach to Prompt Engineering Without Ground Truth Labels
Speaker: Taylor Sorensen

Short: P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks
Speaker: Xiao Liu

Findings: Why Exposure Bias Matters: An Imitation Learning Perspective of Error Accumulation in Language Generation
Speaker: Kushal Arora

Findings: Learning to Robustly Aggregate Labeling Functions for Semi-supervised Data Programming
Speaker: Ayush Maheshwari

Long: bert2BERT: Towards Reusable Pretrained Language Models
Speaker: Cheng Chen

Short: When to Use Multi-Task Learning vs Intermediate Fine-Tuning for Pre-Trained Encoder Transfer Learning
Speaker: Orion Weller

Short: LM-BFF-MS: Improving Few-Shot Fine-tuning of Language Models based on Multiple Soft Demonstration Memory
Speaker: Eunhwan Park

Findings: Efficient, Uncertainty-based Moderation of Neural Networks Text Classifiers
Speaker: Jakob Smedegaard Andersen

Findings: Open Vocabulary Extreme Classification Using Generative Models
Speaker: Daniel Simig

Findings: Towards Adversarially Robust Text Classifiers by Learning to Reweight Clean Examples
Speaker: Jianhan Xu

Long: Leveraging Relaxed Equilibrium by Lazy Transition for Sequence Modeling
Speaker: Xi Ai

Findings: Striking a Balance: Alleviating Inconsistency in Pre-trained Models for Symmetric Classification Tasks
Speaker: Ashutosh Kumar

Long: On Continual Model Refinement in Out-of-Distribution Data Streams
Speaker: Bill Yuchen Lin

Outstanding Paper: Compression of Generative Pre-trained Language Models via Quantization
Speaker: Chaofan Tao

Findings: ELLE: Efficient Lifelong Pre-training for Emerging Data
Speaker: Yujia Qin

Findings: Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models
Speaker: Robert Logan

Findings: Perturbations in the Wild: Leveraging Human-Written Text Perturbations for Realistic Adversarial Attack and Defense
Speaker: Thai Le

Findings: Composing Structure-Aware Batches for Pairwise Sentence Classification
Speaker: Andreas Waldis

Long: Fully Hyperbolic Neural Networks
Speaker: Weize Chen

Short: DMix: Adaptive Distance-aware Interpolative Mixup
Speaker: Ramit Sawhney

Findings: Improving Robustness of Language Models from a Geometry-aware Perspective
Speaker: Bin Zhu

Findings: Task-guided Disentangled Tuning for Pretrained Language Models
Speaker: Jiali Zeng

Findings: Word-level Perturbation Considering Word Length and Compositional Subwords
Speaker: Tatsuya Hiraoka

Findings: MetaWeighting: Learning to Weight Tasks in Multi-Task Learning
Speaker: Yuren Mao

Findings: Prompt Tuning for Discriminative Pre-trained Language Models
Speaker: Yuan Yao

Long: Prototypical Verbalizer for Prompt-based Few-shot Tuning
Speaker: ganqu cui

Long: Incorporating Hierarchy into Text Encoder: a Contrastive Learning Approach for Hierarchical Text Classification
Speaker: Zihan Wang

Long: Transkimmer: Transformer Learns to Layer-wise Skim
Speaker: Yue Guan, Yue Guan

Findings: Platt-Bin: Efficient Posterior Calibrated Training for NLP Classifiers
Speaker: Rishabh Singh

Long: That Is a Suspicious Reaction!: Interpreting Logits Variation to Detect NLP Adversarial Attacks
Speaker: Edoardo Mosca

Long: Using Context-to-Vector with Graph Retrofitting to Improve Word Embeddings
Speaker: Jiangbin Zheng

Short: On the Importance of Effectively Adapting Pretrained Language Models for Active Learning
Speaker: Katerina Margatina

CL: To Augment or Not to Augment? A Comparative Study on Text Augmentation Techniques for Low-Resource NLP
Speaker: Gözde Gül Sahin

VPS3: Machine Translation and Multilinguality

07:30-08:30 (GatherTown)

Findings: Investigating Data Variance in Evaluations of Automatic Machine Translation Metrics
Speaker: Jiannan Xiang

Long: Towards Making the Most of Cross-Lingual Transfer for Zero-Shot Neural Machine Translation
Speaker: Guanhua Chen

Findings: MR-P: A Parallel Decoding Algorithm for Iterative Refinement Non-Autoregressive Translation
Speaker: Hao Cheng

Findings: Better Quality Estimation for Low Resource Corpus Mining
Speaker: Muhammed Yusuf Kocyigit

Findings: Fast Nearest Neighbor Machine Translation
Speaker: Yuxian Meng

Findings: Domain Generalisation of NMT: Fusing Adapters with Leave-One-Domain-Out Training
Speaker: Trang Vu

Findings: Automatic Song Translation for Tonal Languages
Speaker: Fenfei Guo

Long: Flow-Adapter Architecture for Unsupervised Machine Translation
Speaker: Yihong Liu

Long: Bias Mitigation in Machine Translation Quality Estimation
Speaker: Hanna Behnke

Long: Enhancing Cross-lingual Natural Language Inference by Prompt-learning from Cross-lingual Templates
Speaker: Kunxun Qi

Long: Overcoming Catastrophic Forgetting beyond Continual Learning: Balanced Training for Neural Machine Translation
Speaker: Chenze Shao

Findings: Graph Neural Networks for Multiparallel Word Alignment
Speaker: Ayyoob ImaniGooghari

Long: Learning Confidence for Transformer-based Neural Machine Translation
Speaker: Yu Lu

Long: Redistributing Low-Frequency Words: Making the Most of Monolingual Data in Non-Autoregressive Translation
Speaker: Liang Ding

Long: Measuring and Mitigating Name Biases in Neural Machine Translation
Speaker: Jun Wang

Long: Understanding and Improving Sequence-to-Sequence Pretraining for Neural Machine Translation
Speaker: Wenxuan Wang

Findings: DaLC: Domain Adaptation Learning Curve Prediction for Neural Machine Translation
Speaker: Cheonbok Park

Long: Confidence Based Bidirectional Global Context Aware Training Framework for Neural Machine Translation
Speaker: Chulun Zhou

Findings: IndicBART: A Pre-trained Model for Indic Natural Language Generation
Speaker: Raj Dabre, Raj Dabre

Findings: The impact of lexical and grammatical processing on generating code from natural language
Speaker: Nathanaël Beau

Findings: Combining Static and Contextualised Multilingual Embeddings
Speaker: Katharina Hämmerl

Long: Cross-Lingual Phrase Retrieval
Speaker: Heqi Zheng

Long: Scheduled Multi-task Learning for Neural Chat Translation
Speaker: Yunlong Liang

Findings: Why don't people use character-level machine translation?
Speaker: Jindřich Libovický

Long: Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder Translation Models
Speaker: Lorenzo Lupo

Findings: Single Model Ensemble for Subword Regularized Models in Low-Resource Machine Translation
Speaker: Sho Takase

Long: Can Synthetic Translations Improve Bitext Quality?
Speaker: Eleftheria Briakou

Short: S^4-Tuning: A Simple Cross-lingual Sub-network Tuning Method
Speaker: Runxin Xu

Findings: Breaking Down Multilingual Machine Translation
Speaker: Ting-Rui Chiang

Findings: Gaussian Multi-head Attention for Simultaneous Machine Translation
Speaker: Shaolei Zhang

Long: Neural Machine Translation with Phrase-Level Universal Visual Representations
Speaker: Qingkai Fang

Long: MSP: Multi-Stage Prompting for Making Pre-trained Language Models Better Translators
Speaker: Zhixing Tan

Findings: Leveraging Knowledge in Multilingual Commonsense Reasoning
Speaker: Yuwei Fang

Short: Triangular Transfer: Freezing the Pivot for Triangular Machine Translation
Speaker: Meng Zhang

Long: Bridging the Data Gap between Training and Inference for Unsupervised Neural Machine Translation
Speaker: Zhiwei He

Long: Reducing Position Bias in Simultaneous Machine Translation with Length-Aware Framework
Speaker: Shaolei Zhang

Long: STEMM: Self-learning with Speech-text Manifold Mixup for Speech Translation
Speaker: Qingkai Fang

Long: Integrating Vectorized Lexical Constraints for Neural Machine Translation
Speaker: Shuo Wang

Long: Prediction Difference Regularization against Perturbation for Neural Machine Translation
Speaker: Dengji Guo

Long: Learning Adaptive Segmentation Policy for End-to-End Simultaneous Translation
Speaker: Ruiqing Zhang

Long: ODE Transformer: An Ordinary Differential Equation-Inspired Model for Sequence Generation
Speaker: Bei Li

Findings: CrossAligner & Co: Zero-Shot Transfer Methods for Task-Oriented Cross-lingual Natural Language Understanding
Speaker: Milan Gritta

Findings: Structural Supervision for Word Alignment and Machine Translation
Speaker: Lei Li

Long: Bilingual alignment transfers to multilingual alignment for unsupervised parallel text mining
Speaker: Chih-chan Tien

CL: Challenges of Neural Machine Translation for Short Texts
Speaker: Yu Wan

SRW: English-Malay Cross-Lingual Embedding Alignment using Bilingual Lexicon Augmentation
Speaker: Ying Hao Lim

VPS3: NLP Applications

07:30-08:30 (GatherTown)

Short: Automatic Detection of Entity-Manipulated Text using Factual Knowledge
Speaker: Ganesh Jawahar, Ganesh Jawahar

Short: Buy Tesla, Sell Ford: Assessing Implicit Stock Market Preference in Pre-trained Language Models
Speaker: Cheng Yu Chuang

Findings: Learning and Evaluating Character Representations in Novels
Speaker: Naoya Inoue

Findings: MDCSpell: A Multi-task Detector-Corrector Framework for Chinese Spelling Correction
Speaker: Chenxi Zhu

Findings: Constructing Open Cloze Tests Using Generation and Discrimination Capabilities of Transformers
Speaker: Mariano Felice

Findings: A Novel Framework Based on Medical Concept Driven Attention for Explainable Medical Code Prediction via External Knowledge
Speaker: Tao Wang

Long: ClusterFormer: Neural Clustering Attention for Efficient and Effective Transformer
Speaker: Ningning Wang

Findings: Sibylvariant Transformations for Robust Text Classification
Speaker: Fabrice Harel-Canada

Long: KenMeSH: Knowledge-enhanced End-to-end Biomedical Text Labelling
Speaker: Xindi Wang

Findings: Improved Multi-label Classification under Temporal Concept Drift: Rethinking Group-Robust Algorithms in a Label-Wise Setting
Speaker: Ilias Chalkidis

Long: Textomics: A Dataset for Genomics Data Summary Generation
Speaker: Mu-Chun Wang

Long: Educational Question Generation of Children Storybooks via Question Type Distribution Learning and Event-centric Summarization
Speaker: Zhenjie Zhao

Long: A Neural Network Architecture for Program Understanding Inspired by Human Behaviors
Speaker: Renyu Zhu

Findings: Improving Chinese Grammatical Error Detection via Data augmentation by Conditional Error Generation
Speaker: tianchi yue

Findings: CRASpell: A Contextual Typo Robust Approach to Improve Chinese Spelling Correction
Speaker: Shulin Liu

Findings: How Can Cross-lingual Knowledge Contribute Better to Fine-Grained Entity Typing?
Speaker: Hailong Jin

Long: Continual Pre-training of Language Models for Math Problem Understanding with Syntax-Aware Memory Network
Speaker: Zheng Gong

Long: Multitasking Framework for Unsupervised Simple Definition Generation
Speaker: Cunliang Kong

Long: Learning to Reason Deductively: Math Word Problem Solving as Complex Relation Extraction
Speaker: Zhanming Jie

Findings: The Past Mistake is the Future Wisdom: Error-driven Contrastive Probability Optimization for Chinese Spell Checking
Speaker: Li Yinghui

Findings: Type-Driven Multi-Turn Corrections for Grammatical Error Correction
Speaker: Shaopeng Lai

Long: ReACC: A Retrieval-Augmented Code Completion Framework
Speaker: Shuai Lu

Long: Guided Attention Multimodal Multitask Financial Forecasting with Inter-Company Relationships and Global and Local News
Speaker: Gary Ang

Findings: A Neural Pairwise Ranking Model for Readability Assessment
Speaker: Justin Lee

Short: Code Synonyms Do Matter: Multiple Synonyms Matching Network for Automatic ICD Coding
Speaker: Zheng Yuan

Long: Modeling Persuasive Discourse to Adaptively Support Students' Argumentative Writing
Speaker: Thiemo Wambsganss

CL: Novelty Detection: A Perspective from Natural Language Processing
Speaker: Tirthankar Ghosal

VPS3: Phonology, Morphology and Word Segmentation

07:30-08:30 (GatherTown)

Findings: Morphosyntactic Tagging with Pre-trained Language Models for Arabic and its Dialects
Speaker: Go Inoue

Findings: Weighted self Distillation for Chinese word segmentation
Speaker: Rian He

Long: That Slepen Al the Nyght with Open Ye! Cross-era Sequence Segmentation with Switch-memory
Speaker: Xuemei Tang

Findings: Word Segmentation by Separation Inference for East Asian Languages
Speaker: Yu Tong

Findings: Unsupervised Chinese Word Segmentation with BERT Oriented Probing and Transformation
Speaker: Wei Li

VPS3: Question Answering

07:30-08:30 (GatherTown)

Long: Learning to Imagine: Integrating Counterfactual Thinking in Neural Discrete Reasoning
Speaker: Moxin Li

Long: KaFSP: Knowledge-Aware Fuzzy Semantic Parsing for Conversational Question Answering over a Large-Scale Knowledge Base
Speaker: Junzhuo Li

Findings: Plug-and-Play Adaptation for Continuously-updated QA
Speaker: Kyungjae Lee

Long: Learning Disentangled Semantic Representations for Zero-Shot Cross-Lingual Transfer in Multilingual Machine Reading Comprehension
Speaker: Linjuan Wu

Findings: Read before Generate! Faithful Long Form Question Answering with Machine Reading
Speaker: Dan Su

Findings: Fact-Tree Reasoning for N-ary Question Answering over Knowledge Graphs
Speaker: YAO ZHANG

Findings: Answer Uncertainty and Unanswerability in Multiple-Choice Machine Reading Comprehension
Speaker: Vatsal Raina

Long: Answering Open-Domain Multi-Answer Questions via a Recall-then-Verify Framework
Speaker: Zhihong Shao

Findings: Calibration of Machine Reading Systems at Scale
Speaker: Shehzaad Dhuliawala

Findings: Relevant CommonSense Subgraphs for "What if..." Procedural Reasoning
Speaker: Chen Zheng

Findings: Investigating Selective Prediction Approaches Across Several Tasks in IID, OOD, and Adversarial Settings
Speaker: Neeraj Varshney

Long: MMCoQA: Conversational Question Answering over Text, Tables, and Images
Speaker: Yongqi Li

Long: QAConv: Question Answering on Informative Conversations
Speaker: Chien-Sheng Wu

Long: Subgraph Retrieval Enhanced Model for Multi-hop Knowledge Base Question Answering
Speaker: Jing Zhang

Findings: MERIt: Meta-Path Guided Contrastive Learning for Logical Reasoning
Speaker: Fangkai Jiao

Findings: Implicit Relation Linking for Question Answering over Knowledge Graph
Speaker: Yao Zhao

Long: Program Transfer for Answering Complex Questions over Knowledge Bases
Speaker: Shulin Cao

VPS3: Resources and Evaluation

07:30-08:30 (GatherTown)

Findings: RuCCoN: Clinical Concept Normalization in Russian
Speaker: Aleksandr Nesterov

Findings: MIMICause: Representation and automatic extraction of causal relation types from clinical notes
Speaker: Vivek Khetan

Findings: Thai Nested Named Entity Recognition Corpus
Speaker: Weerayut Buaphet

Long: A Taxonomy of Empathetic Questions in Social Dialogs
Speaker: Ekaterina Svikhnushina

Best Resource: DiBiMT: A Novel Benchmark for Measuring Word Sense Disambiguation Biases in Machine Translation
Speaker: Niccolò Campolungo

Findings: xGQA: Cross-Lingual Visual Question Answering
Speaker: Jonas Pfeiffer

Long: WikiDiverse: A Multimodal Entity Linking Dataset with Diversified Contextual Topics and Entity Types
Speaker: Xuwu Wang

Long: FaVIQ: FAct Verification from Information-seeking Questions
Speaker: Jungsoo Park

Findings: ZiNet: Linking Chinese Characters Spanning Three Thousand Years
Speaker: Yang Chi

Long: KQA Pro: A Dataset with Explicit Compositional Programs for Complex Question Answering over Knowledge Base
Speaker: Shulin Cao

Long: Learning to Rank Visual Stories From Human Ranking Data
Speaker: Chi-Yang Hsu

Long: TwittIrish: A Universal Dependencies Treebank of Tweets in Modern Irish
Speaker: Lauren Cassidy

Long: What Makes Reading Comprehension Questions Difficult?
Speaker: Saku Sugawara

Findings: HLDC: Hindi Legal Documents Corpus
Speaker: Arnav Kapoor

Findings: Benchmarking Answer Verification Methods for Question Answering-Based Summarization Evaluation Metrics
Speaker: Daniel Deutsch

Findings: Chinese Synesthesia Detection: New Dataset and Models
Speaker: Xiaotong Jiang

Findings: On the Safety of Conversational Models: Taxonomy, Dataset, and Benchmark
Speaker: Hao Sun

Findings: Translation Error Detection as Rationale Extraction
Speaker: Marina Fomicheva

TACL: LOT: A Story-Centric Benchmark for Evaluating Chinese Long Text Understanding and Generation
Speaker: Jian Guan

SRW: Scoping natural language processing in Indonesian and Malay for education applications
Speaker: Zara Maxwelll-Smith

VPS3: Semantics

07:30-08:30 (GatherTown)

Findings: Sememe Prediction for BabelNet Synsets using Multilingual and Multimodal Information
Speaker: Fanchao Qi

Findings: Open Relation Modeling: Learning to Define Relations between Entities
Speaker: Jie Huang

Long: FaiRR: Faithful and Robust Deductive Reasoning over Natural Language
Speaker: Soumya Sanyal

Findings: To be or not to be an Integer? Encoding Variables for Mathematical Text
Speaker: Deborah Ferreira

Findings: CoCoLM: Complex Commonsense Enhanced Language Model with Discourse Relations
Speaker: Changlong Yu

Findings: Cross-lingual Inference with A Chinese Entailment Graph
Speaker: Tianyi Li

Findings: An Isotropy Analysis in the Multilingual BERT Embedding Space
Speaker: Sara Rajaee

Findings: Learning from Missing Relations: Contrastive Learning with Commonsense Knowledge Graphs for Commonsense Inference
Speaker: Yong-Ho Jung

Long: ExtEnD: Extractive Entity Disambiguation
Speaker: Edoardo Barba, Edoardo Barba

Findings: Capture Human Disagreement Distributions by Calibrated Networks for Natural Language Inference
Speaker: Yuxia Wang

Long: ClarET: Pre-training a Correlation-Aware Context-To-Event Transformer for Event-Centric Generation and Classification
Speaker: Yucheng Zhou

Long: Contextual Representation Learning beyond Masked Language Modeling
Speaker: Zhiyi Fu

Findings: Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models
Speaker: Jianmo Ni

Long: Improving Event Representation via Simultaneous Weakly Supervised Contrastive Learning and Clustering
Speaker: Jun Gao

Findings: Unsupervised Natural Language Inference Using PHL Triplet Generation
Speaker: Neeraj Varshney

Findings: ASCM: An Answer Space Clustered Prompting Method without Answer Engineering
Speaker: Zhen Wang

Long: Rare and Zero-shot Word Sense Disambiguation using Z-Reweighting
Speaker: Ying Su

Long: Nibbling at the Hard Core of Word Sense Disambiguation
Speaker: Marco Maru

Findings: Lacking the Embedding of a Word? Look it up into a Traditional Dictionary
Speaker: Elena Sofia Ruzzetti

Long: Large Scale Substitution-based Word Sense Induction
Speaker: Matan Eyal

Long: A Contrastive Framework for Learning Sentence Representations from Pairwise and Triple-wise Perspective in Angular Space
Speaker: Yuhao Zhang

Long: Bridging the Generalization Gap in Text-to-SQL Parsing with Schema Expansion
Speaker: Chen Zhao

Long: Graph Pre-training for AMR Parsing and Generation
Speaker: Xuefeng Bai

Long: Debiased Contrastive Learning of Unsupervised Sentence Representations
Speaker: Kun Zhou

Findings: Divide and Conquer: Text Semantic Matching with Disentangled Keywords and Intents
Speaker: Yicheng Zou

Long: Learning to Generate Programs for Table Fact Verification via Structure-Aware Semantic Parsing
Speaker: Suixin Ou

Findings: Towards Collaborative Neural-Symbolic Graph Semantic Parsing via Uncertainty
Speaker: Zi Lin

TACL: Is My Model Using The Right Evidence? Systematic Probes for Examining Evidence-Based Tabular Reasoning
Speaker: Vivek Gupta

SRW: Logical Inference for Counting on Semi-structured Tables
Speaker: Tomoya Kurosawa

SRW: Compositional Semantics and Inference System for Temporal Order based on Japanese CCG
Speaker: Tomoki Sugimoto

VPS3: Sentiment Analysis, Stylistic Analysis, and Argument Mining

07:30-08:30 (GatherTown)

Long: JointCL: A Joint Contrastive Learning Framework for Zero-Shot Stance Detection
Speaker: Bin Liang, Bin Liang

Long: Discrete Opinion Tree Induction for Aspect-based Sentiment Analysis
Speaker: Chenhua Chen

Long: Vision-Language Pre-Training for Multimodal Aspect-Based Sentiment Analysis
Speaker: Yan Ling

Long: "You might think about slightly revising the title": Identifying Hedges in Peer-tutoring Interactions
Speaker: Yann Raphalen

Findings: EmoCaps: Emotion Capsule based Model for Conversational Emotion Recognition
Speaker: Zaijing Li

Long: Identifying Chinese Opinion Expressions with Extremely-Noisy Crowdsourcing Annotations
Speaker: Xin Zhang

Findings: Seq2Path: Generating Sentiment Tuples as Paths of a Tree
Speaker: Yue Mao

Long: Incorporating Stock Market Signals for Twitter Stance Detection
Speaker: Costanza Conforti

Long: Effective Token Graph Modeling using a Novel Labeling Strategy for Structured Sentiment Analysis
Speaker: Wenxuan Shi

Long: M3ED: Multi-modal Multi-scene Multi-label Emotional Dialogue Database
Speaker: Jinming Zhao

Long: Multimodal Sarcasm Target Identification in Tweets
Speaker: Jiquan Wang, Jiquan Wang

SRW: Towards Detecting Political Bias in Hindi News Articles
Speaker: Samyak Agrawal

VPS3: Special Theme on Language Diversity: From Low Resource to Endangered

07:30-08:30 (GatherTown)

Findings: Towards Responsible Natural Language Annotation for the Varieties of Arabic
Speaker: Stevie Bergman

Findings: BPE vs. Morphological Segmentation: A Case Study on Machine Translation of Four Polysynthetic Languages
Speaker: Manuel Mager

Findings: Morphological Processing of Low-Resource Languages: Where We Are and What's Next
Speaker: Adam Wiemerslage

Findings: Dim Wihl Gat Tun: The Case for Linguistic Expertise in NLP for Under-Documented Languages
Speaker: Clarissa Forbes

Short: Machine Translation for Livonian: Catering to 20 Speakers
Speaker: Matīss Rikters

Findings: OneAligner: Zero-shot Cross-lingual Transfer with One Rich-Resource Language Pair for Low-Resource Sentence Retrieval
Speaker: Tong Niu

Long: Multilingual unsupervised sequence segmentation transfers to extremely low-resource languages
Speaker: Agatha Downey

VPS3: Summarization

07:30-08:30 (GatherTown)

Long: Modeling Hierarchical Syntax Structure with Triplet Position for Source Code Summarization
Speaker: Juncai Guo

Findings: NEWTS: A Corpus for News Topic-Focused Summarization
Speaker: Seyed Ali Bahrainian

Findings: End-to-End Segmentation-based News Summarization
Speaker: Yang Liu

Long: Unsupervised Extractive Opinion Summarization Using Sparse Coding
Speaker: Somnath Basu Roy Chowdhury

Long: A Variational Hierarchical Model for Neural Cross-Lingual Summarization
Speaker: Yunlong Liang

Findings: Revisiting Automatic Evaluation of Extractive Summarization Task: Can We Do Better than ROUGE?
Speaker: Mousumi Akter

Long: Other Roles Matter! Enhancing Role-Oriented Dialogue Summarization via Role Interactions
Speaker: Haitao Lin

Findings: Training Dynamics for Text Summarization Models
Speaker: Tanya Goyal

Long: Graph Enhanced Contrastive Learning for Radiology Findings Summarization
Speaker: Jinpeng Hu

Long: PRIMERA: Pyramid-based Masked Sentence Pre-training for Multi-document Summarization
Speaker: Wen Xiao

Long: Length Control in Abstractive Summarization by Pretraining Information Selection
Speaker: Yizhu Liu

VPS3: Syntax: Tagging, Chunking and Parsing

07:30-08:30 (GatherTown)

Findings: Challenges to Open-Domain Constituency Parsing
Speaker: Sen Yang, Sen Yang

Findings: SyMCoM - Syntactic Measure of Code Mixing A Study Of English-Hindi Code-Mixing
Speaker: Prashant Kodali

Long: Semi-supervised Domain Adaptation for Dependency Parsing with Dynamic Matching Network
Speaker: Ying Li

Findings: Combining (Second-Order) Graph-Based and Headed-Span-Based Projective Dependency Parsing
Speaker: Songlin Yang

Long: Headed-Span-Based Projective Dependency Parsing
Speaker: Songlin Yang

Long: Bottom-Up Constituency Parsing and Nested Named Entity Recognition with Pointer Networks
Speaker: Songlin Yang

Long: Dependency Parsing as MRC-based Span-Span Prediction
Speaker: Leilei Gan

Findings: Auxiliary tasks to boost Biaffine Semantic Dependency Parsing
Speaker: Marie Candito

Short: Zero-Shot Dependency Parsing with Worst-Case Aware Automated Curriculum Learning
Speaker: Miryam De Lhoneux

Findings: Bridging Pre-trained Language Models and Hand-crafted Features for Unsupervised POS Tagging
Speaker: Houquan Zhou

Long: Phrase-aware Unsupervised Constituency Parsing
Speaker: Xiaotao Gu

Long: Compositional Generalization in Dependency Parsing
Speaker: Emily Goodwin

Findings: Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise
Speaker: Noëmi Aepli

TACL: Instance-Based Dependency Parsing
Speaker: Hiroki Ouchi

Keynote 3: Panel on "Supporting Linguistic Diversity" (chaired by Steven Bird)

09:00-10:15 - **Auditorium** (Auditorium)

Coffee Break

10:15-10:45 - **Auditorium** (Forum)

Session 6 - 10:45-12:15

Computational Social Science and Cultural Analytics

10:45-12:15 (Liffey Hall 1)

10:45-11:00 (Liffey Hall 1)
Reports of personal experiences and stories in argumentation: datasets and analysis
Neele Falk and Gabriella Lapesa
Reports of personal experiences or stories can play a crucial role in argumentation, as they represent an immediate and (often) relatable way to back up one's position with respect to a given topic. They are easy to understand and increase empathy: this makes them powerful in argumentation. The impact of personal reports and stories in argumentation has been studied in the Social Sciences, but it is still largely underexplored in NLP. Our work is the first step towards filling this gap: our goal is to develop robust classifiers to identify documents containing personal experiences and reports. The main challenge is the scarcity of annotated data: our solution is to leverage existing annotations to be able to scale-up the analysis. Our contribution is two-fold. First, we conduct a set of in-domain and cross-domain experiments involving three datasets (two from Argument Mining, one from the Social Sciences), modeling architectures, training setups and fine-tuning options tailored to the involved domains. We show that despite the differences among datasets and annotations, robust cross-domain classification is possible. Second, we employ linear regression for performance mining, identifying performance trends both for overall classification performance and individual classifier predictions.

11:00-11:15 (Liffey Hall 1)
Inducing Positive Perspectives with Text Reframing
Caleb Ziems, Minzhi Li, Anthony Zhang and Diyi Yang
Sentiment transfer is one popular example of a text style transfer task, where the goal is to reverse the sentiment polarity of a text. With a sentiment reversal comes also a reversal in meaning. We introduce a different but related task called positive reframing in which we neutralize a negative point of view and generate a more positive perspective for the author without contradicting the original meaning. Our insistence on meaning preservation makes positive reframing a challenging and semantically rich task. To facilitate rapid progress, we introduce a large-scale benchmark, Positive Psychology Frames, with 8,349 sentence pairs and 12,755 structured annotations to explain positive reframing in terms of six theoretically-motivated reframing strategies. Then we evaluate a set of state-of-the-art text style transfer models, and conclude by discussing key challenges and directions for future work.

11:15-11:30 (Liffey Hall 1)
Automatic Identification and Classification of Bragging in Social Media
Mali Jin, Daniel Preotiuc-Pietro, A. Seza Doğruöz and Nikolaos Aletras
Bragging is a speech act employed with the goal of constructing a favorable self-image through positive statements about oneself. It is widespread in daily communication and especially popular in social media, where users aim to build a positive image of their persona directly or indirectly. In this paper, we present the first large scale study of bragging in computational linguistics, building on previous research in linguistics and pragmatics. To facilitate this, we introduce a new publicly available data set of tweets annotated for bragging and their types. We empirically evaluate different transformer-based models injected with linguistic information in (a) binary bragging classification, i.e., if tweets contain bragging statements or not; and (b) multi-class bragging type prediction including not bragging. Our results show that our models can predict bragging with macro F1 up to 72.42 and 35.95 in the binary and multi-class classification tasks respectively. Finally, we present an extensive linguistic and error analysis of bragging prediction to guide future research on this topic.

11:30-11:45 (Liffey Hall 1)
Should a Chatbot be Sarcastic? Understanding User Preferences Towards Sarcasm Generation
Silviu Vlad Oprea, Steven R. Wilson and Walid Magdy

Previous sarcasm generation research has focused on how to generate text that people perceive as sarcastic to create more human-like interactions. In this paper, we argue that we should first turn our attention to the question of when sarcasm should be generated, finding that humans consider sarcastic responses inappropriate to many input utterances. Next, we use a theory-driven framework for generating sarcastic responses, which allows us to control the linguistic devices included during generation. For each device, we investigate how much humans associate it with sarcasm, finding that pragmatic insincerity and emotional markers are devices crucial for making sarcasm recognisable.

11:45-12:00 (Liffey Hall 1)
Improving the Generalizability of Depression Detection by Leveraging Clinical Questionnaires
Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet and Arman Cohan
Automated methods have been widely used to identify and analyze mental health conditions (e.g., depression) from various sources of information, including social media. Yet, deployment of such models in real-world healthcare applications faces challenges including poor out-of-domain generalization and lack of trust in black box models. In this work, we propose approaches for depression detection that are constrained to different degrees by the presence of symptoms described in PHQ9, a questionnaire used by clinicians in the depression screening process. In dataset-transfer experiments on three social media datasets, we find that grounding the model in PHQ9's symptoms substantially improves its ability to generalize to out-of-distribution data compared to a standard BERT-based approach. Furthermore, this approach can still perform competitively on in-domain data. These results and our qualitative analyses suggest that grounding model predictions in clinically-relevant symptoms can improve generalizability while producing a model that is easier to inspect.

12:00-12:15 (Liffey Hall 1)
Leveraging Wikipedia article evolution for promotional tone detection
Christine De Kock and Andreas Vlachos
Detecting biased language is useful for a variety of applications, such as identifying hyperpartisan news sources or flagging one-sided rhetoric. In this work we introduce WikiEvolve, a dataset for document-level promotional tone detection. Unlike previously proposed datasets, WikiEvolve contains seven versions of the same article from Wikipedia, from different points in its revision history; one with promotional tone, and six without it. This allows for obtaining more precise training signal for learning models from promotional tone detection. We adapt the previously proposed gradient reversal layer framework to encode two article versions simultaneously and thus leverage this additional training signal. In our experiments, our proposed adaptation of gradient reversal improves the accuracy of four different architectures on both in-domain and out-of-domain evaluation.

Dialogue and Interactive Systems 3

10:45-12:15 (The Liffey B)

10:45-11:00 (The Liffey B)
CICERO: A Dataset for Contextualized Commonsense Inference in Dialogues
Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea and Soujanya Poria
This paper addresses the problem of dialogue reasoning with contextualized commonsense inference. We curate CICERO, a dataset of dyadic conversations with five types of utterance-level reasoning-based inferences: cause, subsequent event, prerequisite, motivation, and emotional reaction. The dataset contains 53,105 of such inferences from 5,672 dialogues. We use this dataset to solve relevant generative and discriminative tasks: generation of cause and subsequent event; generation of prerequisite, motivation, and listener's emotional reaction; and selection of plausible alternatives. Our results ascertain the value of such dialogue-centric commonsense knowledge datasets. It is our hope that CICERO will open new research avenues into commonsense-based dialogue reasoning.

11:00-11:15 (The Liffey B)
Dynamic Schema Graph Fusion Network for Multi-Domain Dialogue State Tracking
Yue Feng, Aldo Lipani, Fanghua Ye, Qiang Zhang and Emine Yilmaz
Dialogue State Tracking (DST) aims to keep track of users' intentions during the course of a conversation. In DST, modelling the relations among domains and slots is still an under-studied problem. Existing approaches that have considered such relations generally fall short in: (1) fusing prior slot-domain membership relations and dialogue-aware dynamic slot relations explicitly, and (2) generalizing to unseen domains. To address these issues, we propose a novel **D**ynamic **S**chema **G**raph **F**usion **Net**work (**DSGFNet**), which generates a dynamic schema graph to explicitly fuse the prior slot-domain membership relations and dialogue-aware dynamic slot relations. It also uses the schemata to facilitate knowledge transfer to new domains. DSGFNet consists of a dialogue utterance encoder, a schema graph encoder, a dialogue-aware schema graph evolving network, and a schema graph enhanced dialogue state decoder. Empirical results on benchmark datasets (i.e., SGD, MultiWOZ2.1, and MultiWOZ2.2), show that DSGFNet outperforms existing methods.

11:15-11:30 (The Liffey B)
Where to Go for the Holidays: Towards Mixed-Type Dialogs for Clarification of User Goals
Zeming Liu, Jun Xu, Zeyang Lei, Haifeng Wang, Zheng-Yu Niu and Hua Wu
Most dialog systems posit that users have figured out clear and specific goals before starting an interaction. For example, users have determined the departure, the destination, and the travel time for booking a flight. However, in many scenarios, limited by experience and knowledge, users may know what they need, but still struggle to figure out clear and specific goals by determining all the necessary slots.
In this paper, we identify this challenge, and make a step forward by collecting a new human-to-human mixed-type dialog corpus. It contains 5k dialog sessions and 168k utterances for 4 dialog types and 5 domains. Within each session, an agent first provides user-goal-related knowledge to help figure out clear and specific goals, and then help achieve them.
Furthermore, we propose a mixed-type dialog model with a novel Prompt-based continual learning mechanism. Specifically, the mechanism enables the model to continually strengthen its ability on any specific type by utilizing existing dialog corpora effectively.

11:30-11:45 (The Liffey B)
GlobalWoZ: Globalizing MultiWoZ to Develop Multilingual Task-Oriented Dialogue Systems
Bosheng Ding, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si and Chunyan Miao
Over the last few years, there has been a move towards data curation for multilingual task-oriented dialogue (ToD) systems that can serve

people speaking different languages. However, existing multilingual ToD datasets either have a limited coverage of languages due to the high cost of data curation, or ignore the fact that dialogue entities barely exist in countries speaking these languages. To tackle these limitations, we introduce a novel data curation method that generates GlobalWoZ — a large-scale multilingual ToD dataset globalized from an English ToD dataset for three unexplored use cases of multilingual ToD systems. Our method is based on translating dialogue templates and filling them with local entities in the target-language countries. Besides, we extend the coverage of target languages to 20 languages. We will release our dataset and a set of strong baselines to encourage research on multilingual ToD systems for real use cases.

11:45-12:00 (The Liffey B)
[TACL] **Designing an Automatic Agent for Repeated Language based Persuasion Games**
Roi Reichart, Maya Raifer, Guy Rotman, Reut Apel and Moshe Tennenholtz

11:45-12:00 (The Liffey B)
Think Before You Speak: Explicitly Generating Implicit Commonsense Knowledge for Response Generation
Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu and Dilek Hakkani-Tur
Implicit knowledge, such as common sense, is key to fluid human conversations. Current neural response generation (RG) models are trained to generate responses directly, omitting unstated implicit knowledge. In this paper, we present Think-Before-Speaking (TBS), a generative approach to first externalize implicit commonsense knowledge ($think$) and use this knowledge to generate responses ($speak$). We argue that externalizing implicit knowledge allows more efficient learning, produces more informative responses, and enables more explainable models. We analyze different choices to collect knowledge-aligned dialogues, represent implicit knowledge, and transition between knowledge and dialogues. Empirical results show TBS models outperform end-to-end and knowledge-augmented RG baselines on most automatic metrics and generate more informative, specific, and commonsense-following responses, as evaluated by human annotators. TBS also generates $knowledge$ that makes sense and is relevant to the dialogue around 85

Generation 2

10:45-12:15 (Wicklow Hall 1)

10:45-11:00 (Wicklow Hall 1)
Quality Controlled Paraphrase Generation
Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shnayderman, Noam Slonim and Liat Ein-Dor
Paraphrase generation has been widely used in various downstream tasks. Most tasks benefit mainly from high quality paraphrases, namely those that are semantically similar to, yet linguistically diverse from, the original sentence. Generating high-quality paraphrases is challenging as it becomes increasingly hard to preserve meaning as linguistic diversity increases. Recent works achieve nice results by controlling specific aspects of the paraphrase, such as its syntactic tree. However, they do not allow to directly control the quality of the generated paraphrase, and suffer from low flexibility and scalability. Here we propose QCPG, a quality-guided controlled paraphrase generation model, that allows directly controlling the quality dimensions. Furthermore, we suggest a method that given a sentence, identifies points in the quality control space that are expected to yield optimal generated paraphrases. We show that our method is able to generate paraphrases which maintain the original meaning while achieving higher diversity than the uncontrolled baseline. The models, the code, and the data can be found in https://github.com/IBM/quality-controlled-paraphrase-generation.

11:00-11:15 (Wicklow Hall 1)
An Imitation Learning Curriculum for Text Editing with Non-Autoregressive Models
Sweta Agrawal and Marine Carpuat
We propose a framework for training non-autoregressive sequence-to-sequence models for editing tasks, where the original input sequence is iteratively edited to produce the output. We show that the imitation learning algorithms designed to train such models for machine translation introduces mismatches between training and inference that lead to undertraining and poor generalization in editing scenarios. We address this issue with two complementary strategies: 1) a roll-in policy that exposes the model to intermediate training sequences that it is more likely to encounter during inference, 2) a curriculum that presents easy-to-learn edit operations first, gradually increasing the difficulty of training samples as the model becomes competent. We show the efficacy of these strategies on two challenging English editing tasks: controllable text simplification and abstractive summarization. Our approach significantly improves output quality on both tasks and controls output complexity better on the simplification task.

11:15-11:30 (Wicklow Hall 1)
Updated Headline Generation: Creating Updated Summaries for Evolving News Stories
Sheena Panthaplackel, Adrian Benton and Mark Dredze
We propose the task of updated headline generation, in which a system generates a headline for an updated article, considering both the previous article and headline. The system must identify the novel information in the article update, and modify the existing headline accordingly. We create data for this task using the NewsEdits corpus by automatically identifying contiguous article versions that are likely to require a substantive headline update. We find that models conditioned on the prior headline and body revisions produce headlines judged by humans to be as factual as gold headlines while making fewer unnecessary edits compared to a standard headline generation model. Our experiments establish benchmarks for this new contextual summarization task.

11:30-11:45 (Wicklow Hall 1)
Hierarchical Sketch Induction for Paraphrase Generation
Tom Hosking, Hao Tang and Mirella Lapata
We propose a generative model of paraphrase generation, that encourages syntactic diversity by conditioning on an explicit syntactic sketch. We introduce Hierarchical Refinement Quantized Variational Autoencoders (HRQ-VAE), a method for learning decompositions of dense encodings as a sequence of discrete latent variables that make iterative refinements of increasing granularity. This hierarchy of codes is learned through end-to-end training, and represents fine-to-coarse grained information about the input. We use HRQ-VAE to encode the syntactic form of an input sentence as a path through the hierarchy, allowing us to more easily predict syntactic sketches at test time. Extensive experiments, including a human evaluation, confirm that HRQ-VAE learns a hierarchical representation of the input space, and generates paraphrases of

higher quality than previous systems.

11:45-12:00 (Wicklow Hall 1)
Few-shot Controllable Style Transfer for Low-Resource Multilingual Settings
Kalpesh Krishna, Deepak Nathani, Xavier Garcia, Bidisha Samanta and Partha Talukdar
Style transfer is the task of rewriting a sentence into a target style while approximately preserving content. While most prior literature assumes access to a large style-labelled corpus, recent work (Riley et al. 2021) has attempted "few-shot" style transfer using only 3-10 sentences at inference for style extraction. In this work we study a relevant low-resource setting: style transfer for languages where no style-labelled corpora are available. We notice that existing few-shot methods perform this task poorly, often copying inputs verbatim. We push the state-of-the-art for few-shot style transfer with a new method modeling the stylistic difference between paraphrases. When compared to prior work, our model achieves 2-3x better performance in formality transfer and code-mixing addition across seven languages. Moreover, our method is better at controlling the style transfer magnitude using an input scalar knob. We report promising qualitative results for several attribute transfer tasks (sentiment transfer, simplification, gender neutralization, text anonymization) all without retraining the model. Finally, we find model evaluation to be difficult due to the lack of datasets and metrics for many languages. To facilitate future research we crowdsource formality annotations for 4000 sentence pairs in four Indic languages, and use this data to design our automatic evaluations.

12:00-12:15 (Wicklow Hall 1)
Dependency-based Mixture Language Models
Zhixian Yang and Xiaojun Wan
Various models have been proposed to incorporate knowledge of syntactic structures into neural language models. However, previous works have relied heavily on elaborate components for a specific language model, usually recurrent neural network (RNN), which makes themselves unwieldy in practice to fit into other neural language models, such as Transformer and GPT-2. In this paper, we introduce the Dependency-based Mixture Language Models. In detail, we first train neural language models with a novel dependency modeling objective to learn the probability distribution of future dependent tokens given context. We then formulate the next-token probability by mixing the previous dependency modeling probability distributions with self-attention. Extensive experiments and human evaluations show that our method can be easily and effectively applied to different neural language models while improving neural text generation on various tasks.

Interpretability and Analysis of Models for NLP 3

10:45-12:15 (The Liffey A)

10:45-11:00 (The Liffey A)
The Paradox of the Compositionality of Natural Language: A Neural Machine Translation Case Study
Verna Dankers, Elia Bruni and Dieuwke Hupkes
Obtaining human-like performance in NLP is often argued to require compositional generalisation. Whether neural networks exhibit this ability is usually studied by training models on highly compositional synthetic data. However, compositionality in natural language is much more complex than the rigid, arithmetic-like version such data adheres to, and artificial compositionality tests thus do not allow us to determine how neural models deal with more realistic forms of compositionality. In this work, we re-instantiate three compositionality tests from the literature and reformulate them for neural machine translation (NMT). Our results highlight that: i) unfavourably, models trained on more data are more compositional; ii) models are sometimes less compositional than expected, but sometimes more, exemplifying that different levels of compositionality are required, and models are not always able to modulate between them correctly; iii) some of the non-compositional behaviours are mistakes, whereas others reflect the natural variation in data. Apart from an empirical study, our work is a call to action: we should rethink the evaluation of compositionality in neural networks and develop benchmarks using real data to evaluate compositionality on natural language, where composing meaning is not as straightforward as doing the math.

11:00-11:15 (The Liffey A)
Low-Rank Softmax Can Have Unargmaxable Classes in Theory but Rarely in Practice
Andreas Grivas, Nikolay Bogoychev and Adam Lopez
Classifiers in natural language processing (NLP) often have a large number of output classes. For example, neural language models (LMs) and machine translation (MT) models both predict tokens from a vocabulary of thousands. The Softmax output layer of these models typically receives as input a dense feature representation, which has much lower dimensionality than the output. In theory, the result is some words may be impossible to be predicted via argmax, irrespective of input features, and empirically, there is evidence this happens in small language models (Demeter et al., 2020). In this paper we ask whether it can happen in practical large language models and translation models. To do so, we develop algorithms to detect such unargmaxable tokens in public models. We find that 13 out of 150 models do indeed have such tokens; however, they are very infrequent and unlikely to impact model quality. We release our algorithms and code to the public.

11:15-11:30 (The Liffey A)
Finding Structural Knowledge in Multimodal-BERT
Victor Siemen Janusz Milewski, Miryam de Lhoneux and Marie-Francine Moens
In this work, we investigate the knowledge learned in the embeddings of multimodal-BERT models. More specifically, we probe their capabilities of storing the grammatical structure of linguistic data and the structure learned over objects in visual data. To reach that goal, we first make the inherent structure of language and visuals explicit by a dependency parse of the sentences that describe the image and by the dependencies between the object regions in the image, respectively. We call this explicit visual structure the scene tree, that is based on the dependency tree of the language description. Extensive probing experiments show that the multimodal-BERT models do not encode these scene trees.

11:30-11:45 (The Liffey A)
Toward Interpretable Semantic Textual Similarity via Optimal Transport-based Contrastive Sentence Learning
Seonghyeon Lee, Dongha Lee, Seongbo Jang and Hwanjo Yu
Recently, finetuning a pretrained language model to capture the similarity between sentence embeddings has shown the state-of-the-art performance on the semantic textual similarity (STS) task. However, the absence of an interpretation method for the sentence similarity makes

it difficult to explain the model output. In this work, we explicitly describe the sentence distance as the weighted sum of contextualized token distances on the basis of a transportation problem, and then present the optimal transport-based distance measure, named RCMD; it identifies and leverages semantically-aligned token pairs. In the end, we propose CLRCMD, a contrastive learning framework that optimizes RCMD of sentence pairs, which enhances the quality of sentence similarity and their interpretation. Extensive experiments demonstrate that our learning framework outperforms other baselines on both STS and interpretable-STS benchmarks, indicating that it computes effective sentence similarity and also provides interpretation consistent with human judgement.

11:45-12:00 (The Liffey A)
Pretraining with Artificial Language: Studying Transferable Knowledge in Language Models
Ryokan Ri and Yoshimasa Tsuruoka
We investigate what kind of structural knowledge learned in neural network encoders is transferable to processing natural language. We design *artificial languages* with structural properties that mimic natural language, pretrain encoders on the data, and see how much performance the encoder exhibits on downstream tasks in natural language. Our experimental results show that pretraining with an artificial language with a nesting dependency structure provides some knowledge transferable to natural language. A follow-up probing analysis indicates that its success in the transfer is related to the amount of encoded contextual information and what is transferred is the knowledge of *position-aware context dependence* of language. Our results provide insights into how neural network encoders process human languages and the source of cross-lingual transferability of recent multilingual language models.

12:00-12:10 (The Liffey A)
On the Effect of Isotropy on VAE Representations of Text
Lan Zhang, Wray Buntine and Ehsan Shareghi
Injecting desired geometric properties into text representations has attracted a lot of attention. A property that has been argued for, due to its better utilisation of representation space, is isotropy. In parallel, VAEs have been successful in areas of NLP, but are known for their sub-optimal utilisation of the representation space. To address an aspect of this, we investigate the impact of injecting isotropy during training of VAEs. We achieve this by using an isotropic Gaussian posterior (IGP) instead of the ellipsoidal Gaussian posterior. We illustrate that IGP effectively encourages isotropy in the representations, inducing a more discriminative latent space. Compared to vanilla VAE, this translates into a much better classification performance, robustness to input perturbation, and generative behavior. Additionally, we offer insights about the representational properties encouraged by IGP.

NLP Applications 3

10:45-12:15 (Liffey Hall 2)

10:45-11:00 (Liffey Hall 2)
Clickbait Spoiling via Question Answering and Passage Retrieval
Matthias Hagen, Maik Fröbe, Artur Jurk and Martin Potthast
We introduce and study the task of clickbait spoiling: generating a short text that satisfies the curiosity induced by a clickbait post. Clickbait links to a web page and advertises its contents by arousing curiosity instead of providing an informative summary. Our contributions are approaches to classify the type of spoiler needed (i.e., a phrase or a passage), and to generate appropriate spoilers. A large-scale evaluation and error analysis on a new corpus of 5,000 manually spoiled clickbait posts—the Webis Clickbait Spoiling Corpus 2022—shows that our spoiler type classifier achieves an accuracy of 80%, while the question answering model DeBERTa-large outperforms all others in generating spoilers for both types.

11:00-11:15 (Liffey Hall 2)
From the Detection of Toxic Spans in Online Discussions to the Analysis of Toxic-to-Civil Transfer
John Pavlopoulos, Leo Laugier, Alexandros Xenos, Jeffrey S. Sorensen and Ion Androutsopoulos
We study the task of toxic spans detection, which concerns the detection of the spans that make a text toxic, when detecting such spans is possible. We introduce a dataset for this task, ToxicSpans, which we release publicly. By experimenting with several methods, we show that sequence labeling models perform best, but methods that add generic rationale extraction mechanisms on top of classifiers trained to predict if a post is toxic or not are also surprisingly promising. Finally, we use ToxicSpans and systems trained on it, to provide further analysis of state-of-the-art toxic to non-toxic transfer systems, as well as of human performance on that latter task. Our work highlights challenges in finer toxicity detection and mitigation.

11:15-11:30 (Liffey Hall 2)
On the Robustness of Offensive Language Classifiers
Jonathan Rusert, Zubair Shafiq and Padmini Srinivasan
Social media platforms are deploying machine learning based offensive language classification systems to combat hateful, racist, and other forms of offensive speech at scale. However, despite their real-world deployment, we do not yet comprehensively understand the extent to which offensive language classifiers are robust against adversarial attacks. Prior work in this space is limited to studying robustness of offensive language classifiers against primitive attacks such as misspellings and extraneous spaces. To address this gap, we systematically analyze the robustness of state-of-the-art offensive language classifiers against more crafty adversarial attacks that leverage greedy- and attention-based word selection and context-aware embeddings for word replacement. Our results on multiple datasets show that these crafty adversarial attacks can degrade the accuracy of offensive language classifiers by more than 50

11:30-11:45 (Liffey Hall 2)
Improving Generalizability in Implicitly Abusive Language Detection with Concept Activation Vectors
Isar Nejadgholi, Kathleen C. Fraser and Svetlana Kiritchenko
Robustness of machine learning models on ever-changing real-world data is critical, especially for applications affecting human well-being such as content moderation. New kinds of abusive language continually emerge in online discussions in response to current events (e.g., COVID-19), and the deployed abuse detection systems should be updated regularly to remain accurate. In this paper, we show that general abusive language classifiers tend to be fairly reliable in detecting out-of-domain explicitly abusive utterances but fail to detect new types of more subtle, implicit abuse. Next, we propose an interpretability technique, based on the Testing Concept Activation Vector (TCAV) method

from computer vision, to quantify the sensitivity of a trained model to the human-defined concepts of explicit and implicit abusive language, and use that to explain the generalizability of the model on new data, in this case, COVID-related anti-Asian hate speech. Extending this technique, we introduce a novel metric, Degree of Explicitness, for a single instance and show that the new metric is beneficial in suggesting out-of-domain unlabeled examples to effectively enrich the training data with informative, implicitly abusive texts.

11:45-12:00 (Liffey Hall 2)
FairLex: A Multilingual Benchmark for Evaluating Fairness in Legal Text Processing
Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Felix Schwemer and Anders Søgaard
We present a benchmark suite of four datasets for evaluating the fairness of pre-trained language models and the techniques used to fine-tune them for downstream tasks. Our benchmarks cover four jurisdictions (European Council, USA, Switzerland, and China), five languages (English, German, French, Italian and Chinese) and fairness across five attributes (gender, age, region, language, and legal area). In our experiments, we evaluate pre-trained language models using several group-robust fine-tuning techniques and show that performance group disparities are vibrant in many cases, while none of these techniques guarantee fairness, nor consistently mitigate group disparities. Furthermore, we provide a quantitative and qualitative analysis of our results, highlighting open challenges in the development of robustness methods in legal NLP.

12:00-12:15 (Liffey Hall 2)
[TACL] **A Neighbourhood Framework for Resource-Lean Content Flagging**
Momchil Hardalov, Sheikh Sarwar, Dimitrina Zlatkova, Yoan Dinkov, Isabelle Augenstein and Preslav Nakov

Semantics 3

10:45-12:15 (Wicklow Hall 2a)

10:45-11:00 (Wicklow Hall 2a)
Generating Data to Mitigate Spurious Correlations in Natural Language Inference Datasets
Yuxiang Wu, Matt Gardner, Pontus Stenetorp and Pradeep Dasigi
Natural language processing models often exploit spurious correlations between task-independent features and labels in datasets to perform well only within the distributions they are trained on, while not generalising to different task distributions. We propose to tackle this problem by generating a debiased version of a dataset, which can then be used to train a debiased, off-the-shelf model, by simply replacing its training data. Our approach consists of 1) a method for training data generators to generate high-quality, label-consistent data samples; and 2) a filtering mechanism for removing data points that contribute to spurious correlations, measured in terms of z-statistics. We generate debiased versions of the SNLI and MNLI datasets, and we evaluate on a large suite of debiased, out-of-distribution, and adversarial test sets. Results show that models trained on our debiased datasets generalise better than those trained on the original datasets in all settings. On the majority of the datasets, our method outperforms or performs comparably to previous state-of-the-art debiasing strategies, and when combined with an orthogonal technique, product-of-experts, it improves further and outperforms previous best results of SNLI-hard and MNLI-hard.

11:00-11:15 (Wicklow Hall 2a)
Modeling Syntactic-Semantic Dependency Correlations in Semantic Role Labeling Using Mixture Models
Junjie Chen, Xiangheng He and Yusuke Miyao
In this paper, we propose a mixture model-based end-to-end method to model the syntactic-semantic dependency correlation in Semantic Role Labeling (SRL). Semantic dependencies in SRL are modeled as a distribution over semantic dependency labels conditioned on a predicate and an argument word. The semantic label distribution varies depending on Shortest Syntactic Dependency Path (SSDP) hop patterns. We target the variation of semantic label distributions using a mixture model, separately estimating semantic label distributions for different hop patterns and probabilistically clustering hop patterns with similar semantic label distributions. Experiments show that the proposed method successfully learns a cluster assignment reflecting the variation of semantic label distributions. Modeling the variation improves performance in predicting short distance semantic dependencies, in addition to the improvement on long distance semantic dependencies that previous syntax-aware methods have achieved. The proposed method achieves a small but statistically significant improvement over baseline methods in English, German, and Spanish and obtains competitive performance with state-of-the-art methods in English.

11:15-11:30 (Wicklow Hall 2a)
[TACL] **Is My Model Using The Right Evidence? Systematic Probes for Examining Evidence-Based Tabular Reasoning**
Vivek Gupta, Riyaz Bhat, Atreya Ghosal, Manish Shrivastava, Maneesh Singh and Vivek Srikumar

11:30-11:45 (Wicklow Hall 2a)
Right for the Right Reason: Evidence Extraction for Trustworthy Tabular Reasoning
Vivek Gupta, Shuo Zhang, Alakananda Vempala, Yujie He, Temma Choji and Vivek Srikumar
When pre-trained contextualized embedding-based models developed for unstructured data are adapted for structured tabular data, they perform admirably. However, recent probing studies show that these models use spurious correlations, and often predict inference labels by focusing on false evidence or ignoring it altogether. To study this issue, we introduce the task of Trustworthy Tabular Reasoning, where a model needs to extract evidence to be used for reasoning, in addition to predicting the label. As a case study, we propose a two-stage sequential prediction approach, which includes an evidence extraction and an inference stage. First, we crowdsource evidence row labels and develop several unsupervised and supervised evidence extraction strategies for InfoTabS, a tabular NLI benchmark. Our evidence extraction strategy outperforms earlier baselines. On the downstream tabular inference task, using only the automatically extracted evidence as the premise, our approach outperforms prior benchmarks.

11:45-12:00 (Wicklow Hall 2a)
WatClaimCheck: A new Dataset for Claim Entailment and Inference
Kashif Khan, Ruizhe Wang and Pascal Poupart
We contribute a new dataset for the task of automated fact checking and an evaluation of state of the art algorithms. The dataset includes

claims (from speeches, interviews, social media and news articles), review articles published by professional fact checkers and premise articles used by those professional fact checkers to support their review and verify the veracity of the claims. An important challenge in the use of premise articles is the identification of relevant passages that will help to infer the veracity of a claim. We show that transferring a dense passage retrieval model trained with review articles improves the retrieval quality of passages in premise articles. We report results for the prediction of claim veracity by inference from premise articles.

12:00-12:15 (Wicklow Hall 2a)
EPT-X: An Expression-Pointer Transformer model that generates eXplanations for numbers
Bugeun Kim, Kyung Seo Ki, Sangkyu Rhim and Gahgene Gweon
In this paper, we propose a neural model EPT-X (Expression-Pointer Transformer with Explanations), which utilizes natural language explanations to solve an algebraic word problem. To enhance the explainability of the encoding process of a neural model, EPT-X adopts the concepts of plausibility and faithfulness which are drawn from math word problem solving strategies by humans. A plausible explanation is one that includes contextual information for the numbers and variables that appear in a given math word problem. A faithful explanation is one that accurately represents the reasoning process behind the model's solution equation. The EPT-X model yields an average baseline performance of 69.59

Linguistic Theories, Cognitive Modeling and Psycholinguistics

10:45-12:15 (Wicklow Hall 2b)

10:45-11:00 (Wicklow Hall 2b)
Do Transformer Models Show Similar Attention Patterns to Task-Specific Human Gaze?
Oliver Eberle, Stephanie Brandl, Jonas Pilot and Anders Søgaard
Learned self-attention functions in state-of-the-art NLP models often correlate with human attention. We investigate whether self-attention in large-scale pre-trained language models is as predictive of human eye fixation patterns during task-reading as classical cognitive models of human attention. We compare attention functions across two task-specific reading datasets for sentiment analysis and relation extraction. We find the predictiveness of large-scale pre-trained self-attention for human attention depends on 'what is in the tail', e.g., the syntactic nature of rare contexts. Further, we observe that task-specific fine-tuning does not increase the correlation with human task-specific reading. Through an input reduction experiment we give complementary insights on the sparsity and fidelity trade-off, showing that lower-entropy attention vectors are more faithful.

11:00-11:15 (Wicklow Hall 2b)
Speaker Information Can Guide Models to Better Inductive Biases: A Case Study On Predicting Code-Switching
Alissa Ostapenko, Shuly Wintner, Melinda Fricke and Yulia Tsvetkov
Natural language processing (NLP) models trained on people-generated data can be unreliable because, without any constraints, they can learn from spurious correlations that are not relevant to the task. We hypothesize that enriching models with speaker information in a controlled, educated way can guide them to pick up on relevant inductive biases. For the speaker-driven task of predicting code-switching points in English–Spanish bilingual dialogues, we show that adding sociolinguistically-grounded speaker features as prepended prompts significantly improves accuracy. We find that by adding influential phrases to the input, speaker-informed models learn useful and explainable linguistic information. To our knowledge, we are the first to incorporate speaker characteristics in a neural model for code-switching, and more generally, take a step towards developing transparent, personalized models that use speaker information in a controlled way.

11:15-11:30 (Wicklow Hall 2b)
[TACL] **A Biologically Plausible Parser**
Daniel Mitropolsky, Michael Collins and Christos Papadimitriou

11:30-11:45 (Wicklow Hall 2b)
Characterizing Idioms: Conventionality and Contingency
Michaela Socolof, Jackie CK Cheung, Michael Wagner and Timothy J. O'Donnell
Idioms are unlike most phrases in two important ways. First, words in an idiom have non-canonical meanings. Second, the non-canonical meanings of words in an idiom are contingent on the presence of other words in the idiom. Linguistic theories differ on whether these properties depend on one another, as well as whether special theoretical machinery is needed to accommodate idioms. We define two measures that correspond to the properties above, and we show that idioms fall at the expected intersection of the two dimensions, but that the dimensions themselves are not correlated. Our results suggest that introducing special machinery to handle idioms may not be warranted.

11:45-12:00 (Wicklow Hall 2b)
[TACL] **Quantifying Cognitive Factors in Lexical Decline**
Ella Rabinovich, David Francis, Samir Farhan, David Mortensen and Suzanne Stevenson

12:00-12:10 (Wicklow Hall 2b)
Analyzing Wrap-Up Effects through an Information-Theoretic Lens
Clara Isabel Meister, Tiago Pimentel, Thomas Hikaru Clark, Ryan D Cotterell and Roger P. Levy
Numerous analyses of reading time (RT) data have been undertaken in the effort to learn more about the internal processes that occur during reading comprehension. However, data measured on words at the end of a sentence–or even clause–is often omitted due to the confounding factors introduced by so-called "wrap-up effects," which manifests as a skewed distribution of RTs for these words. Consequently, the understanding of the cognitive processes that might be involved in these effects is limited. In this work, we attempt to learn more about these processes by looking for the existence–or absence–of a link between wrap-up effects and information theoretic quantities, such as word and context information content. We find that the information distribution of prior context is often predictive of sentence- and clause-final RTs (while not of sentence-medial RTs), which lends support to several prior hypotheses about the processes involved in wrap-up effects.

Poster Session 6: Resources and Evaluation

10:45-12:15 (Forum)

10:45-12:15 (Forum)
#1 Identifying Moments of Change from Longitudinal User Text
Adam Tsakalidis, Jenny Chim, Anthony Hills, Maria Liakata, Federico Nanni and Jiayu Song
Identifying changes in individuals' behaviour and mood, as observed via content shared on online platforms, is increasingly gaining importance. Most research to-date on this topic focuses on either: (a) identifying individuals at risk or with a certain mental health condition given a batch of posts or (b) providing equivalent labels at the post level. A disadvantage of such work is the lack of a strong temporal component and the inability to make longitudinal assessments following an individual's trajectory and allowing timely interventions. Here we define a new task, that of identifying moments of change in individuals on the basis of their shared content online. The changes we consider are sudden shifts in mood (switches) or gradual mood progression (escalations). We have created detailed guidelines for capturing moments of change and a corpus of 500 manually annotated user timelines (18.7K posts). We have developed a variety of baseline models drawing inspiration from related tasks and show that the best performance is obtained through context aware sequential modelling. We also introduce new metrics for capturing rare events in temporal windows.

10:45-12:15 (Forum)
#2 QuoteR: A Benchmark of Quote Recommendation for Writing
Fanchao Qi, Zhili Cheng, Zhiyuan Liu, Maosong Sun, Yanhui Yang and Jing Yi
It is very common to use quotations (quotes) to make our writings more elegant or convincing. To help people find appropriate quotes efficiently, the task of quote recommendation is presented, aiming to recommend quotes that fit the current context of writing. There have been various quote recommendation approaches, but they are evaluated on different unpublished datasets. To facilitate the research on this task, we build a large and fully open quote recommendation dataset called QuoteR, which comprises three parts including English, standard Chinese and classical Chinese. Any part of it is larger than previous unpublished counterparts. We conduct an extensive evaluation of existing quote recommendation methods on QuoteR. Furthermore, we propose a new quote recommendation model that significantly outperforms previous methods on all three parts of QuoteR. All the code and data of this paper can be obtained at https://github.com/thunlp/QuoteR.

10:45-12:15 (Forum)
#3 k-Rater Reliability: The Correct Unit of Reliability for Aggregated Human Annotations
Ka Wong and Praveen Paritosh
Since the inception of crowdsourcing, aggregation has been a common strategy for dealing with unreliable data. Aggregate ratings are more reliable than individual ones. However, many Natural Language Processing (NLP) applications that rely on aggregate ratings only report the reliability of individual ratings, which is the incorrect unit of analysis. In these instances, the data reliability is under-reported, and a proposed k-rater reliability (kRR) should be used as the correct data reliability for aggregated datasets. It is a multi-rater generalization of inter-rater reliability (IRR). We conducted two replications of the WordSim-353 benchmark, and present empirical, analytical, and bootstrap-based methods for computing kRR on WordSim-353. These methods produce very similar results. We hope this discussion will nudge researchers to report kRR in addition to IRR.

10:45-12:15 (Forum)
#4 FaVIQ: FAct Verification from Information-seeking Questions
Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer and Hannaneh Hajishirzi
Despite significant interest in developing general purpose fact checking models, it is challenging to construct a large-scale fact verification dataset with realistic real-world claims. Existing claims are either authored by crowdworkers, thereby introducing subtle biases that are difficult to control for, or manually verified by professional fact checkers, causing them to be expensive and limited in scale. In this paper, we construct a large-scale challenging fact verification dataset called FAVIQ, consisting of 188k claims derived from an existing corpus of ambiguous information-seeking questions. The ambiguities in the questions enable automatically constructing true and false claims that reflect user confusions (e.g., the year of the movie being filmed vs. being released). Claims in FAVIQ are verified to be natural, contain little lexical bias, and require a complete understanding of the evidence for verification. Our experiments show that the state-of-the-art models are far from solving our new task. Moreover, training on our data helps in professional fact-checking, outperforming models trained on the widely used dataset FEVER or in-domain data by up to 17

10:45-12:15 (Forum)
#5 RELiC: Retrieving Evidence for Literary Claims
Katherine Thai, Yapei Chang, Kalpesh Krishna and Mohit Iyyer
Humanities scholars commonly provide evidence for claims that they make about a work of literature (e.g., a novel) in the form of quotations from the work. We collect a large-scale dataset (RELiC) of 78K literary quotations and surrounding critical analysis and use it to formulate the novel task of literary evidence retrieval, in which models are given an excerpt of literary analysis surrounding a masked quotation and asked to retrieve the quoted passage from the set of all passages in the work. Solving this retrieval task requires a deep understanding of complex literary and linguistic phenomena, which proves challenging to methods that overwhelmingly rely on lexical and semantic similarity matching. We implement a RoBERTa-based dense passage retriever for this task that outperforms existing pretrained information retrieval baselines; however, experiments and analysis by human domain experts indicate that there is substantial room for improvement.

10:45-12:15 (Forum)
#6 IAM: A Comprehensive and Large-Scale Dataset for Integrated Argument Mining Tasks
Liying Cheng, Lidong Bing, Ruidan He, Qian Yu, Yan Zhang and Luo Si
Traditionally, a debate usually requires a manual preparation process, including reading plenty of articles, selecting the claims, identifying the stances of the claims, seeking the evidence for the claims, etc. As the AI debate attracts more attention these years, it is worth exploring the methods to automate the tedious process involved in the debating system. In this work, we introduce a comprehensive and large dataset named IAM, which can be applied to a series of argument mining tasks, including claim extraction, stance classification, evidence extraction, etc. Our dataset is collected from over 1k articles related to 123 topics. Near 70k sentences in the dataset are fully annotated based on their argument properties (e.g., claims, stances, evidence, etc.). We further propose two new integrated argument mining tasks associated with the debate preparation process: (1) claim extraction with stance classification (CESC) and (2) claim-evidence pair extraction (CEPE). We adopt a pipeline approach and an end-to-end method for each integrated task separately. Promising experimental results are reported to show the

values and challenges of our proposed tasks, and motivate future research on argument mining.

10:45-12:15 (Forum)
#7 AlephBERT: Language Model Pre-training and Evaluation from Sub-Word to Sentence Level
Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Shaked Greenfeld and Reut Tsarfaty
Large Pre-trained Language Models (PLMs) have become ubiquitous in the development of language understanding technology and lie at the heart of many artificial intelligence advances. While advances reported for English using PLMs are unprecedented, reported advances using PLMs for Hebrew are few and far between. The problem is twofold. First, so far, Hebrew resources for training large language models are not of the same magnitude as their English counterparts. Second, most benchmarks available to evaluate progress in Hebrew NLP require morphological boundaries which are not available in the output of standard PLMs. In this work we remedy both aspects. We present AlephBERT, a large PLM for Modern Hebrew, trained on larger vocabulary and a larger dataset than any Hebrew PLM before. Moreover, we introduce a novel neural architecture that recovers the morphological segments encoded in contextualized embedding vectors. Based on this new morphological component we offer an evaluation suite consisting of multiple tasks and benchmarks that cover sentence-level, word-level and sub-word level analyses. On all tasks, AlephBERT obtains state-of-the-art results beyond contemporary Hebrew baselines. We make our AlephBERT model, the morphological extraction model, and the Hebrew evaluation suite publicly available, for evaluating future Hebrew PLMs.

10:45-12:15 (Forum)
[CL] **#8 Annotation Curricula to Implicitly Train Non-Expert Annotators**
Ji-Ung Lee, Jan-Christoph Klie, Iryna Gurevych

10:45-12:15 (Forum)
#9 Thai Nested Named Entity Recognition Corpus
Weerayut Buaphet, Can Udomcharoenchaikit, Peerat Limkonchotiwat, Attapol Rutherford and Sarana Nutanong
This paper presents the first Thai Nested Named Entity Recognition (N-NER) dataset. Thai N-NER consists of 264,798 mentions, 104 classes, and a maximum depth of 8 layers obtained from 4,894 documents in the domains of news articles and restaurant reviews. Our work, to the best of our knowledge, presents the largest non-English N-NER dataset and the first non-English one with fine-grained classes. To understand the new challenges our proposed dataset brings to the field, we conduct an experimental study on (i) cutting edge N-NER models with the state-of-the-art accuracy in English and (ii) baseline methods based on well-known language model architectures. From the experimental results, we obtained two key findings. First, all models produced poor F1 scores in the tail region of the class distribution. There is little or no performance improvement provided by these models with respect to the baseline methods with our Thai dataset. These findings suggest that further investigation is required to make a multilingual N-NER solution that works well across different languages.

10:45-12:15 (Forum)
[TACL] **#10 Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics**
Paula Czarnowska, Yogarshi Vyas and Kashif Shah

10:45-12:15 (Forum)
[TACL] **#11 Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations**
Aida Mostafazadeh Davani, Mark Díaz and Vinodkumar Prabhakaran

10:45-12:15 (Forum)
#12 Quantified Reproducibility Assessment of NLP Results
Anya Belz, Maja Popovic and Simon Mille
This paper describes and tests a method for carrying out quantified reproducibility assessment (QRA) that is based on concepts and definitions from metrology. QRA produces a single score estimating the degree of reproducibility of a given system and evaluation measure, on the basis of the scores from, and differences between, different reproductions. We test QRA on 18 different system and evaluation measure combinations (involving diverse NLP tasks and types of evaluation), for each of which we have the original results and one to seven reproduction results. The proposed QRA method produces degree-of-reproducibility scores that are comparable across multiple reproductions not only of the same, but also of different, original studies. We find that the proposed method facilitates insights into causes of variation between reproductions, and as a result, allows conclusions to be drawn about what aspects of system and/or evaluation design need to be changed in order to improve reproducibility.

10:45-12:15 (Forum)
#13 A Taxonomy of Empathetic Questions in Social Dialogs
Ekaterina Svikhnushina, Iuliana Voinea, Anuradha Welivita and Pearl Pu
Effective question-asking is a crucial component of a successful conversational chatbot. It could help the bots manifest empathy and render the interaction more engaging by demonstrating attention to the speaker's emotions. However, current dialog generation approaches do not model this subtle emotion regulation technique due to the lack of a taxonomy of questions and their purpose in social chitchat. To address this gap, we have developed an empathetic question taxonomy (EQT), with special attention paid to questions' ability to capture communicative acts and their emotion-regulation intents. We further design a crowd-sourcing task to annotate a large subset of the EmpatheticDialogues dataset with the established labels. We use the crowd-annotated data to develop automatic labeling tools and produce labels for the whole dataset. Finally, we employ information visualization techniques to summarize co-occurrences of question acts and intents and their role in regulating interlocutor's emotion. These results reveal important question-asking strategies in social dialogs. The EQT classification scheme can facilitate computational analysis of questions in datasets. More importantly, it can inform future efforts in empathetic question generation using neural or hybrid methods.

10:45-12:15 (Forum)
#14 Detecting Unassimilated Borrowings in Spanish: An Annotated Corpus and Approaches to Modeling
Elena Álvarez-Mellado and Constantine Lignos
This work presents a new resource for borrowing identification and analyzes the performance and errors of several models on this task. We

introduce a new annotated corpus of Spanish newswire rich in unassimilated lexical borrowings—words from one language that are introduced into another without orthographic adaptation—and use it to evaluate how several sequence labeling models (CRF, BiLSTM-CRF, and Transformer-based models) perform. The corpus contains 370,000 tokens and is larger, more borrowing-dense, OOV-rich, and topic-varied than previous corpora available for this task. Our results show that a BiLSTM-CRF model fed with subword embeddings along with either Transformer-based embeddings pretrained on codeswitched data or a combination of contextualized word embeddings outperforms results obtained by a multilingual BERT-based model.

10:45-12:15 (Forum)
#15 DiBiMT: A Novel Benchmark for Measuring Word Sense Disambiguation Biases in Machine Translation
Niccolò Campolungo, Federico Martelli, Francesco Saina and Roberto Navigli
Lexical ambiguity poses one of the greatest challenges in the field of Machine Translation. Over the last few decades, multiple efforts have been undertaken to investigate incorrect translations caused by the polysemous nature of words. Within this body of research, some studies have posited that models pick up semantic biases existing in the training data, thus producing translation errors. In this paper, we present DiBiMT, the first entirely manually-curated evaluation benchmark which enables an extensive study of semantic biases in Machine Translation of nominal and verbal words in five different language combinations, namely, English and one or other of the following languages: Chinese, German, Italian, Russian and Spanish. Furthermore, we test state-of-the-art Machine Translation systems, both commercial and non-commercial ones, against our new test bed and provide a thorough statistical and linguistic analysis of the results. We release DiBiMT at https://nlp.uniroma1.it/dibimt as a closed benchmark with a public leaderboard.

10:45-12:15 (Forum)
#16 BenchIE: A Framework for Multi-Faceted Fact-Based Open Information Extraction Evaluation
Kiril Gashteovski, Mingying Yu, Bhushan Kotnis, Carolin Lawrence, Mathias Niepert and Goran Glavaš
Intrinsic evaluations of OIE systems are carried out either manually—with human evaluators judging the correctness of extractions—or automatically, on standardized benchmarks. The latter, while much more cost-effective, is less reliable, primarily because of the incompleteness of the existing OIE benchmarks: the ground truth extractions do not include all acceptable variants of the same fact, leading to unreliable assessment of the models' performance. Moreover, the existing OIE benchmarks are available for English only. In this work, we introduce BenchIE: a benchmark and evaluation framework for comprehensive evaluation of OIE systems for English, Chinese, and German. In contrast to existing OIE benchmarks, BenchIE is fact-based, i.e., it takes into account informational equivalence of extractions: our gold standard consists of *fact synsets*, clusters in which we exhaustively list all acceptable surface forms of the same fact. Moreover, having in mind common downstream applications for OIE, we make BenchIE multi-faceted; i.e., we create benchmark variants that focus on different facets of OIE evaluation, e.g., compactness or minimality of extractions. We benchmark several state-of-the-art OIE systems using BenchIE and demonstrate that these systems are significantly less effective than indicated by existing OIE benchmarks. We make BenchIE (data and evaluation code) publicly available.

10:45-12:15 (Forum)
#17 RoMe: A Robust Metric for Evaluating Natural Language Generation
Md Rashad Al Hasan Rony, Liubov Kovriguina, Debanjan Chaudhuri, Ricardo Usbeck and Jens Lehmann
Evaluating Natural Language Generation (NLG) systems is a challenging task. Firstly, the metric should ensure that the generated hypothesis reflects the reference's semantics. Secondly, it should consider the grammatical quality of the generated sentence. Thirdly, it should be robust enough to handle various surface forms of the generated sentence. Thus, an effective evaluation metric has to be multifaceted. In this paper, we propose an automatic evaluation metric incorporating several core aspects of natural language understanding (language competence, syntactic and semantic variation). Our proposed metric, RoMe, is trained on language features such as semantic similarity combined with tree edit distance and grammatical acceptability, using a self-supervised neural network to assess the overall quality of the generated sentence. Moreover, we perform an extensive robustness analysis of the state-of-the-art methods and RoMe. Empirical results suggest that RoMe has a stronger correlation to human judgment over state-of-the-art metrics in evaluating system-generated sentences across several NLG tasks.

10:45-12:15 (Forum)
#18 PriMock57: A Dataset Of Primary Care Mock Consultations
Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac and Aleksandar Savkov
Recent advances in Automatic Speech Recognition (ASR) have made it possible to reliably produce automatic transcripts of clinician-patient conversations. However, access to clinical datasets is heavily restricted due to patient privacy, thus slowing down normal research practices. We detail the development of a public access, high quality dataset comprising of 57 mocked primary care consultations, including audio recordings, their manual utterance-level transcriptions, and the associated consultation notes. Our work illustrates how the dataset can be used as a benchmark for conversational medical ASR as well as consultation note generation from transcripts.

10:45-12:15 (Forum)
#19 A Statutory Article Retrieval Dataset in French
Antoine Louis and Gerasimos Spanakis
Statutory article retrieval is the task of automatically retrieving law articles relevant to a legal question. While recent advances in natural language processing have sparked considerable interest in many legal tasks, statutory article retrieval remains primarily untouched due to the scarcity of large-scale and high-quality annotated datasets. To address this bottleneck, we introduce the Belgian Statutory Article Retrieval Dataset (BSARD), which consists of 1,100+ French native legal questions labeled by experienced jurists with relevant articles from a corpus of 22,600+ Belgian law articles. Using BSARD, we benchmark several state-of-the-art retrieval approaches, including lexical and dense architectures, both in zero-shot and supervised setups. We find that fine-tuned dense retrieval models significantly outperform other systems. Our best performing baseline achieves 74.8

10:45-12:15 (Forum)
#20 CoDA21: Evaluating Language Understanding Capabilities of NLP Models With Context-Definition Alignment
Lütfi Kerem Senel, Timo Schick and Hinrich Schuetze
Pretrained language models (PLMs) have achieved superhuman performance on many benchmarks, creating a need for harder tasks. We introduce CoDA21 (Context Definition Alignment), a challenging benchmark that measures natural language understanding (NLU) capabilities of PLMs: Given a definition and a context each for k words, but not the words themselves, the task is to align the k definitions with the k contexts. CoDA21 requires a deep understanding of contexts and definitions, including complex inference and world knowledge. We find that there is a large gap between human and PLM performance, suggesting that CoDA21 measures an aspect of NLU that is not sufficiently

covered in existing benchmarks.

10:45-12:15 (Forum)
#21 Active Evaluation: Efficient NLG Evaluation with Few Pairwise Comparisons
Akash Kumar Mohankumar and Mitesh M Khapra
Recent studies have shown the advantages of evaluating NLG systems using pairwise comparisons as opposed to direct assessment. Given k systems, a naive approach for identifying the top-ranked system would be to uniformly obtain pairwise comparisons from all $\binom{k}{2}$ pairs of systems. However, this can be very expensive as the number of human annotations required would grow quadratically with k. In this work, we introduce Active Evaluation, a framework to efficiently identify the top-ranked system by actively choosing system pairs for comparison using dueling bandit algorithms. We perform extensive experiments with 13 dueling bandits algorithms on 13 NLG evaluation datasets spanning 5 tasks and show that the number of human annotations can be reduced by 80

10:45-12:15 (Forum)
#22 Human Evaluation and Correlation with Automatic Metrics in Consultation Note Generation
Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz and Aleksandar Savkov
In recent years, machine learning models have rapidly become better at generating clinical consultation notes; yet, there is little work on how to properly evaluate the generated consultation notes to understand the impact they may have on both the clinician using them and the patient's clinical safety. To address this we present an extensive human evaluation study of consultation notes where 5 clinicians (i) listen to 57 mock consultations, (ii) write their own notes, (iii) post-edit a number of automatically generated notes, and (iv) extract all the errors, both quantitative and qualitative. We then carry out a correlation study with 18 automatic quality metrics and the human judgements. We find that a simple, character-based Levenshtein distance metric performs on par if not better than common model-based metrics like BertScore. All our findings and annotations are open-sourced.

10:45-12:15 (Forum)
#23 NumGLUE: A Suite of Fundamental yet Challenging Mathematical Reasoning Tasks
Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Singh Sachdeva, Peter Clark, Chitta Baral and Ashwin Kalyan
Given the ubiquitous nature of numbers in text, reasoning with numbers to perform simple calculations is an important skill of AI systems. While many datasets and models have been developed to this end, state-of-the-art AI systems are brittle; failing to perform the underlying mathematical reasoning when they appear in a slightly different scenario. Drawing inspiration from GLUE that was proposed in the context of natural language understanding, we propose NumGLUE, a multi-task benchmark that evaluates the performance of AI systems on eight different tasks, that at their core require simple arithmetic understanding. We show that this benchmark is far from being solved with neural models including state-of-the-art large-scale language models performing significantly worse than humans (lower by 46.4

10:45-12:15 (Forum)
#24 Sense Embeddings are also Biased – Evaluating Social Biases in Static and Contextualised Sense Embeddings
Yi Zhou, Masahiro Kaneko and Danushka Bollegala
Sense embedding learning methods learn different embeddings for the different senses of an ambiguous word. One sense of an ambiguous word might be socially biased while its other senses remain unbiased. In comparison to the numerous prior work evaluating the social biases in pretrained word embeddings, the biases in sense embeddings have been relatively understudied. We create a benchmark dataset for evaluating the social biases in sense embeddings and propose novel sense-specific bias evaluation measures. We conduct an extensive evaluation of multiple static and contextualised sense embeddings for various types of social biases using the proposed measures. Our experimental results show that even in cases where no biases are found at word-level, there still exist worrying levels of social biases at sense-level, which are often ignored by the word-level bias evaluation measures.

10:45-12:15 (Forum)
#25 ILDAE: Instance-Level Difficulty Analysis of Evaluation Data
Neeraj Varshney, Swaroop Mishra and Chitta Baral
Knowledge of difficulty level of questions helps a teacher in several ways, such as estimating students' potential quickly by asking carefully selected questions and improving quality of examination by modifying trivial and hard questions. Can we extract such benefits of instance difficulty in Natural Language Processing? To this end, we conduct Instance-Level Difficulty Analysis of Evaluation data (ILDAE) in a large-scale setup of 23 datasets and demonstrate its five novel applications: 1) conducting efficient-yet-accurate evaluations with fewer instances saving computational cost and time, 2) improving quality of existing evaluation datasets by repairing erroneous and trivial instances, 3) selecting the best model based on application requirements, 4) analyzing dataset characteristics for guiding future data creation, 5) estimating Out-of-Domain performance reliably. Comprehensive experiments for these applications lead to several interesting results, such as evaluation using just 5

10:45-12:15 (Forum)
#26 Evaluating Extreme Hierarchical Multi-label Classification
Enrique Amigo and Agustín D. Delgado
Several natural language processing (NLP) tasks are defined as a classification problem in its most complex form: Multi-label Hierarchical Extreme classification, in which items may be associated with multiple classes from a set of thousands of possible classes organized in a hierarchy and with a highly unbalanced distribution both in terms of class frequency and the number of labels per item. We analyze the state of the art of evaluation metrics based on a set of formal properties and we define an information theoretic based metric inspired by the Information Contrast Model (ICM). Experiments on synthetic data and a case study on real data show the suitability of the ICM for such scenarios.

10:45-12:15 (Forum)
#27 Understanding Iterative Revision from Human-Written Text
Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez and Dongyeop Kang
Writing is, by nature, a strategic, adaptive, and, more importantly, an iterative process. A crucial part of writing is editing and revising the text. Previous works on text revision have focused on defining edit intention taxonomies within a single domain or developing computational models with a single level of edit granularity, such as sentence-level edits, which differ from human's revision cycles. This work describes IteraTeR: the first large-scale, multi-domain, edit-intention annotated corpus of iteratively revised text. In particular, IteraTeR is collected based on a new framework to comprehensively model the iterative text revisions that generalizes to a variety of domains, edit intentions,

revision depths, and granularities. When we incorporate our annotated edit intentions, both generative and action-based text revision models significantly improve automatic evaluations. Through our work, we better understand the text revision process, making vital connections between edit intentions and writing quality, enabling the creation of diverse corpora to support computational modeling of iterative text revisions.

10:45-12:15 (Forum)
#28 TwittIrish: A Universal Dependencies Treebank of Tweets in Modern Irish
Lauren Cassidy, Teresa Lynn, James Barry and Jennifer Foster
Modern Irish is a minority language lacking sufficient computational resources for the task of accurate automatic syntactic parsing of user-generated content such as tweets. Although language technology for the Irish language has been developing in recent years, these tools tend to perform poorly on user-generated content. As with other languages, the linguistic style observed in Irish tweets differs, in terms of orthography, lexicon, and syntax, from that of standard texts more commonly used for the development of language models and parsers. We release the first Universal Dependencies treebank of Irish tweets, facilitating natural language processing of user-generated content in Irish. In this paper, we explore the differences between Irish tweets and standard Irish text, and the challenges associated with dependency parsing of Irish tweets. We describe our bootstrapping method of treebank development and report on preliminary parsing experiments.

10:45-12:15 (Forum)
#29 Analyzing Dynamic Adversarial Training Data in the Limit
Eric Wallace, Adina Williams, Robin Jia and Douwe Kiela
To create models that are robust across a wide range of test inputs, training datasets should include diverse examples that span numerous phenomena. Dynamic adversarial data collection (DADC), where annotators craft examples that challenge continually improving models, holds promise as an approach for generating such diverse training sets. Prior work has shown that running DADC over 1-3 rounds can help models fix some error types, but it does not necessarily lead to better generalization beyond adversarial test data. We argue that running DADC over many rounds maximizes its training-time benefits, as the different rounds can together cover many of the task-relevant phenomena. We present the first study of longer-term DADC, where we collect 20 rounds of NLI examples for a small set of premise paragraphs, with both adversarial and non-adversarial approaches. Models trained on DADC examples make 26% fewer errors on our expert-curated test set compared to models trained on non-adversarial data. Our analysis shows that DADC yields examples that are more difficult, more lexically and syntactically diverse, and contain fewer annotation artifacts compared to non-adversarial examples.

10:45-12:15 (Forum)
#30 Mukayese: Turkish NLP Strikes Back
Ali Safaya, Emirhan Kurtuluş, Arda Goktogan and Deniz Yuret
Having sufficient resources for language X lifts it from the under-resourced languages class, but not necessarily from the under-researched class. In this paper, we address the problem of the absence of organized benchmarks in the Turkish language. We demonstrate that languages such as Turkish are left behind the state-of-the-art in NLP applications. As a solution, we present Mukayese, a set of NLP benchmarks for the Turkish language that contains several NLP tasks. We work on one or more datasets for each benchmark and present two or more baselines. Moreover, we present four new benchmarking datasets in Turkish for language modeling, sentence segmentation, and spell checking. All datasets and baselines are available under: https://github.com/alisafaya/mukayese

10:45-12:15 (Forum)
#31 Benchmarking Answer Verification Methods for Question Answering-Based Summarization Evaluation Metrics
Daniel Deutsch and Dan Roth
Question answering-based summarization evaluation metrics must automatically determine whether the QA model's prediction is correct or not, a task known as answer verification. In this work, we benchmark the lexical answer verification methods which have been used by current QA-based metrics as well as two more sophisticated text comparison methods, BERTScore and LERC. We find that LERC out-performs the other methods in some settings while remaining statistically indistinguishable from lexical overlap in others. However, our experiments reveal that improved verification performance does not necessarily translate to overall QA-based metric quality: In some scenarios, using a worse verification method — or using none at all — has comparable performance to using the best verification method, a result that we attribute to properties of the datasets.

10:45-12:15 (Forum)
#32 MIMICause: Representation and automatic extraction of causal relation types from clinical notes
Vivek Khetan, Md Imbesat Hassan Rizvi, Jessica Huber, Paige Bartusiak, Bogdan Eugen Sacaleanu and Andrew Fano
Understanding causal narratives communicated in clinical notes can help make strides towards personalized healthcare. Extracted causal information from clinical notes can be combined with structured EHR data such as patients' demographics, diagnoses, and medications. This will enhance healthcare providers' ability to identify aspects of a patient's story communicated in the clinical notes and help make more informed decisions.
In this work, we propose annotation guidelines, develop an annotated corpus and provide baseline scores to identify types and direction of causal relations between a pair of biomedical concepts in clinical notes; communicated implicitly or explicitly, identified either in a single sentence or across multiple sentences.
We annotate a total of 2714 de-identified examples sampled from the 2018 n2c2 shared task dataset and train four different language model based architectures. Annotation based on our guidelines achieved a high inter-annotator agreement i.e. Fleiss' kappa (κ) score of 0.72, and our model for identification of causal relations achieved a macro F1 score of 0.56 on the test data. The high inter-annotator agreement for clinical text shows the quality of our annotation guidelines while the provided baseline F1 score sets the direction for future research towards understanding narratives in clinical texts.

10:45-12:15 (Forum)
#33 xGQA: Cross-Lingual Visual Question Answering
Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O. Steitz, Stefan Roth, Ivan Vulić and Iryna Gurevych
Recent advances in multimodal vision and language modeling have predominantly focused on the English language, mostly due to the lack of multilingual multimodal datasets to steer modeling efforts. In this work, we address this gap and provide xGQA, a new multilingual evaluation benchmark for the visual question answering task. We extend the established English GQA dataset to 7 typologically diverse languages, enabling us to detect and explore crucial challenges in cross-lingual visual question answering. We further propose new adapter-based approaches to adapt multimodal transformer-based models to become multilingual, and—vice versa—multilingual models to become

multimodal. Our proposed methods outperform current state-of-the-art multilingual multimodal models (e.g., M3P) in zero-shot cross-lingual settings, but the accuracy remains low across the board; a performance drop of around 38 accuracy points in target languages showcases the difficulty of zero-shot cross-lingual transfer for this task. Our results suggest that simple cross-lingual transfer of multimodal models yields latent multilingual multimodal misalignment, calling for more sophisticated methods for vision and multilingual language modeling.

10:45-12:15 (Forum)
#34 End-to-End Speech Translation for Code Switched Speech
Orion Weller, Matthias Sperber, Telmo Pires, Hendra Setiawan, Christian Gollan, Dominic C Telaar and Matthias Paulik
Code switching (CS) refers to the phenomenon of interchangeably using words and phrases from different languages. CS can pose significant accuracy challenges to NLP, due to the often monolingual nature of the underlying systems. In this work, we focus on CS in the context of English/Spanish conversations for the task of speech translation (ST), generating and evaluating both transcript and translation. To evaluate model performance on this task, we create a novel ST corpus derived from existing public data sets. We explore various ST architectures across two dimensions: cascaded (transcribe then translate) vs end-to-end (jointly transcribe and translate) and unidirectional (source -> target) vs bidirectional (source <-> target). We show that our ST architectures, and especially our bidirectional end-to-end architecture, perform well on CS speech, even when no CS training data is used.

10:45-12:15 (Forum)
#35 Nibbling at the Hard Core of Word Sense Disambiguation
Marco Maru, Simone Conia, Michele Bevilacqua and Roberto Navigli
With state-of-the-art systems having finally attained estimated human performance, Word Sense Disambiguation (WSD) has now joined the array of Natural Language Processing tasks that have seemingly been solved, thanks to the vast amounts of knowledge encoded into Transformer-based pre-trained language models. And yet, if we look below the surface of raw figures, it is easy to realize that current approaches still make trivial mistakes that a human would never make. In this work, we provide evidence showing why the F1 score metric should not simply be taken at face value and present an exhaustive analysis of the errors that seven of the most representative state-of-the-art systems for English all-words WSD make on traditional evaluation benchmarks. In addition, we produce and release a collection of test sets featuring (a) an amended version of the standard evaluation benchmark that fixes its lexical and semantic inaccuracies, (b) 42D, a challenge set devised to assess the resilience of systems with respect to least frequent word senses and senses not seen at training time, and (c) hardEN, a challenge set made up solely of instances which none of the investigated state-of-the-art systems can solve. We make all of the test sets and model predictions available to the research community at https://github.com/SapienzaNLP/wsd-hard-benchmark.

10:45-12:15 (Forum)
#36 ParaDetox: Detoxification with Parallel Data
Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov and Alexander Panchenko
We present a novel pipeline for the collection of parallel data for the detoxification task. We collect non-toxic paraphrases for over 10,000 English toxic sentences. We also show that this pipeline can be used to distill a large existing corpus of paraphrases to get toxic-neutral sentence pairs. We release two parallel corpora which can be used for the training of detoxification models. To the best of our knowledge, these are the first parallel datasets for this task. We describe our pipeline in detail to make it fast to set up for a new language or domain, thus contributing to faster and easier development of new parallel resources.
We train several detoxification models on the collected data and compare them with several baselines and state-of-the-art unsupervised approaches. We conduct both automatic and manual evaluations. All models trained on parallel data outperform the state-of-the-art unsupervised models by a large margin. This suggests that our novel datasets can boost the performance of detoxification systems.

10:45-12:15 (Forum)
#37 On the Safety of Conversational Models: Taxonomy, Dataset, and Benchmark
Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu and Minlie Huang
Dialogue safety problems severely limit the real-world deployment of neural conversational models and have attracted great research interests recently. However, dialogue safety problems remain under-defined and the corresponding dataset is scarce. We propose a taxonomy for dialogue safety specifically designed to capture unsafe behaviors in human-bot dialogue settings, with focuses on context-sensitive unsafety, which is under-explored in prior works. To spur research in this direction, we compile DiaSafety, a dataset with rich context-sensitive unsafe examples. Experiments show that existing safety guarding tools fail severely on our dataset. As a remedy, we train a dialogue safety classifier to provide a strong baseline for context-sensitive dialogue unsafety detection. With our classifier, we perform safety evaluations on popular conversational models and show that existing dialogue systems still exhibit concerning context-sensitive safety problems.

10:45-12:15 (Forum)
#38 SRL4E – Semantic Role Labeling for Emotions: A Unified Evaluation Framework
Cesare Campagnano, Simone Conia and Roberto Navigli
In the field of sentiment analysis, several studies have highlighted that a single sentence may express multiple, sometimes contrasting, sentiments and emotions, each with its own experiencer, target and/or cause. To this end, over the past few years researchers have started to collect and annotate data manually, in order to investigate the capabilities of automatic systems not only to distinguish between emotions, but also to capture their semantic constituents. However, currently available gold datasets are heterogeneous in size, domain, format, splits, emotion categories and role labels, making comparisons across different works difficult and hampering progress in the area. In this paper, we tackle this issue and present a unified evaluation framework focused on Semantic Role Labeling for Emotions (SRL4E), in which we unify several datasets tagged with emotions and semantic roles by using a common labeling scheme. We use SRL4E as a benchmark to evaluate how modern pretrained language models perform and analyze where we currently stand in this task, hoping to provide the tools to facilitate studies in this complex area.

10:45-12:15 (Forum)
[DEMO] **AnnIE: An Annotation Platform for Constructing Complete Open Information Extraction Benchmark**
Niklas Friedrich, Kiril Gashteovski, Mingying Yu, Bhushan Kotnis, Carolin Lawrence, Mathias Niepert and Goran Glavaš
Open Information Extraction (OIE) is the task of extracting facts from sentences in the form of relations and their corresponding arguments in schema-free manner. Intrinsic performance of OIE systems is difficult to measure due to the incompleteness of existing OIE benchmarks: ground truth extractions do not group all acceptable surface realizations of the same fact that can be extracted from a sentence. To measure

performance of OIE systems more realistically, it is necessary to manually annotate complete facts (i.e., clusters of all acceptable surface realizations of the same fact) from input sentences. We propose AnnIE: an interactive annotation platform that facilitates such challenging annotation tasks and supports creation of complete fact-oriented OIE evaluation benchmarks. AnnIE is modular and flexible in order to support different use case scenarios (i.e., benchmarks covering different types of facts) and different languages. We use AnnIE to build two complete OIE benchmarks: one with verb-mediated facts and another with facts encompassing named entities. We evaluate several OIE systems on our complete benchmarks created with AnnIE. We publicly release AnnIE (and all gold datasets generated with it) under non-restrictive license.

10:45-12:15 (Forum)
[DEMO] **AdapterHub Playground: Simple and Flexible Few-Shot Learning with Adapters**
Tilman Beck, Bela Bohlender, Christina Viehmann, Vincent Hane, Yanik Adamson, Jaber Khuri, Jonas Brossmann, Jonas Pfeiffer and Iryna Gurevych
The open-access dissemination of pretrained language models through online repositories has led to a democratization of state-of-the-art natural language processing (NLP) research. This also allows people outside of NLP to use such models and adapt them to specific use-cases. However, a certain amount of technical proficiency is still required which is an entry barrier for users who want to apply these models to a certain task but lack the necessary knowledge or resources. In this work, we aim to overcome this gap by providing a tool which allows researchers to leverage pretrained models without writing a single line of code. Built upon the parameter-efficient adapter modules for transfer learning, our AdapterHub Playground provides an intuitive interface, allowing the usage of adapters for prediction, training and analysis of textual data for a variety of NLP tasks. We present the tool's architecture and demonstrate its advantages with prototypical use-cases, where we show that predictive performance can easily be increased in a few-shot learning scenario. Finally, we evaluate its usability in a user study. We provide the code and a live interface at https://adapter-hub.github.io/playground.

10:45-12:15 (Forum)
[DEMO] **TS-ANNO: An Annotation Tool to Build, Annotate and Evaluate Text Simplification Corpora**
Regina Stodden and Laura Kallmeyer
We introduce TS-ANNO, an open-source web application for manual creation and for evaluation of parallel corpora for text simplification. TS-ANNO can be used for i) sentence–wise alignment, ii) rating alignment pairs (e.g., w.r.t. grammaticality, meaning preservation, ...), iii) annotating alignment pairs w.r.t. simplification transformations (e.g., lexical substitution, sentence splitting, ...), and iv) manual simplification of complex documents. For evaluation, TS-ANNO calculates inter-annotator agreement of alignments (i) and annotations (ii).

10:45-12:15 (Forum)
[DEMO] **Dynatask: A Framework for Creating Dynamic AI Benchmark Tasks**
Tristan Thrush, Kushal Tirumala, Anmol Gupta, Max Bartolo, Pedro Rodriguez, Tariq Kane, William Gaviria Rojas, Peter Mattson, Adina Williams and Douwe Kiela
We introduce Dynatask: an open source system for setting up custom NLP tasks that aims to greatly lower the technical knowledge and effort required for hosting and evaluating state-of-the-art NLP models, as well as for conducting model in the loop data collection with crowdworkers. Dynatask is integrated with Dynabench, a research platform for rethinking benchmarking in AI that facilitates human and model in the loop data collection and evaluation. To create a task, users only need to write a short task configuration file from which the relevant web interfaces and model hosting infrastructure are automatically generated. The system is available at https://dynabench.org/ and the full library can be found at https://github.com/facebookresearch/dynabench.

Poster Session 6: Machine Translation and Multilinguality

10:45-12:15 (Forum)

10:45-12:15 (Forum)
#39 Multilingual Document-Level Translation Enables Zero-Shot Transfer From Sentences to Documents
Biao Zhang, Ankur Bapna, Melvin Johnson, Ali Dabirmoghaddam, Naveen Arivazhagan and Orhan Firat
Document-level neural machine translation (DocNMT) achieves coherent translations by incorporating cross-sentence context. However, for most language pairs there's a shortage of parallel documents, although parallel sentences are readily available. In this paper, we study whether and how contextual modeling in DocNMT is transferable via multilingual modeling. We focus on the scenario of zero-shot transfer from teacher languages with document level data to student languages with no documents but sentence level data, and for the first time treat document-level translation as a transfer learning problem. Using simple concatenation-based DocNMT, we explore the effect of 3 factors on the transfer: the number of teacher languages with document level data, the balance between document and sentence level data at training, and the data condition of parallel documents (genuine vs. back-translated). Our experiments on Europarl-7 and IWSLT-10 show the feasibility of multilingual transfer for DocNMT, particularly on document-specific metrics. We observe that more teacher languages and adequate data balance both contribute to better transfer quality. Surprisingly, the transfer is less sensitive to the data condition, where multilingual DocNMT delivers decent performance with either back-translated or genuine document pairs.

10:45-12:15 (Forum)
#40 Bilingual alignment transfers to multilingual alignment for unsupervised parallel text mining
Chih-chan Tien and Shane Steinert-Threlkeld
This work presents methods for learning cross-lingual sentence representations using paired or unpaired bilingual texts. We hypothesize that the cross-lingual alignment strategy is transferable, and therefore a model trained to align only two languages can encode multilingually more aligned representations. We thus introduce dual-pivot transfer: training on one language pair and evaluating on other pairs. To study this theory, we design unsupervised models trained on unpaired sentences and single-pair supervised models trained on bitexts, both based on the unsupervised language model XLM-R with its parameters frozen. The experiments evaluate the models as universal sentence encoders on the task of unsupervised bitext mining on two datasets, where the unsupervised model reaches the state of the art of unsupervised retrieval, and the alternative single-pair supervised model approaches the performance of multilingually supervised models. The results suggest that bilingual training techniques as proposed can be applied to get sentence representations with multilingual alignment.

10:45-12:15 (Forum)
#41 Combining Static and Contextualised Multilingual Embeddings
Katharina Hämmerl, Jindřich Libovický and Alexander Fraser
Static and contextual multilingual embeddings have complementary strengths. Static embeddings, while less expressive than contextual language models, can be more straightforwardly aligned across multiple languages. We combine the strengths of static and contextual models to improve multilingual representations. We extract static embeddings for 40 languages from XLM-R, validate those embeddings with cross-lingual word retrieval, and then align them using VecMap. This results in high-quality, highly multilingual static embeddings. Then we apply a novel continued pre-training approach to XLM-R, leveraging the high quality alignment of our static embeddings to better align the representation space of XLM-R. We show positive results for multiple complex semantic tasks. We release the static embeddings and the continued pre-training code. Unlike most previous work, our continued pre-training approach does not require parallel text.

10:45-12:15 (Forum)
#42 Better Quality Estimation for Low Resource Corpus Mining
Muhammed Yusuf Kocyigit, Jiho Lee and Derry Wijaya
Quality Estimation (QE) models have the potential to change how we evaluate and maybe even train machine translation models. However, these models still lack the robustness to achieve general adoption. We show that Stateof-the-art QE models, when tested in a Parallel Corpus Mining (PCM) setting, perform unexpectedly bad due to a lack of robustness to out-of-domain examples. We propose a combination of multitask training, data augmentation and contrastive learning to achieve better and more robust QE performance. We show that our method improves QE performance significantly in the MLQE challenge and the robustness of QE models when tested in the Parallel Corpus Mining setup. We increase the accuracy in PCM by more than 0.80, making it on par with state-of-the-art PCM methods that use millions of sentence pairs to train their models. In comparison, we use a thousand times less data, 7K parallel sentences in total, and propose a novel low resource PCM method.

10:45-12:15 (Forum)
#43 Detecting Various Types of Noise for Neural Machine Translation
Christian Herold, Jan Rosendahl, Joris Vanvinckenroye and Hermann Ney
The filtering and/or selection of training data is one of the core aspects to be considered when building a strong machine translation system. In their influential work, Khayrallah and Koehn (2018) investigated the impact of different types of noise on the performance of machine translation systems. In the same year the WMT introduced a shared task on parallel corpus filtering, which went on to be repeated in the following years, and resulted in many different filtering approaches being proposed. In this work we aim to combine the recent achievements in data filtering with the original analysis of Khayrallah and Koehn (2018) and investigate whether state-of-the-art filtering systems are capable of removing all the suggested noise types. We observe that most of these types of noise can be detected with an accuracy of over 90However, we also find that when confronted with more refined noise categories or when working with a less common language pair, the performance of the filtering systems is far from optimal, showing that there is still room for improvement in this area of research.

10:45-12:15 (Forum)
[TACL] **#44 Samanantar: The Largest Publicly Available Parallel Corpora Collection For 11 Indic Languages**
Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh Khapra and Srihari Nagaraj

10:45-12:15 (Forum)
[TACL] **#45 ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models**
Noah Constant, Linting Xue, Aditya Barua, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts and Colin Raffel

10:45-12:15 (Forum)
[TACL] **#46 Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets**
Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Rubungo Niyongabo, Toan Nguyen, Mathias Müller, André Müller, Shamsuddeen Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal and Mofetoluwa Adeyemi

10:45-12:15 (Forum)
#47 Overlap-based Vocabulary Generation Improves Cross-lingual Transfer Among Related Languages
Vaidehi Patil, Partha Talukdar and Sunita Sarawagi
Pre-trained multilingual language models such as mBERT and XLM-R have demonstrated great potential for zero-shot cross-lingual transfer to low web-resource languages (LRL). However, due to limited model capacity, the large difference in the sizes of available monolingual corpora between high web-resource languages (HRL) and LRLs does not provide enough scope of co-embedding the LRL with the HRL, thereby affecting the downstream task performance of LRLs. In this paper, we argue that relatedness among languages in a language family along the dimension of lexical overlap may be leveraged to overcome some of the corpora limitations of LRLs. We propose Overlap BPE (OBPE), a simple yet effective modification to the BPE vocabulary generation algorithm which enhances overlap across related languages. Through extensive experiments on multiple NLP tasks and datasets, we observe that OBPE generates a vocabulary that increases the representation of LRLs via tokens shared with HRLs. This results in improved zero-shot transfer from related HRLs to LRLs without reducing HRL representation and accuracy. Unlike previous studies that dismissed the importance of token-overlap, we show that in the low-resource related language setting, token overlap matters. Synthetically reducing the overlap to zero can cause as much as a four-fold drop in zero-shot transfer accuracy.

10:45-12:15 (Forum)
#48 Language-agnostic BERT Sentence Embedding

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan and Wei Wang
While BERT is an effective method for learning monolingual sentence embeddings for semantic similarity and embedding based transfer learning BERT based cross-lingual sentence embeddings have yet to be explored. We systematically investigate methods for learning multilingual sentence embeddings by combining the best methods for learning monolingual and cross-lingual representations including: masked language modeling (MLM), translation language modeling (TLM), dual encoder translation ranking, and additive margin softmax. We show that introducing a pre-trained multilingual language model dramatically reduces the amount of parallel training data required to achieve good performance by 80dev/google/LaBSE.

10:45-12:15 (Forum)
#49 Match the Script, Adapt if Multilingual: Analyzing the Effect of Multilingual Pretraining on Cross-lingual Transferability
Yoshinari Fujinuma, Jordan Lee Boyd-Graber and Katharina Kann
Pretrained multilingual models enable zero-shot learning even for unseen languages, and that performance can be further improved via adaptation prior to finetuning. However, it is unclear how the number of pretraining languages influences a model's zero-shot learning for languages unseen during pretraining. To fill this gap, we ask the following research questions: (1) How does the number of pretraining languages influence zero-shot performance on unseen target languages? (2) Does the answer to that question change with model adaptation? (3) Do the findings for our first question change if the languages used for pretraining are all related? Our experiments on pretraining with related languages indicate that choosing a diverse set of languages is crucial. Without model adaptation, surprisingly, increasing the number of pretraining languages yields better results up to adding related languages, after which performance plateaus. In contrast, with model adaptation via continued pretraining, pretraining on a larger number of languages often gives further improvement, suggesting that model adaptation is crucial to exploit additional pretraining languages.

10:45-12:15 (Forum)
#50 Composable Sparse Fine-Tuning for Cross-Lingual Transfer
Alan Ansell, Edoardo Ponti, Anna Korhonen and Ivan Vulić
Fine-tuning the entire set of parameters of a large pretrained model has become the mainstream approach for transfer learning. To increase its efficiency and prevent catastrophic forgetting and interference, techniques like adapters and sparse fine-tuning have been developed. Adapters are modular, as they can be combined to adapt a model towards different facets of knowledge (e.g., dedicated language and/or task adapters). Sparse fine-tuning is expressive, as it controls the behavior of all model components. In this work, we introduce a new fine-tuning method with both these desirable properties. In particular, we learn sparse, real-valued masks based on a simple variant of the Lottery Ticket Hypothesis. Task-specific masks are obtained from annotated data in a source language, and language-specific masks from masked language modeling in a target language. Both these masks can then be composed with the pretrained model. Unlike adapter-based fine-tuning, this method neither increases the number of parameters at inference time nor alters the original model architecture. Most importantly, it outperforms adapters in zero-shot cross-lingual transfer by a large margin in a series of multilingual benchmarks, including Universal Dependencies, MasakhaNER, and AmericasNLI. Based on an in-depth analysis, we additionally find that sparsity is crucial to prevent both 1) interference between the fine-tunings to be composed and 2) overfitting. We release the code and models at https://github.com/cambridgeltl/composable-sft.

10:45-12:15 (Forum)
#51 Under the Morphosyntactic Lens: A Multifaceted Evaluation of Gender Bias in Speech Translation
Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri and Marco Turchi
Gender bias is largely recognized as a problematic phenomenon affecting language technologies, with recent studies underscoring that it might surface differently across languages. However, most of current evaluation practices adopt a word-level focus on a narrow set of occupational nouns under synthetic conditions. Such protocols overlook key features of grammatical gender languages, which are characterized by morphosyntactic chains of gender agreement, marked on a variety of lexical items and parts-of-speech (POS). To overcome this limitation, we enrich the natural, gender-sensitive MuST-SHE corpus (Bentivogli et al., 2020) with two new linguistic annotation layers (POS and agreement chains), and explore to what extent different lexical categories and agreement phenomena are impacted by gender skews. Focusing on speech translation, we conduct a multifaceted evaluation on three language directions (English-French/Italian/Spanish), with models trained on varying amounts of data and different word segmentation techniques. By shedding light on model behaviours, gender bias, and its detection at several levels of granularity, our findings emphasize the value of dedicated analyses beyond aggregated overall results.

10:45-12:15 (Forum)
#52 Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder Translation Models
Lorenzo Lupo, Marco Dinarelli and Laurent Besacier
Multi-encoder models are a broad family of context-aware neural machine translation systems that aim to improve translation quality by encoding document-level contextual information alongside the current sentence. The context encoding is undertaken by contextual parameters, trained on document-level data. In this work, we discuss the difficulty of training these parameters effectively, due to the sparsity of the words in need of context (i.e., the training signal), and their relevant context. We propose to pre-train the contextual parameters over split sentence pairs, which makes an efficient use of the available data for two reasons. Firstly, it increases the contextual training signal by breaking intra-sentential syntactic relations, and thus pushing the model to search the context for disambiguating clues more frequently. Secondly, it eases the retrieval of relevant context, since context segments become shorter. We propose four different splitting methods, and evaluate our approach with BLEU and contrastive test sets. Results show that it consistently improves learning of contextual parameters, both in low and high resource settings.

10:45-12:15 (Forum)
#53 Multilingual Generative Language Models for Zero-Shot Cross-Lingual Event Argument Extraction
Kuan-Hao Huang, I-Hung Hsu, Prem Natarajan, Kai-Wei Chang and Nanyun Peng
We present a study on leveraging multilingual pre-trained generative language models for zero-shot cross-lingual event argument extraction (EAE). By formulating EAE as a language generation task, our method effectively encodes event structures and captures the dependencies between arguments. We design language-agnostic templates to represent the event argument structures, which are compatible with any language, hence facilitating the cross-lingual transfer. Our proposed model finetunes multilingual pre-trained generative language models to generate sentences that fill in the language-agnostic template with arguments extracted from the input passage. The model is trained on source languages and is then directly applied to target languages for event argument extraction. Experiments demonstrate that the proposed model outperforms the current state-of-the-art models on zero-shot cross-lingual EAE. Comprehensive studies and error analyses are presented to better understand the advantages and the current limitations of using generative language models for zero-shot cross-lingual transfer EAE.

10:45-12:15 (Forum)
#54 From Simultaneous to Streaming Machine Translation by Leveraging Streaming History
Javier Iranzo Sanchez, Jorge Civera and Alfons Juan-Císcar
Simultaneous machine translation has recently gained traction thanks to significant quality improvements and the advent of streaming applications. Simultaneous translation systems need to find a trade-off between translation quality and response time, and with this purpose multiple latency measures have been proposed. However, latency evaluations for simultaneous translation are estimated at the sentence level, not taking into account the sequential nature of a streaming scenario. Indeed, these sentence-level latency measures are not well suited for continuous stream translation, resulting in figures that are not coherent with the simultaneous translation policy of the system being assessed. This work proposes a stream-level adaptation of the current latency measures based on a re-segmentation approach applied to the output translation, that is successfully evaluated on streaming conditions for a reference IWSLT task

10:45-12:15 (Forum)
#55 Accurate Online Posterior Alignments for Principled Lexically-Constrained Decoding
Soumya Chatterjee, Sunita Sarawagi and Preethi Jyothi
Online alignment in machine translation refers to the task of aligning a target word to a source word when the target sequence has only been partially decoded. Good online alignments facilitate important applications such as lexically constrained translation where user-defined dictionaries are used to inject lexical constraints into the translation model. We propose a novel posterior alignment technique that is truly online in its execution and superior in terms of alignment error rates compared to existing methods. Our proposed inference technique jointly considers alignment and token probabilities in a principled manner and can be seamlessly integrated within existing constrained beam-search decoding algorithms. On five language pairs, including two distant language pairs, we achieve consistent drop in alignment error rates. When deployed on seven lexically constrained translation tasks, we achieve significant improvements in BLEU specifically around the constrained positions.

10:45-12:15 (Forum)
#56 As Little as Possible, as Much as Necessary: Detecting Over- and Undertranslations with Contrastive Conditioning
Jannis Vamvas and Rico Sennrich
Omission and addition of content is a typical issue in neural machine translation. We propose a method for detecting such phenomena with off-the-shelf translation models. Using contrastive conditioning, we compare the likelihood of a full sequence under a translation model to the likelihood of its parts, given the corresponding source or target sequence. This allows to pinpoint superfluous words in the translation and untranslated words in the source even in the absence of a reference translation. The accuracy of our method is comparable to a supervised method that requires a custom quality estimation model.

10:45-12:15 (Forum)
#57 CipherDAug: Ciphertext based Data Augmentation for Neural Machine Translation
Nishant Kambhatla, Logan Born and Anoop Sarkar
We propose a novel data-augmentation technique for neural machine translation based on ROT-k ciphertexts. ROT-k is a simple letter substitution cipher that replaces a letter in the plaintext with the kth letter after it in the alphabet. We first generate multiple ROT-k ciphertexts using different values of k for the plaintext which is the source side of the parallel data. We then leverage this enciphered training data along with the original parallel data via multi-source training to improve neural machine translation. Our method, CipherDAug, uses a co-regularization-inspired training procedure, requires no external data sources other than the original training data, and uses a standard Transformer to outperform strong data augmentation techniques on several datasets by a significant margin. This technique combines easily with existing approaches to data augmentation, and yields particularly strong results in low-resource settings.

10:45-12:15 (Forum)
#58 Graph Neural Networks for Multiparallel Word Alignment
Ayyoob Imani, Lütfi Kerem Senel, Masoud Jalili Sabet, François Yvon and Hinrich Schuetze
After a period of decrease, interest in word alignments is increasing again for their usefulness in domains such as typological research, cross-lingual annotation projection and machine translation. Generally, alignment algorithms only use bitext and do not make use of the fact that many parallel corpora are multiparallel. Here, we compute high-quality word alignments between multiple language pairs by considering all language pairs together. First, we create a multiparallel word alignment graph, joining all bilingual word alignment pairs in one graph. Next, we use graph neural networks (GNNs) to exploit the graph structure. Our GNN approach (i) utilizes information about the meaning, position and language of the input words, (ii) incorporates information from multiple parallel sentences, (iii) adds and removes edges from the initial alignments, and (iv) yields a prediction model that can generalize beyond the training sentences. We show that community detection algorithms can provide valuable information for multiparallel word alignment. Our method outperforms previous work on three word alignment datasets and on a downstream task.

10:45-12:15 (Forum)
#59 IndicBART: A Pre-trained Model for Indic Natural Language Generation
Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra and Pratyush Kumar
In this paper, we study pre-trained sequence-to-sequence models for a group of related languages, with a focus on Indic languages. We present IndicBART, a multilingual, sequence-to-sequence pre-trained model focusing on 11 Indic languages and English. IndicBART utilizes the orthographic similarity between Indic scripts to improve transfer learning between similar Indic languages. We evaluate IndicBART on two NLG tasks: Neural Machine Translation (NMT) and extreme summarization. Our experiments on NMT and extreme summarization show that a model specific to related languages like IndicBART is competitive with large pre-trained models like mBART50 despite being significantly smaller. It also performs well on very low-resource translation scenarios where languages are not included in pre-training or fine-tuning. Script sharing, multilingual training, and better utilization of limited model capacity contribute to the good performance of the compact IndicBART model.

10:45-12:15 (Forum)
#60 Meta-X$_{NLG}$: A Meta-Learning Approach Based on Language Clustering for Zero-Shot Cross-Lingual Transfer and Generation
Kaushal Kumar Maurya and Maunendra Sankar Desarkar

Recently, the NLP community has witnessed a rapid advancement in multilingual and cross-lingual transfer research where the supervision is transferred from high-resource languages (HRLs) to low-resource languages (LRLs). However, the cross-lingual transfer is not uniform across languages, particularly in the zero-shot setting. Towards this goal, one promising research direction is to learn shareable structures across multiple tasks with limited annotated data. The downstream multilingual applications may benefit from such a learning setup as most of the languages across the globe are low-resource and share some structures with other languages. In this paper, we propose a novel meta-learning framework (called Meta-X$_{NLG}$) to learn shareable structures from typologically diverse languages based on meta-learning and language clustering. This is a step towards uniform cross-lingual transfer for unseen languages. We first cluster the languages based on language representations and identify the centroid language of each cluster. Then, a meta-learning algorithm is trained with all centroid languages and evaluated on the other languages in the zero-shot setting. We demonstrate the effectiveness of this modeling on two NLG tasks (Abstractive Text Summarization and Question Generation), 5 popular datasets and 30 typologically diverse languages. Consistent improvements over strong baselines demonstrate the efficacy of the proposed framework. The careful design of the model makes this end-to-end NLG setup less vulnerable to the accidental translation problem, which is a prominent concern in zero-shot cross-lingual NLG tasks.

10:45-12:15 (Forum)
#61 A Natural Diet: Towards Improving Naturalness of Machine Translation Output
Markus Freitag, David Vilar, David Grangier, Colin Cherry and George Foster
Machine translation (MT) evaluation often focuses on accuracy and fluency, without paying much attention to translation style. This means that, even when considered accurate and fluent, MT output can still sound less natural than high quality human translations or text originally written in the target language. Machine translation output notably exhibits lower lexical diversity, and employs constructs that mirror those in the source sentence. In this work we propose a method for training MT systems to achieve a more natural style, i.e. mirroring the style of text originally written in the target language. Our method tags parallel training data according to the naturalness of the target side by contrasting language models trained on natural and translated data. Tagging data allows us to put greater emphasis on target sentences originally written in the target language. Automatic metrics show that the resulting models achieve lexical richness on par with human translations, mimicking a style much closer to sentences originally written in the target language. Furthermore, we find that their output is preferred by human experts when compared to the baseline translations.

10:45-12:15 (Forum)
#62 The impact of lexical and grammatical processing on generating code from natural language
Nathanaël Beau and Benoit Crabbé
Considering the seq2seq architecture of Yin and Neubig (2018) for natural language to code translation, we identify four key components of importance: grammatical constraints, lexical preprocessing, input representations, and copy mechanisms. To study the impact of these components, we use a state-of-the-art architecture that relies on BERT encoder and a grammar-based decoder for which a formalization is provided. The paper highlights the importance of the lexical substitution component in the current natural language to code systems.

10:45-12:15 (Forum)
#63 First the Worst: Finding Better Gender Translations During Beam Search
Danielle Saunders, Rosie Sallis and Bill Byrne
Generating machine translations via beam search seeks the most likely output under a model. However, beam search has been shown to amplify demographic biases exhibited by a model. We aim to address this, focusing on gender bias resulting from systematic errors in grammatical gender translation. Almost all prior work on this problem adjusts the training data or the model itself. By contrast, our approach changes only the inference procedure.
We constrain beam search to improve gender diversity in n-best lists, and rerank n-best lists using gender features obtained from the source sentence. Combining these strongly improves WinoMT gender translation accuracy for three language pairs without additional bilingual data or retraining. We also demonstrate our approach's utility for consistently gendering named entities, and its flexibility to handle new gendered language beyond the binary.

10:45-12:15 (Forum)
#64 Why don't people use character-level machine translation?
Jindřich Libovický, Helmut Schmid and Alexander Fraser
We present a literature and empirical survey that critically assesses the state of the art in character-level modeling for machine translation (MT). Despite evidence in the literature that character-level systems are comparable with subword systems, they are virtually never used in competitive setups in WMT competitions. We empirically show that even with recent modeling innovations in character-level natural language processing, character-level MT systems still struggle to match their subword-based counterparts. Character-level MT systems show neither better domain robustness, nor better morphological generalization, despite being often so motivated. However, we are able to show robustness towards source side noise and that translation quality does not degrade with increasing beam size at decoding time.

10:45-12:15 (Forum)
#65 Domain Generalisation of NMT: Fusing Adapters with Leave-One-Domain-Out Training
Thuy-Trang Vu, Shahram Khadivi, Dinh Phung and Gholamreza Haffari
Generalising to unseen domains is under-explored and remains a challenge in neural machine translation. Inspired by recent research in parameter-efficient transfer learning from pretrained models, this paper proposes a fusion-based generalisation method that learns to combine domain-specific parameters. We propose a leave-one-domain-out training strategy to avoid information leaking to address the challenge of not knowing the test domain during training time. Empirical results on three language pairs show that our proposed fusion method outperforms other baselines up to +0.8 BLEU score on average.

10:45-12:15 (Forum)
[TACL] **#66 The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation**
Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán and Angela Fan

10:45-12:15 (Forum)
[TACL] **#67 Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation**

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan and Wolfgang Macherey

10:45-12:15 (Forum)
#68 Can Synthetic Translations Improve Bitext Quality?
Eleftheria Briakou and Marine Carpuat
Synthetic translations have been used for a wide range of NLP tasks primarily as a means of data augmentation. This work explores, instead, how synthetic translations can be used to revise potentially imperfect reference translations in mined bitext. We find that synthetic samples can improve bitext quality without any additional bilingual supervision when they replace the originals based on a semantic equivalence classifier that helps mitigate NMT noise. The improved quality of the revised bitext is confirmed intrinsically via human evaluation and extrinsically through bilingual induction and MT tasks.

10:45-12:15 (Forum)
#69 Multi Task Learning For Zero Shot Performance Prediction of Multilingual Models
Kabir Ahuja, Shanu Kumar, Sandipan Dandapat and Monojit Choudhury
Massively Multilingual Transformer based Language Models have been observed to be surprisingly effective on zero-shot transfer across languages, though the performance varies from language to language depending on the pivot language(s) used for fine-tuning. In this work, we build upon some of the existing techniques for predicting the zero-shot performance on a task, by modeling it as a multi-task learning problem. We jointly train predictive models for different tasks which helps us build more accurate predictors for tasks where we have test data in very few languages to measure the actual performance of the model. Our approach also lends us the ability to perform a much more robust feature selection, and identify a common set of features that influence zero-shot performance across a variety of tasks.

10:45-12:15 (Forum)
#70 Flow-Adapter Architecture for Unsupervised Machine Translation
Yihong Liu, Haris Jabbar and Hinrich Schuetze
In this work, we propose a flow-adapter architecture for unsupervised NMT. It leverages normalizing flows to explicitly model the distributions of sentence-level latent representations, which are subsequently used in conjunction with the attention mechanism for the translation task. The primary novelties of our model are: (a) capturing language-specific sentence representations separately for each language using normalizing flows and (b) using a simple transformation of these latent representations for translating from one language to another. This architecture allows for unsupervised training of each language independently. While there is prior work on latent variables for supervised MT, to the best of our knowledge, this is the first work that uses latent variables and normalizing flows for unsupervised MT. We obtain competitive results on several unsupervised MT benchmarks.

10:45-12:15 (Forum)
#71 Automatic Song Translation for Tonal Languages
Fenfei Guo, Chen Zhang, Zhirui Zhang, Qixin He, Kejun Zhang, Jun Xie and Jordan Lee Boyd-Graber
This paper develops automatic song translation (AST) for tonal languages and addresses the unique challenge of aligning words' tones with melody of a song in addition to conveying the original meaning. We propose three criteria for effective AST—preserving meaning, singability and intelligibility—and design metrics for these criteria. We develop a new benchmark for English–Mandarin song translation and develop an unsupervised AST system, Guided AliGnment for Automatic Song Translation (GagaST), which combines pre-training with three decoding constraints. Both automatic and human evaluations show GagaST successfully balances semantics and singability.

10:45-12:15 (Forum)
[DEMO] **Language Diversity: Visible to Humans, Exploitable by Machines**
Gábor Bella, Erdenebileg Byambadorj, Yamini Chandrashekar, Khuyagbaatar Batsuren, Danish Cheema and Fausto Giunchiglia
The Universal Knowledge Core (UKC) is a large multilingual lexical database with a focus on language diversity and covering over two thousand languages. The aim of the database, as well as its tools and data catalogue, is to make the abstract notion of linguistic diversity visually understandable for humans and formally exploitable by machines. The UKC website lets users explore millions of individual words and their meanings, but also phenomena of cross-lingual convergence and divergence, such as shared interlingual meanings, lexicon similarities, cognate clusters, or lexical gaps. The UKC LiveLanguage Catalogue, in turn, provides access to the underlying lexical data in a computer-processable form, ready to be reused in cross-lingual applications.

Poster Session 6: Special Theme on Language Diversity: From Low Resource to Endangered Languages

10:45-12:15 (Forum)

10:45-12:15 (Forum)
#72 Multilingual unsupervised sequence segmentation transfers to extremely low-resource languages
C. M. Downey, Shannon Drizin, Levon Haroutunian and Shivin Thukral
We show that unsupervised sequence-segmentation performance can be transferred to extremely low-resource languages by pre-training a Masked Segmental Language Model (Downey et al., 2021) multilingually. Further, we show that this transfer can be achieved by training over a collection of low-resource languages that are typologically similar (but phylogenetically unrelated) to the target language. In our experiments, we transfer from a collection of 10 Indigenous American languages (AmericasNLP, Mager et al., 2021) to K'iche', a Mayan language. We compare our multilingual model to a monolingual (from-scratch) baseline, as well as a model pre-trained on Quechua only. We show that the multilingual pre-trained approach yields consistent segmentation quality across target dataset sizes, exceeding the monolingual baseline in 6/10 experimental settings. Our model yields especially strong results at small target sizes, including a zero-shot performance of 20.6 F1. These results have promising implications for low-resource NLP pipelines involving human-like linguistic units, such as the sparse transcription framework proposed by Bird (2020).

10:45-12:15 (Forum)
#73 One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia
Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau and Sebastian Ruder
NLP research is impeded by a lack of resources and awareness of the challenges presented by underrepresented languages and dialects. Focusing on the languages spoken in Indonesia, the second most linguistically diverse and the fourth most populous nation of the world, we provide an overview of the current state of NLP research for Indonesia's 700+ languages. We highlight challenges in Indonesian NLP and how these affect the performance of current NLP systems. Finally, we provide general recommendations to help develop NLP technology not only for languages of Indonesia but also other underrepresented languages.

10:45-12:15 (Forum)
#74 Learning From Failure: Data Capture in an Australian Aboriginal Community
Eric Le Ferrand, Steven Bird and Laurent BesacierUGA
Most low resource language technology development is premised on the need to collect data for training statistical models. When we follow the typical process of recording and transcribing text for small Indigenous languages, we hit up against the so-called "transcription bottleneck." Therefore it is worth exploring new ways of engaging with speakers which generate data while avoiding the transcription bottleneck. We have deployed a prototype app for speakers to use for confirming system guesses in an approach to transcription based on word spotting. However, in the process of testing the app we encountered many new problems for engagement with speakers. This paper presents a close-up study of the process of deploying data capture technology on the ground in an Australian Aboriginal community. We reflect on our interactions with participants and draw lessons that apply to anyone seeking to develop methods for language data collection in an Indigenous community.

10:45-12:15 (Forum)
#75 Weakly Supervised Word Segmentation for Computational Language Documentation
Shu Okabe, Laurent BesacierUGA and François Yvon
Word and morpheme segmentation are fundamental steps of language documentation as they allow to discover lexical units in a language for which the lexicon is unknown. However, in most language documentation scenarios, linguists do not start from a blank page: they may already have a pre-existing dictionary or have initiated manual segmentation of a small part of their data. This paper studies how such a weak supervision can be taken advantage of in Bayesian non-parametric models of segmentation. Our experiments on two very low resource languages (Mboshi and Japhug), whose documentation is still in progress, show that weak supervision can be beneficial to the segmentation quality. In addition, we investigate an incremental learning scenario where manual segmentations are provided in a sequential manner. This work opens the way for interactive annotation tools for documentary linguists.

10:45-12:15 (Forum)
#76 Cree Corpus: A Collection of nêhiyawêwin Resources
Daniela Teodorescu, Josie Matalski, Delaney Alexa Lothian, Denilson Barbosa and Carrie Demmans Epp
Plains Cree (nêhiyawêwin) is an Indigenous language that is spoken in Canada and the USA. It is the most widely spoken dialect of Cree and a morphologically complex language that is polysynthetic, highly inflective, and agglutinative. It is an extremely low resource language, with no existing corpus that is both available and prepared for supporting the development of language technologies. To support nêhiyawêwin revitalization and preservation, we developed a corpus covering diverse genres, time periods, and texts for a variety of intended audiences. The data has been verified and cleaned; it is ready for use in developing language technologies for nêhiyawêwin. The corpus includes the corresponding English phrases or audio files where available. We demonstrate the utility of the corpus through its community use and its use to build language technologies that can provide the types of support that community members have expressed are desirable. The corpus is available for public use.

10:45-12:15 (Forum)
#77 Machine Translation for Livonian: Catering to 20 Speakers
Matīss Rikters, Marili Tomingas, Tuuli Tuisk, Valts Ernštreits and Mark Fishel
Livonian is one of the most endangered languages in Europe with just a tiny handful of speakers and virtually no publicly available corpora. In this paper we tackle the task of developing neural machine translation (NMT) between Livonian and English, with a two-fold aim: on one hand, preserving the language and on the other – enabling access to Livonian folklore, lifestories and other textual intangible heritage as well as making it easier to create further parallel corpora. We rely on Livonian's linguistic similarity to Estonian and Latvian and collect parallel and monolingual data for the four languages for translation experiments. We combine different low-resource NMT techniques like zero-shot translation, cross-lingual transfer and synthetic data creation to reach the highest possible translation quality as well as to find which base languages are empirically more helpful for transfer to Livonian. The resulting NMT systems and the collected monolingual and parallel data, including a manually translated and verified translation benchmark, are publicly released via OPUS and Huggingface repositories.

10:45-12:15 (Forum)
#78 Automatic Speech Recognition and Query By Example for Creole Languages Documentation
Cécile Macaire, Didier Schwab, Benjamin Lecouteux and Emmanuel Schang
We investigate the exploitation of self-supervised models for two Creole languages with few resources: Gwadloupéyen and Morisien. Automatic language processing tools are almost non-existent for these two languages. We propose to use about one hour of annotated data to design an automatic speech recognition system for each language. We evaluate how much data is needed to obtain a query-by-example system that is usable by linguists. Moreover, our experiments show that multilingual self-supervised models are not necessarily the most efficient for Creole languages.

10:45-12:15 (Forum)
#79 Towards Responsible Natural Language Annotation for the Varieties of Arabic
A. Stevie Bergman and Mona T. Diab
When building NLP models, there is a tendency to aim for broader coverage, often overlooking cultural and (socio)linguistic nuance. In this position paper, we make the case for care and attention to such nuances, particularly in dataset annotation, as well as the inclusion of cultural and linguistic expertise in the process. We present a playbook for responsible dataset creation for polyglossic, multidialectal languages. This work is informed by a study on Arabic annotation of social media content.

10:45-12:15 (Forum)
#80 Interactive Word Completion for Plains Cree
William Abbott Lane, Atticus Galvin Harrigan and Antti Arppe
The composition of richly-inflected words in morphologically complex languages can be a challenge for language learners developing literacy. Accordingly, Lane and Bird (2020) proposed a finite state approach which maps prefixes in a language to a set of possible completions up to the next morpheme boundary, for the incremental building of complex words. In this work, we develop an approach to morph-based auto-completion based on a finite state morphological analyzer of Plains Cree (nêhiyawêwin), showing the portability of the concept to a much larger, more complete morphological transducer. Additionally, we propose and compare various novel ranking strategies on the morph auto-complete output. The best weighting scheme ranks the target completion in the top 10 results in 64.9

10:45-12:15 (Forum)
#81 Towards Afrocentric NLP for African Languages: Where We Are and Where We Can Go
Ife Adebara and Muhammad Abdul-Mageed
Aligning with ACL 2022 special Theme on "Language Diversity: from Low Resource to Endangered Languages", we discuss the major linguistic and sociopolitical challenges facing development of NLP technologies for African languages. Situating African languages in a typological framework, we discuss how the particulars of these languages can be harnessed. To facilitate future research, we also highlight current efforts, communities, venues, datasets, and tools. Our main objective is to motivate and advocate for an Afrocentric approach to technology development. With this in mind, we recommend *what* technologies to build and *how* to build, evaluate, and deploy them based on the needs of local African communities.

10:45-12:15 (Forum)
#82 Phone-ing it in: Towards Flexible Multi-Modal Language Model Training by Phonetic Representations of Data
Colin Leong and Daniel Lee Whitenack
Multi-modal techniques offer significant untapped potential to unlock improved NLP technology for local languages. However, many advances in language model pre-training are focused on text, a fact that only increases systematic inequalities in the performance of NLP tasks across the world's languages. In this work, we propose a multi-modal approach to train language models using whatever text and/or audio data might be available in a language. Initial experiments using Swahili and Kinyarwanda data suggest the viability of the approach for downstream Named Entity Recognition (NER) tasks, with models pre-trained on phone data showing an improvement of up to 6% F1-score above models that are trained from scratch. Preprocessing and training code will be uploaded to https://github.com/sil-ai/phone-it-in.

10:45-12:15 (Forum)
#83 Requirements and Motivations of Low-Resource Speech Synthesis for Language Revitalization
Aidan Pine, Dan Wells, Nathan Brinklow, Patrick William Littell and Korin Richmond
This paper describes the motivation and development of speech synthesis systems for the purposes of language revitalization. By building speech synthesis systems for three Indigenous languages spoken in Canada, Kanien'kéha, Gitksan & SENĆOTEN, we re-evaluate the question of how much data is required to build low-resource speech synthesis systems featuring state-of-the-art neural models. For example, preliminary results with English data show that a FastSpeech2 model trained with 1 hour of training data can produce speech with comparable naturalness to a Tacotron2 model trained with 10 hours of data. Finally, we motivate future research in evaluation and classroom integration in the field of speech synthesis for language revitalization.

10:45-12:15 (Forum)
#84 Make the Best of Cross-lingual Transfer: Evidence from POS Tagging with over 100 Languages
Wietse de Vries, Martijn Wieling and Malvina Nissim
Cross-lingual transfer learning with large multilingual pre-trained models can be an effective approach for low-resource languages with no labeled training data. Existing evaluations of zero-shot cross-lingual generalisability of large pre-trained models use datasets with English training data, and test data in a selection of target languages. We explore a more extensive transfer learning setup with 65 different source languages and 105 target languages for part-of-speech tagging. Through our analysis, we show that pre-training of both source and target language, as well as matching language families, writing systems, word order systems, and lexical-phonetic distance significantly impact cross-lingual performance. The findings described in this paper can be used as indicators of which factors are important for effective zero-shot cross-lingual transfer to zero- and low-resource languages.

10:45-12:15 (Forum)
#85 Local Languages, Third Spaces, and other High-Resource Scenarios
Steven Bird
How can language technology address the diverse situations of the world's languages? In one view, languages exist on a resource continuum and the challenge is to scale existing solutions, bringing under-resourced languages into the high-resource world. In another view, presented here, the world's language ecology includes standardised languages, local languages, and contact languages. These are often subsumed under the label of "under-resourced languages" even though they have distinct functions and prospects. I explore this position and propose some ecologically-aware language technology agendas.

10:45-12:15 (Forum)
#86 From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology
Mark Dingemanse and Andreas Liesenfeld
Informal social interaction is the primordial home of human language. Linguistically diverse conversational corpora are an important and largely untapped resource for computational linguistics and language technology. Through the efforts of a worldwide language documentation movement, such corpora are increasingly becoming available. We show how interactional data from 63 languages (26 families) harbours insights about turn-taking, timing, sequential structure and social action, with implications for language technology, natural language understanding, and the design of conversational interfaces. Harnessing linguistically diverse conversational corpora will provide the empirical foundations for flexible, localizable, humane language technologies of the future.

10:45-12:15 (Forum)
#87 Challenges and Strategies in Cross-Cultural NLP

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust and Anders Søgaard
Various efforts in the Natural Language Processing (NLP) community have been made to accommodate linguistic diversity and serve speakers of many different languages. However, it is important to acknowledge that speakers and the content they produce and require, vary not just by language, but also by culture. Although language and culture are tightly linked, there are important differences. Analogous to cross-lingual and multilingual NLP, cross-cultural and multicultural NLP considers these differences in order to better serve users of NLP systems. We propose a principled framework to frame these efforts, and survey existing and potential strategies.

10:45-12:15 (Forum)
#88 Dataset Geography: Mapping Language Data to Language Users
Fahim Faisal, Yinkai Wang and Antonios Anastasopoulos
As language technologies become more ubiquitous, there are increasing efforts towards expanding the language diversity and coverage of natural language processing (NLP) systems. Arguably, the most important factor influencing the quality of modern NLP systems is data availability. In this work, we study the geographical representativeness of NLP datasets, aiming to quantify if and by how much do NLP datasets match the expected needs of the language speakers. In doing so, we use entity recognition and linking systems, also making important observations about their cross-lingual consistency and giving suggestions for more robust evaluation. Last, we explore some geographical and economic factors that may explain the observed dataset distributions.

10:45-12:15 (Forum)
#89 Systematic Inequalities in Language Technology Performance across the World's Languages
Damian E Blasi, Antonios Anastasopoulos and Graham Neubig
Natural language processing (NLP) systems have become a central technology in communication, education, medicine, artificial intelligence, and many other domains of research and development. While the performance of NLP methods has grown enormously over the last decade, this progress has been restricted to a minuscule subset of the world's ≈6,500 languages. We introduce a framework for estimating the global utility of language technologies as revealed in a comprehensive snapshot of recent publications in NLP. Our analyses involve the field at large, but also more in-depth studies on both user-facing technologies (machine translation, language understanding, question answering, text-to-speech synthesis) as well as foundational NLP tasks (dependency parsing, morphological inflection). In the process, we (1) quantify disparities in the current state of NLP research, (2) explore some of its associated societal and academic factors, and (3) produce tailored recommendations for evidence-based policy making aimed at promoting more global and equitable language technologies. Data and code to reproduce the findings discussed in this paper are available on GitHub (https://github.com/neubig/globalutility).

10:45-12:15 (Forum)
#90 Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism
Lane Schwartz
In this paper, we challenge the ACL community to reckon with historical and ongoing colonialism by adopting a set of ethical obligations and best practices drawn from the Indigenous studies literature. While the vast majority of NLP research focuses on a very small number of very high resource languages (English, Chinese, etc), some work has begun to engage with Indigenous languages. No research involving Indigenous language data can be considered ethical without first acknowledging that Indigenous languages are not merely very low resource languages. The toxic legacy of colonialism permeates every aspect of interaction between Indigenous communities and outside researchers. To this end, we propose that the ACL draft and adopt an ethical framework for NLP researchers and computational linguists wishing to engage in research involving Indigenous languages.

10:45-12:15 (Forum)
#91 AmericasNLI: Evaluating Zero-shot Natural Language Understanding of Pretrained Multilingual Models in Truly Low-resource Languages
Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo A. Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu and Katharina Kann
Pretrained multilingual models are able to perform cross-lingual transfer in a zero-shot setting, even for languages unseen during pretraining. However, prior work evaluating performance on unseen languages has largely been limited to low-level, syntactic tasks, and it remains unclear if zero-shot learning of high-level, semantic tasks is possible for unseen languages. To explore this question, we present AmericasNLI, an extension of XNLI (Conneau et al., 2018) to 10 Indigenous languages of the Americas. We conduct experiments with XLM-R, testing multiple zero-shot and translation-based approaches. Additionally, we explore model adaptation via continued pretraining and provide an analysis of the dataset by considering hypothesis-only models. We find that XLM-R's zero-shot performance is poor for all 10 languages, with an average performance of 38.48

10:45-12:15 (Forum)
#92 Dim Wihl Gat Tun: The Case for Linguistic Expertise in NLP for Under-Documented Languages
Clarissa Forbes, Farhan Samir, Bruce Harold Oliver, Changbing Yang, Edith Coates, Garrett Nicolai and Miikka Silfverberg
Recent progress in NLP is driven by pretrained models leveraging massive datasets and has predominantly benefited the world's political and economic superpowers. Technologically underserved languages are left behind because they lack such resources. Hundreds of underserved languages, nevertheless, have available data sources in the form of interlinear glossed text (IGT) from language documentation efforts. IGT remains underutilized in NLP work, perhaps because its annotations are only semi-structured and often language-specific. With this paper, we make the case that IGT data can be leveraged successfully provided that target language expertise is available. We specifically advocate for collaboration with documentary linguists. Our paper provides a roadmap for successful projects utilizing IGT data: (1) It is essential to define which NLP tasks can be accomplished with the given IGT data and how these will benefit the speech community. (2) Great care and target language expertise is required when converting the data into structured formats commonly employed in NLP. (3) Task-specific and user-specific evaluation can help to ascertain that the tools which are created benefit the target language speech community. We illustrate each step through a case study on developing a morphological reinflection system for the Tsimchianic language Gitksan.

10:45-12:15 (Forum)
#93 BPE vs. Morphological Segmentation: A Case Study on Machine Translation of Four Polysynthetic Languages
Manuel Mager, Arturo Oncevay, Elisabeth Mager, Katharina Kann and Thang Vu
Morphologically-rich polysynthetic languages present a challenge for NLP systems due to data sparsity, and a common strategy to handle this

issue is to apply subword segmentation. We investigate a wide variety of supervised and unsupervised morphological segmentation methods for four polysynthetic languages: Nahuatl, Raramuri, Shipibo-Konibo, and Wixarika. Then, we compare the morphologically inspired segmentation methods against Byte-Pair Encodings (BPEs) as inputs for machine translation (MT) when translating to and from Spanish. We show that for all language pairs except for Nahuatl, an unsupervised morphological segmentation algorithm outperforms BPEs consistently and that, although supervised methods achieve better segmentation scores, they under-perform in MT challenges. Finally, we contribute two new morphological segmentation datasets for Raramuri and Shipibo-Konibo, and a parallel corpus for Raramuri–Spanish.

10:45-12:15 (Forum)
#94 Pre-Trained Multilingual Sequence-to-Sequence Models: A Hope for Low-Resource Language Translation?
En-Shiun Annie Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Ifeoluwa Adelani, Ruisi Su and Arya D. McCarthy
What can pre-trained multilingual sequence-to-sequence models like mBART contribute to translating low-resource languages? We conduct a thorough empirical experiment in 10 languages to ascertain this, considering five factors: (1) the amount of fine-tuning data, (2) the noise in the fine-tuning data, (3) the amount of pre-training data in the model, (4) the impact of domain mismatch, and (5) language typology. In addition to yielding several heuristics, the experiments form a framework for evaluating the data sensitivities of machine translation systems. While mBART is robust to domain differences, its translations for unseen and typologically distant languages remain below 3.0 BLEU. In answer to our title's question, mBART is not a low-resource panacea; we therefore encourage shifting the emphasis from new models to new data.

10:45-12:15 (Forum)
#95 Morphological Processing of Low-Resource Languages: Where We Are and What's Next
Adam Wiemerslage, Miikka Silfverberg, Changbing Yang, Arya D. McCarthy, Garrett Nicolai, Eliana Colunga and Katharina Kann
Automatic morphological processing can aid downstream natural language processing applications, especially for low-resource languages, and assist language documentation efforts for endangered languages. Having long been multilingual, the field of computational morphology is increasingly moving towards approaches suitable for languages with minimal or no annotated resources. First, we survey recent developments in computational morphology with a focus on low-resource languages. Second, we argue that the field is ready to tackle the logical next challenge: understanding a language's morphology from raw text alone. We perform an empirical study on a truly unsupervised version of the paradigm completion task and show that, while existing state-of-the-art models bridged by two newly proposed models we devise perform reasonably, there is still much room for improvement. The stakes are high: solving this task will increase the language coverage of morphological resources by a number of magnitudes.

10:45-12:15 (Forum)
#96 Toward More Meaningful Resources for Lower-resourced Languages
Constantine Lignos, Nolan Holley, Chester Palen-Michel and Jonne Sälevä
In this position paper, we describe our perspective on how meaningful resources for lower-resourced languages should be developed in connection with the speakers of those languages. Before advancing that position, we first examine two massively multilingual resources used in language technology development, identifying shortcomings that limit their usefulness. We explore the contents of the names stored in Wikidata for a few lower-resourced languages and find that many of them are not in fact in the languages they claim to be, requiring non-trivial effort to correct. We discuss quality issues present in WikiAnn and evaluate whether it is a useful supplement to hand-annotated data. We then discuss the importance of creating annotations for lower-resourced languages in a thoughtful and ethical way that includes the language speakers as part of the development process. We conclude with recommended guidelines for resource development.

10:45-12:15 (Forum)
#97 OCR Improves Machine Translation for Low-Resource Languages
Oana Ignat, Jean Maillard, Vishrav Chaudhary and Francisco Guzmán
We aim to investigate the performance of current OCR systems on low resource languages and low resource scripts. We introduce and make publicly available a novel benchmark, OCR4MT, consisting of real and synthetic data, enriched with noise, for 60 low-resource languages in low resource scripts. We evaluate state-of-the-art OCR systems on our benchmark and analyse most common errors. We show that OCR monolingual data is a valuable resource that can increase performance of Machine Translation models, when used in backtranslation. We then perform an ablation study to investigate how OCR errors impact Machine Translation performance and determine what is the minimum level of OCR quality needed for the monolingual data to be useful for Machine Translation.

Lunch Break
12:15-13:30 - **Auditorium** (Lunch is not served)

Session 7 - 13:30-14:30

Language Grounding, Speech and Multimodality 3
13:30-14:30 (Liffey Hall 1)

13:30-13:45 (Liffey Hall 1)
Understanding Multimodal Procedural Knowledge by Sequencing Multimodal Instructional Manuals
Te-Lin Wu, Alex Spangher, Pegah Alipoormolabashi, Marjorie Freedman, Ralph M. Weischedel and Nanyun Peng
The ability to sequence unordered events is evidence of comprehension and reasoning about real world tasks/procedures. It is essential for applications such as task planning and multi-source instruction summarization. It often requires thorough understanding of temporal common sense and multimodal information, since these procedures are often conveyed by a combination of texts and images. While humans are capable of reasoning about and sequencing unordered procedural instructions, the extent to which the current machine learning methods possess such capability is still an open question. In this work, we benchmark models' capability of reasoning over and sequencing unordered multimodal instructions by curating datasets from online instructional manuals and collecting comprehensive human annotations. We find current state-of-the-art models not only perform significantly worse than humans but also seem incapable of efficiently utilizing multimodal information. To improve machines' performance on multimodal event sequencing, we propose sequence-aware pretraining techniques exploiting the sequential alignment properties of both texts and images, resulting in > 5

13:45-14:00 (Liffey Hall 1)
[TACL] **Word Representation Learning in Multimodal Pre-Trained Transformers: An Intrinsic Evaluation**
Sandro Pezzelle, Ece Takmaz and Raquel Fernández

14:00-14:15 (Liffey Hall 1)
There's a Time and Place for Reasoning Beyond the Image
Xingyu Fu, Ben Zhou, Ishaan Preetam Chandratreya, Carl Vondrick and Dan Roth
Images are often more significant than only the pixels to human eyes, as we can infer, associate, and reason with contextual information from other sources to establish a more complete picture. For example, in Figure 1, we can find a way to identify the news articles related to the picture through segment-wise understandings of the signs, the buildings, the crowds, and more. This reasoning could provide the time and place the image was taken, which will help us in subsequent tasks, such as automatic storyline construction, correction of image source in intended effect photographs, and upper-stream processing such as image clustering for certain location or time.
In this work, we formulate this problem and introduce TARA: a dataset with 16k images with their associated news, time, and location, automatically extracted from New York Times, and an additional 61k examples as distant supervision from WIT. On top of the extractions, we present a crowdsourced subset in which we believe it is possible to find the images' spatio-temporal information for evaluation purpose. We show that there exists a 70% gap between a state-of-the-art joint model and human performance, which is slightly filled by our proposed model that uses segment-wise reasoning, motivating higher-level vision-language joint models that can conduct open-ended reasoning with world knowledge. The data and code are publicly available at https://github.com/zeyofu/TARA.

14:15-14:25 (Liffey Hall 1)
Voxel-informed Language Grounding
Rodolfo Corona, Shizhan Zhu, Dan Klein and Trevor Darrell
Natural language applied to natural 2D images describes a fundamentally 3D world. We present the Voxel-informed Language Grounder (VLG), a language grounding model that leverages 3D geometric information in the form of voxel maps derived from the visual input using a volumetric reconstruction model. We show that VLG significantly improves grounding accuracy on SNARE, an object reference game task. At the time of writing, VLG holds the top place on the SNARE leaderboard, achieving SOTA results with a 2.0

Machine Learning for NLP 5

13:30-14:30 (The Liffey B)

13:30-13:45 (The Liffey B)
[TACL] **CANINE: Pre-training an Efficient Tokenization-Free Encoder for Language Representation**
Jonathan Clark, Dan Garrette, Iulia Turc and John Wieting

13:45-14:00 (The Liffey B)
Word2Box: Capturing Set-Theoretic Semantics of Words using Box Embeddings
Shib Sankar Dasgupta, Michael Boratko, Siddhartha Mishra, Shriya Atmakuri, Dhruvesh Patel, Xiang Lorraine Li and Andrew McCallum
Learning representations of words in a continuous space is perhaps the most fundamental task in NLP, however words interact in ways much richer than vector dot product similarity can provide. Many relationships between words can be expressed set-theoretically, for example, adjective-noun compounds (eg. "red cars"$\subseteq$"cars") and homographs (eg. "tongue"$\cap$"body" should be similar to "mouth", while "tongue"$\cap$"language" should be similar to "dialect") have natural set-theoretic interpretations. Box embeddings are a novel region-based representation which provide the capability to perform these set-theoretic operations. In this work, we provide a fuzzy-set interpretation of box embeddings, and learn box representations of words using a set-theoretic training objective. We demonstrate improved performance on various word similarity tasks, particularly on less common words, and perform a quantitative and qualitative analysis exploring the additional unique expressivity provided by Word2Box.

14:00-14:15 (The Liffey B)
[TACL] **Towards General Natural Language Understanding with Probabilistic Worldbuilding**
Abulhair Saparov and Tom Mitchell

14:15-14:30 (The Liffey B)
Adapting Coreference Resolution Models through Active Learning
Michelle Yuan, Patrick Xia, Chandler May, Benjamin Van Durme and Jordan Lee Boyd-Graber
Neural coreference resolution models trained on one dataset may not transfer to new, low-resource domains. Active learning mitigates this problem by sampling a small subset of data for annotators to label. While active learning is well-defined for classification tasks, its application to coreference resolution is neither well-defined nor fully understood. This paper explores how to actively label coreference, examining sources of model uncertainty and document reading costs. We compare uncertainty sampling strategies and their advantages through thorough error analysis. In both synthetic and human experiments, labeling spans within the same document is more effective than annotating spans across documents. The findings contribute to a more realistic development of coreference resolution models.

Machine Translation and Multilinguality 4

13:30-14:30 (The Liffey A)

13:30-13:45 (The Liffey A)
Language-agnostic BERT Sentence Embedding

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan and Wei Wang
While BERT is an effective method for learning monolingual sentence embeddings for semantic similarity and embedding based transfer learning BERT based cross-lingual sentence embeddings have yet to be explored. We systematically investigate methods for learning multilingual sentence embeddings by combining the best methods for learning monolingual and cross-lingual representations including: masked language modeling (MLM), translation language modeling (TLM), dual encoder translation ranking, and additive margin softmax. We show that introducing a pre-trained multilingual language model dramatically reduces the amount of parallel training data required to achieve good performance by 80dev/google/LaBSE.

13:45-14:00 (The Liffey A)
Overlap-based Vocabulary Generation Improves Cross-lingual Transfer Among Related Languages
Vaidehi Patil, Partha Talukdar and Sunita Sarawagi
Pre-trained multilingual language models such as mBERT and XLM-R have demonstrated great potential for zero-shot cross-lingual transfer to low web-resource languages (LRL). However, due to limited model capacity, the large difference in the sizes of available monolingual corpora between high web-resource languages (HRL) and LRLs does not provide enough scope of co-embedding the LRL with the HRL, thereby affecting the downstream task performance of LRLs. In this paper, we argue that relatedness among languages in a language family along the dimension of lexical overlap may be leveraged to overcome some of the corpora limitations of LRLs. We propose Overlap BPE (OBPE), a simple yet effective modification to the BPE vocabulary generation algorithm which enhances overlap across related languages. Through extensive experiments on multiple NLP tasks and datasets, we observe that OBPE generates a vocabulary that increases the representation of LRLs via tokens shared with HRLs. This results in improved zero-shot transfer from related HRLs to LRLs without reducing HRL representation and accuracy. Unlike previous studies that dismissed the importance of token-overlap, we show that in the low-resource related language setting, token overlap matters. Synthetically reducing the overlap to zero can cause as much as a four-fold drop in zero-shot transfer accuracy.

14:00-14:15 (The Liffey A)
[TACL] **Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets**
Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Rubungo Niyongabo, Toan Nguyen, Mathias Müller, André Müller, Shamsuddeen Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal and Mofetoluwa Adeyemi

14:15-14:30 (The Liffey A)
[TACL] **Samanantar: The Largest Publicly Available Parallel Corpora Collection For 11 Indic Languages**
Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh Khapra and Srihari Nagaraj

Resources and Evaluation 4

13:30-14:30 (Liffey Hall 2)

13:30-13:45 (Liffey Hall 2)
NumGLUE: A Suite of Fundamental yet Challenging Mathematical Reasoning Tasks
Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Singh Sachdeva, Peter Clark, Chitta Baral and Ashwin Kalyan
Given the ubiquitous nature of numbers in text, reasoning with numbers to perform simple calculations is an important skill of AI systems. While many datasets and models have been developed to this end, state-of-the-art AI systems are brittle; failing to perform the underlying mathematical reasoning when they appear in a slightly different scenario. Drawing inspiration from GLUE that was proposed in the context of natural language understanding, we propose NumGLUE, a multi-task benchmark that evaluates the performance of AI systems on eight different tasks, that at their core require simple arithmetic understanding. We show that this benchmark is far from being solved with neural models including state-of-the-art large-scale language models performing significantly worse than humans (lower by 46.4

13:45-14:00 (Liffey Hall 2)
LexGLUE: A Benchmark Dataset for Legal Language Understanding in English
Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz and Nikolaos Aletras
Laws and their interpretations, legal arguments and agreements are typically expressed in writing, leading to the production of vast corpora of legal text. Their analysis, which is at the center of legal practice, becomes increasingly elaborate as these collections grow in size. Natural language understanding (NLU) technologies can be a valuable tool to support legal practitioners in these endeavors. Their usefulness, however, largely depends on whether current state-of-the-art models can generalize across various tasks in the legal domain. To answer this currently open question, we introduce the Legal General Language Understanding Evaluation (LexGLUE) benchmark, a collection of datasets for evaluating model performance across a diverse set of legal NLU tasks in a standardized way. We also provide an evaluation and analysis of several generic and legal-oriented models demonstrating that the latter consistently offer performance improvements across multiple tasks.

14:00-14:15 (Liffey Hall 2)
IAM: A Comprehensive and Large-Scale Dataset for Integrated Argument Mining Tasks
Liying Cheng, Lidong Bing, Ruidan He, Qian Yu, Yan Zhang and Luo Si
Traditionally, a debate usually requires a manual preparation process, including reading plenty of articles, selecting the claims, identifying the stances of the claims, seeking the evidence for the claims, etc. As the AI debate attracts more attention these years, it is worth exploring the methods to automate the tedious process involved in the debating system. In this work, we introduce a comprehensive and large dataset named IAM, which can be applied to a series of argument mining tasks, including claim extraction, stance classification, evidence extraction, etc. Our dataset is collected from over 1k articles related to 123 topics. Near 70k sentences in the dataset are fully annotated based on their

argument properties (e.g., claims, stances, evidence, etc.). We further propose two new integrated argument mining tasks associated with the debate preparation process: (1) claim extraction with stance classification (CESC) and (2) claim-evidence pair extraction (CEPE). We adopt a pipeline approach and an end-to-end method for each integrated task separately. Promising experimental results are reported to show the values and challenges of our proposed tasks, and motivate future research on argument mining.

14:15-14:25 (Liffey Hall 2)
CoDA21: Evaluating Language Understanding Capabilities of NLP Models With Context-Definition Alignment
Lütfi Kerem Senel, Timo Schick and Hinrich Schuetze
Pretrained language models (PLMs) have achieved superhuman performance on many benchmarks, creating a need for harder tasks. We introduce CoDA21 (Context Definition Alignment), a challenging benchmark that measures natural language understanding (NLU) capabilities of PLMs: Given a definition and a context each for k words, but not the words themselves, the task is to align the k definitions with the k contexts. CoDA21 requires a deep understanding of contexts and definitions, including complex inference and world knowledge. We find that there is a large gap between human and PLM performance, suggesting that CoDA21 measures an aspect of NLU that is not sufficiently covered in existing benchmarks.

Question Answering 3

13:30-14:30 (Wicklow Hall 1)

13:30-13:45 (Wicklow Hall 1)
Automated Crossword Solving
Eric Wallace, Nicholas Tomlin, Albert Xu, Kevin Yang, Eshaan Pathak, Matthew L. Ginsberg and Dan Klein
We present the Berkeley Crossword Solver, a state-of-the-art approach for automatically solving crossword puzzles. Our system works by generating answer candidates for each crossword clue using neural question answering models and then combines loopy belief propagation with local search to find full puzzle solutions. Compared to existing approaches, our system improves exact puzzle accuracy from 57

13:45-14:00 (Wicklow Hall 1)
QAConv: Question Answering on Informative Conversations
Chien-Sheng Wu, Andrea Madotto, Wenhao Liu, Pascale Fung and Caiming Xiong
This paper introduces QAConv, a new question answering (QA) dataset that uses conversations as a knowledge source. We focus on informative conversations, including business emails, panel discussions, and work channels. Unlike open-domain and task-oriented dialogues, these conversations are usually long, complex, asynchronous, and involve strong domain knowledge. In total, we collect 34,608 QA pairs from 10,259 selected conversations with both human-written and machine-generated questions. We use a question generator and a dialogue summarizer as auxiliary tools to collect and recommend questions. The dataset has two testing scenarios: chunk mode and full mode, depending on whether the grounded partial conversation is provided or retrieved. Experimental results show that state-of-the-art pretrained QA systems have limited zero-shot performance and tend to predict our questions as unanswerable. Our dataset provides a new training and evaluation testbed to facilitate QA on conversations research.

14:00-14:15 (Wicklow Hall 1)
Answer-level Calibration for Free-form Multiple Choice Question Answering
Sawan Kumar
Pre-trained language models have recently shown that training on large corpora using the language modeling objective enables few-shot and zero-shot capabilities on a variety of NLP tasks, including commonsense reasoning tasks. This is achieved using text interactions with the model, usually by posing the task as a natural language text completion problem. While using language model probabilities to obtain task specific scores has been generally useful, it often requires task-specific heuristics such as length normalization, or probability calibration. In this work, we consider the question answering format, where we need to choose from a set of (free-form) textual choices of unspecified lengths given a context. We present ALC (Answer-Level Calibration), where our main suggestion is to model context-independent biases in terms of the probability of a choice without the associated context and to subsequently remove it using an unsupervised estimate of similarity with the full context. We show that our unsupervised answer-level calibration consistently improves over or is competitive with baselines using standard evaluation metrics on a variety of tasks including commonsense reasoning tasks. Further, we show that popular datasets potentially favor models biased towards easy cues which are available independent of the context. We analyze such biases using an associated F1-score. Our analysis indicates that answer-level calibration is able to remove such biases and leads to a more robust measure of model capability.

14:15-14:25 (Wicklow Hall 1)
Predicting Difficulty and Discrimination of Natural Language Questions
Matthew Alexander Byrd and Shashank Srivastava
Item Response Theory (IRT) has been extensively used to numerically characterize question difficulty and discrimination for human subjects in domains including cognitive psychology and education (Primi et al., 2014; Downing, 2003). More recently, IRT has been used to similarly characterize item difficulty and discrimination for natural language models across various datasets (Lalor et al., 2019; Vania et al., 2021; Rodriguez et al., 2021). In this work, we explore predictive models for directly estimating and explaining these traits for natural language questions in a question-answering context. We use HotpotQA for illustration. Our experiments show that it is possible to predict both difficulty and discrimination parameters for new questions, and these traits are correlated with features of questions, answers, and associated contexts. Our findings can have significant implications for the creation of new datasets and tests on the one hand and strategies such as active learning and curriculum learning on the other.

Summarization 2

13:30-14:30 (Wicklow Hall 2a)

13:30-13:45 (Wicklow Hall 2a)

[TACL] **SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization**
Philippe Laban, Tobias Schnabel, Paul Bennett and Marti Hearst

13:45-14:00 (Wicklow Hall 2a)
ASPECTNEWS: Aspect-Oriented Summarization of News Documents
Ojas Ahuja, Jiacheng Xu, Akshay Kumar Gupta, Kevin Horecka and Greg Durrett
Generic summaries try to cover an entire document and query-based summaries try to answer document-specific questions. But real users' needs often fall in between these extremes and correspond to aspects, high-level topics discussed among similar types of documents. In this paper, we collect a dataset of realistic aspect-oriented summaries, AspectNews, which covers different subtopics about articles in news sub-domains. We annotate data across two domains of articles, earthquakes and fraud investigations, where each article is annotated with two distinct summaries focusing on different aspects for each domain. A system producing a single generic summary cannot concisely satisfy both aspects. Our focus in evaluation is how well existing techniques can generalize to these domains without seeing in-domain training data, so we turn to techniques to construct synthetic training data that have been used in query-focused summarization work. We compare several training schemes that differ in how strongly keywords are used and how oracle summaries are extracted. Our evaluation shows that our final approach yields (a) focused summaries, better than those from a generic summarization system or from keyword matching; (b) a system sensitive to the choice of keywords.

14:00-14:15 (Wicklow Hall 2a)
A Well-Composed Text is Half Done! Composition Sampling for Diverse Conditional Generation
Shashi Narayan, Gonçalo Simões, Yao Zhao, Joshua Maynez, Dipanjan Das, Michael Collins and Mirella Lapata
We propose Composition Sampling, a simple but effective method to generate diverse outputs for conditional generation of higher quality compared to previous stochastic decoding strategies. It builds on recently proposed plan-based neural generation models (FROST, Narayan et al, 2021) that are trained to first create a composition of the output and then generate by conditioning on it and the input. Our approach avoids text degeneration by first sampling a composition in the form of an entity chain and then using beam search to generate the best possible text grounded to this entity chain. Experiments on summarization (CNN/DailyMail and XSum) and question generation (SQuAD), using existing and newly proposed automatic metrics together with human-based evaluation, demonstrate that Composition Sampling is currently the best available decoding strategy for generating diverse meaningful outputs.

14:15-14:30 (Wicklow Hall 2a)
BRIO: Bringing Order to Abstractive Summarization
Yixin Liu, Pengfei Liu, Dragomir Radev and Graham Neubig
Abstractive summarization models are commonly trained using maximum likelihood estimation, which assumes a deterministic (one-point) target distribution in which an ideal model will assign all the probability mass to the reference summary. This assumption may lead to performance degradation during inference, where the model needs to compare several system-generated (candidate) summaries that have deviated from the reference summary. To address this problem, we propose a novel training paradigm which assumes a non-deterministic distribution so that different candidate summaries are assigned probability mass according to their quality. Our method achieves a new state-of-the-art result on the CNN/DailyMail (47.78 ROUGE-1) and XSum (49.07 ROUGE-1) datasets. Further analysis also shows that our model can estimate probabilities of candidate summaries that are more correlated with their level of quality.

Sentiment Analysis, Stylistic Analysis, and Argument Mining 2

13:30-14:30 (Wicklow Hall 2b)

13:30-13:45 (Wicklow Hall 2b)
[CL] **Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis**
Saif M. Mohammad

13:45-14:00 (Wicklow Hall 2b)
A Rationale-Centric Framework for Human-in-the-loop Machine Learning
Jinghui Lu, Linyi Yang, Brian Mac Namee and Yue Zhang
We present a novel rational-centric framework with human-in-the-loop – Rationales-centric Double-robustness Learning (RDL) – to boost model out-of-distribution performance in few-shot learning scenarios. By using static semi-factual generation and dynamic human-intervened correction, RDL, acting like a sensible "inductive bias", exploits rationales (i.e. phrases that cause the prediction), human interventions and semi-factual augmentations to decouple spurious associations and bias models towards generally applicable underlying distributions, which enables fast and accurate generalisation. Experimental results show that RDL leads to significant prediction benefits on both in-distribution and out-of-distribution tests, especially for few-shot learning scenarios, compared to many state-of-the-art benchmarks. We also perform extensive ablation studies to support in-depth analyses of each component in our framework.

14:00-14:15 (Wicklow Hall 2b)
DoCoGen: Domain Counterfactual Generation for Low Resource Domain Adaptation
Nitay Calderon, Eyal Ben-David, Amir Feder and Roi Reichart
Natural language processing (NLP) algorithms have become very successful, but they still struggle when applied to out-of-distribution examples. In this paper we propose a controllable generation approach in order to deal with this domain adaptation (DA) challenge. Given an input text example, our DoCoGen algorithm generates a domain-counterfactual textual example (D-con) - that is similar to the original in all aspects, including the task label, but its domain is changed to a desired one. Importantly, DoCoGen is trained using only unlabeled examples from multiple domains - no NLP task labels or parallel pairs of textual examples and their domain-counterfactuals are required. We show that DoCoGen can generate coherent counterfactuals consisting of multiple sentences. We use the D-cons generated by DoCoGen to augment a sentiment classifier and a multi-label intent classifier in 20 and 78 DA setups, respectively, where source-domain labeled data is scarce. Our model outperforms strong baselines and improves the accuracy of a state-of-the-art unsupervised DA algorithm.

14:15-14:30 (Wicklow Hall 2b)
So Different Yet So Alike! Constrained Unsupervised Text Style Transfer
Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, Roger Zimmermann and Soujanya Poria
Automatic transfer of text between domains has become popular in recent times. One of its aims is to preserve the semantic content while adapting to the target domain. However, it does not explicitly maintain other attributes between the source and translated text: e.g., text length and descriptiveness. Maintaining constraints in transfer has several downstream applications, including data augmentation and debiasing. We introduce a method for such constrained unsupervised text style transfer by introducing two complementary losses to the generative adversarial network (GAN) family of models. Unlike the competing losses used in GANs, we introduce cooperative losses where the discriminator and the generator cooperate and reduce the same loss. The first is a contrastive loss and the second is a classification loss — aiming to regularize the latent space further and bring similar sentences closer together. We demonstrate that such training retains lexical, syntactic and domain-specific constraints between domains for multiple benchmark datasets, including ones where more than one attribute change. We show that the complementary cooperative losses improve text quality, according to both automated and human evaluation measures.

Mini Break
14:30-14:45 - **Auditorium** (Forum)

Best Paper Awards
14:45-16:00 - **Auditorium** (Auditorium)

Coffee Break
16:00-16:30 - **Auditorium** (Forum)

ACL Awards
16:30-18:00 - **Auditorium** (Auditorium)

Closing Session
18:00-18:30 - **Auditorium** (Auditorium)

Virtual Poster Session 4 - 19:00-20:00

VPS4: Computational Social Science and Cultural Analytics
19:00-20:00 (GatherTown)

Long: Which side are you on? Insider-Outsider classification in conspiracy-theoretic social media
Speaker: Pavan Holur

Findings: EnCBP: A New Benchmark Dataset for Finer-Grained Cultural Background Prediction in English
Speaker: Weicheng Ma

Findings: From Stance to Concern: Adaptation of Propositional Analysis to New Tasks and Domains
Speaker: Brodie Mather

SRW: A Checkpoint on Multilingual Misogyny Identification
Speaker: Arianna Muti

VPS4: Dialogue and Interactive Systems
19:00-20:00 (GatherTown)

Long: [CASPI] Causal-aware Safe Policy Improvement for Task-oriented Dialogue
Speaker: Govardana Sachithanandam Ramachandran

Long: New Intent Discovery with Pre-training and Contrastive Learning
Speaker: Yuwei Zhang

Long: ProphetChat: Enhancing Dialogue Generation with Simulation of Future Conversation
Speaker: Chang Liu

Long: Think Before You Speak: Explicitly Generating Implicit Commonsense Knowledge for Response Generation
Speaker: Pei Zhou

Outstanding Paper: Online Semantic Parsing for Latency Reduction in Task-Oriented Dialogue
Speaker: Jiawei Zhou

Findings: C^3KG: A Chinese Commonsense Conversation Knowledge Graph
Speaker: Dawei Li

Short: Investigating person-specific errors in chat-oriented dialogue systems
Speaker: Koh Mitsuda

Findings: Long Time No See! Open-Domain Conversation with Long-Term Persona Memory
Speaker: Xinchao Xu

Long: Beyond Goldfish Memory: Long-Term Open-Domain Conversation
Speaker: Jing Xu

Findings: Selecting Stickers in Open-Domain Dialogue through Multitask Learning
Speaker: Zhexin Zhang

Short: Probing the Robustness of Trained Metrics for Conversational Dialogue Systems
Speaker: Jan Deriu

Long: Lexical Knowledge Internalization for Neural Dialog Generation
Speaker: Zhiyong Wu

Long: A Model-agnostic Data Manipulation Method for Persona-based Dialogue Generation
Speaker: Yu Cao

VPS4: Discourse and Pragmatics & Ethics in NLP

19:00-20:00 (GatherTown)

Short: Predicting Sentence Deletions for Text Simplification Using a Functional Discourse Structure
Speaker: Bohan Zhang

Long: Learning to Mediate Disparities Towards Pragmatic Communication
Speaker: Yuwei Bao

Long: How Do We Answer Complex Questions: Discourse Structure of Long-form Answers
Speaker: Fangyuan Xu

SRW: Towards Unification of Discourse Annotation Frameworks
Speaker: Yingxue Fu

VPS4: Ethics in NLP

19:00-20:00 (GatherTown)

Findings: Mitigating Gender Bias in Distilled Language Models via Counterfactual Role Reversal
Speaker: Umang Gupta

Findings: Learning Bias-reduced Word Embeddings Using Dictionary Definitions
Speaker: Haozhe An

Long: ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection
Speaker: Thomas Hartvigsen

Long: Upstream Mitigation Is extitNot All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models
Speaker: Ryan Steed

Findings: Your fairness may vary: Pretrained language model fairness in toxic text classification
Speaker: Ioana Baldini

Findings: Square One Bias in NLP: Towards a Multi-Dimensional Exploration of the Research Manifold
Speaker: Sebastian Ruder

Long: Reinforcement Guided Multi-Task Learning Framework for Low-Resource Stereotype Detection
Speaker: Rajkumar Pujari

SRW: Ethical Considerations for Low-resourced Machine Translation
Speaker: Levon Haroutunian

VPS4: Generation

19:00-20:00 (GatherTown)

Long: Mix and Match: Learning-free Controllable Text Generationusing Energy Language Models
Speaker: Fatemehsadat Mireshghalla

Long: Explanation Graph Generation via Pre-trained Language Models: An Empirical Study with Contrastive Learning
Speaker: Swarnadeep Saha

Long: Keywords and Instances: A Hierarchical Contrastive Learning Framework Unifying Hybrid Granularities for Text Generation
Speaker: Mingzhe Li

Short: A Recipe for Arbitrary Text Style Transfer with Large Language Models
Speaker: Emily Reif

VPS4: Information Extraction

19:00-20:00 (GatherTown)

Long: Domain Adaptation in Multilingual and Multi-Domain Monolingual Settings for Complex Word Identification
Speaker: George-Eduard Zaharia

Findings: Eider: Empowering Document-level Relation Extraction with Efficient Evidence Extraction and Inference-stage Fusion
Speaker: Yiqing Xie

Long: Multilingual Knowledge Graph Completion with Self-Supervised Adaptive Graph Alignment
Speaker: Zijie Huang

Long: De-Bias for Generative Extraction in Unified NER Task
Speaker: Shuai Zhang

Long: Text-to-Table: A New Way of Information Extraction
Speaker: Xueqing Wu

Findings: Document-Level Event Argument Extraction via Optimal Transport
Speaker: Amir Pouran Ben Veyseh

Long: A Meta-framework for Spatiotemporal Quantity Extraction from Text
Speaker: Qiang Ning

Long: Continual Few-shot Relation Learning via Embedding Space Regularization and Data Augmentation
Speaker: Chengwei Qin

Findings: Leveraging Expert Guided Adversarial Augmentation For Improving Generalization in Named Entity Recognition
Speaker: Aaron Reich

Findings: Detection, Disambiguation, Re-ranking: Autoregressive Entity Linking as a Multi-Task Problem
Speaker: Khalil Mrini

Long: FormNet: Structural Encoding beyond Sequential Modeling in Form Document Information Extraction
Speaker: Chen-Yu Lee

Long: Automatic Error Analysis for Document-level Information Extraction
Speaker: Aliva Das

Findings: Learn and Review: Enhancing Continual Named Entity Recognition via Reviewing Synthetic Samples
Speaker: Yu Xia

Findings: Deep Reinforcement Learning for Entity Alignment
Speaker: Lingbing Guo

Long: MINER: Improving Out-of-Vocabulary Named Entity Recognition from an Information Theoretic Perspective
Speaker: Xiao Wang

Long: Does Recommend-Revise Produce Reliable Annotations? An Analysis on Missing Instances in DocRED
Speaker: Quzhe Huang

Findings: Cross-Lingual UMLS Named Entity Linking using UMLS Dictionary Fine-Tuning
Speaker: Rina Galprein

Findings: Consistent Representation Learning for Continual Relation Extraction
Speaker: Kang Zhao

Short: DiS-ReX: A Multilingual Dataset for Distantly Supervised Relation Extraction
Speaker: Abhyuday Bhartiya

TACL: VILA: Improving Structured Content Extraction from Scientific PDFs Using Visual Layout Groups
Speaker: Zejiang Shen

SRW: GNNer: Reducing Overlapping in Span-based NER Using Graph Neural Networks
Speaker: Urchade Zaratiana

VPS4: Information Retrieval and Text Mining

19:00-20:00 (GatherTown)

Long: Compact Token Representations with Contextual Quantization for Efficient Document Re-ranking
Speaker: Yingrui Yang

Findings: Compressing Sentence Representation for Semantic Retrieval via Homomorphic Projective Distillation
Speaker: Xuandong Zhao

Long: UCTopic: Unsupervised Contrastive Learning for Phrase Representations and Topic Mining
Speaker: Jiacheng Li

Findings: Two Birds with One Stone: Unified Model Learning for Both Recall and Ranking in News Recommendation
Speaker: Chuhan Wu

Findings: ED2LM: Encoder-Decoder to Language Model for Faster Document Re-ranking Inference
Speaker: Kai Hui

VPS4: Interpretability and Analysis of Models for NLP

19:00-20:00 (GatherTown)

Findings: "Is Whole Word Masking Always Better for Chinese BERT?": Probing on Chinese Grammatical Error Correction
Speaker: Yong Dai

Findings: AbductionRules: Training Transformers to Explain Unexpected Inputs
Speaker: Nathan Young

Short: Are Shortest Rationales the Best Explanations for Human Understanding?
Speaker: Hua Shen

Findings: Visualizing the Relationship Between Encoded Linguistic Information and Task Performance
Speaker: Jiannan Xiang

Long: A Closer Look at How Fine-tuning Changes BERT
Speaker: Yichu Zhou

Findings: Explaining Classes through Stable Word Attributions
Speaker: Samuel Rönnqvist

Long: ProtoTEx: Explaining Model Decisions with Prototype Tensors
Speaker: Anubrata Das

Long: Life after BERT: What do Other Muppets Understand about Language?
Speaker: Vladislav Lialin

Short: Problems with Cosine as a Measure of Embedding Similarity for High Frequency Words
Speaker: Kaitlyn Zhou

Findings: extitGeneralized but not Robust? Comparing the Effects of Data Modification Methods on Out-of-Domain Generalization and Adversarial Robustness
Speaker: Tejas Gokhale

Long: Low-Rank Softmax Can Have Unargmaxable Classes in Theory but Rarely in Practice
Speaker: Andreas Grivas

Short: Efficient Classification of Long Documents Using Transformers
Speaker: Hyunji Hayley Park

Findings: Discontinuous Constituency and BERT: A Case Study of Dutch
Speaker: Konstantinos Kogkalidis

Long: Overcoming a Theoretical Limitation of Self-Attention
Speaker: David Chiang

Findings: Interpreting the Robustness of Neural NLP Models to Textual Perturbations
Speaker: Yunxiang Zhang

Findings: On the data requirements of probing
Speaker: Zining Zhu

VPS4: Language Groundings, Speech and Multimodality

19:00-20:00 (GatherTown)

Findings: Debiasing Event Understanding for Visual Commonsense Tasks
Speaker: Minji Seo

Long: Unified Speech-Text Pre-training for Speech Translation and Recognition
Speaker: Yun Tang

Findings: Semantically Distributed Robust Optimization for Vision-and-Language Inference
Speaker: Tejas Gokhale

Long: Direct Speech-to-Speech Translation With Discrete Units
Speaker: Ann Lee

Findings: Enabling Multimodal Generation on CLIP via Vision-Language Knowledge Distillation
Speaker: Wenliang Dai

Long: Leveraging Unimodal Self-Supervised Learning for Multimodal Audio-Visual Speech Recognition
Speaker: Xichen Pan

Long: Understanding Multimodal Procedural Knowledge by Sequencing Multimodal Instructional Manuals
Speaker: Te-Lin Wu

Findings: Modeling Intensification for Sign Language Generation: A Computational Approach
Speaker: Mert Inan

Long: Language-Agnostic Meta-Learning for Low-Resource Text-to-Speech with Articulatory Features
Speaker: Florian Lux

Findings: Prior Knowledge and Memory Enriched Transformer for Sign Language Translation
Speaker: Tao Jin

Long: Self-supervised Semantic-driven Phoneme Discovery for Zero-resource Speech Recognition
Speaker: Liming Wang

SRW: Combine to Describe: Evaluating Compositional Generalization in Image Captioning
Speaker: George Pantazopoulos

SRW: What do Models Learn From Training on More Than Text? Measuring Visual Commonsense Knowledge
Speaker: Lovisa Hagström

On the Effects of Video Grounding on Language Models
Speaker: nan

SRW: Explicit Object Relation Alignment for Vision and Language Navigation
Speaker: Yue Zhang

SRW: Discourse on ASR Measurement: Introducing the ARPOCA Assessment Tool
Speaker: Megan Merz

VPS4: Linguistic Theories, Cognitive Modeling and Psycholinguistics

19:00-20:00 (GatherTown)

Long: GPT-D: Inducing Dementia-related Linguistic Anomalies by Deliberate Degradation of Artificial Neural Language Models
Speaker: Changye Li

Long: Speaker Information Can Guide Models to Better Inductive Biases: A Case Study On Predicting Code-Switching
Speaker: Alissa Ostapenko

Long: Neural reality of argument structure constructions
Speaker: Bai Li

Long: Flexible Generation from Fragmentary Linguistic Input
Speaker: Peng Qian

CL: Assessing corpus evidence for formal and psycholinguistic constraints on nonprojectivity
Speaker: Himanshu Yadav

VPS4: Machine Learning for NLP

19:00-20:00 (GatherTown)

Short: BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models
Speaker: Elad Ben Zaken

Long: Multi-Granularity Structural Knowledge Distillation for Language Model Compression
Speaker: Chang Liu

Findings: Virtual Augmentation Supported Contrastive Learning of Sentence Representations
Speaker: Dejiao Zhang

Long: Better Language Model with Hypernym Class Prediction
Speaker: He Bai

Findings: Learning Adaptive Axis Attentions in Fine-tuning: Beyond Fixed Sparse Attention Patterns
Speaker: Zihan Wang

Short: DQ-BART: Efficient Sequence-to-Sequence Model via Joint Distillation and Quantization
Speaker: Zheng Li

Short: DQ-BART: Efficient Sequence-to-Sequence Model via Joint Distillation and Quantization
Speaker: Zheng Li

Short: Kronecker Decomposition for GPT Compression
Speaker: Ali Edalati

Findings: When Chosen Wisely, More Data Is What You Need: A Universal Sample-Efficient Strategy For Data Augmentation
Speaker: Ehsan Kamalloo

Findings: Ranking-Constrained Learning with Rationales for Text Classification
Speaker: Juanyan Wang

Short: SCD: Self-Contrastive Decorrelation of Sentence Embeddings
Speaker: Tassilo Klein

Long: Prompt-free and Efficient Few-shot Learning with Language Models
Speaker: Rabeeh Karimi Mahabadi

Long: Token Dropping for Efficient BERT Pretraining
Speaker: Le Hou

Findings: Revisiting Uncertainty-based Query Strategies for Active Learning with Transformers
Speaker: Christopher Schröder

Long: Bag-of-Words vs. Graph vs. Sequence in Text Classification: Questioning the Necessity of Text-Graphs and the Surprising Strength of a Wide MLP
Speaker: Lukas Galke

Long: Noisy Channel Language Model Prompting for Few-Shot Text Classification
Speaker: Sewon Min

Findings: Improving Robustness of Language Models from a Geometry-aware Perspective
Speaker: Bin Zhu

Findings: THE-X: Privacy-Preserving Transformer Inference with Homomorphic Encryption
Speaker: Tianyu Chen

Long: CAMERO: Consistency Regularized Ensemble of Perturbed Language Models with Weight Sharing
Speaker: Chen Liang

Long: Sharpness-Aware Minimization Improves Language Model Generalization
Speaker: Dara Bahri

Long: Adapting Coreference Resolution Models through Active Learning
Speaker: Michelle Yuan

Long: Softmax Bottleneck Makes Language Models Unable to Represent Multi-mode Word Distributions
Speaker: Haw-Shiuan Chang

Long: Coherence boosting: When your pretrained language model is not paying enough attention
Speaker: Nikolay Malkin

Long: Label Semantic Aware Pre-training for Few-shot Text Classification
Speaker: Aaron Mueller

Long: A Comparison of Strategies for Source-Free Domain Adaptation
Speaker: Xin Su

Long: PPT: Pre-trained Prompt Tuning for Few-shot Learning
Speaker: Yuxian Gu

Short: On the Importance of Effectively Adapting Pretrained Language Models for Active Learning
Speaker: Katerina Margatina

TACL: Compressing Large-Scale Transformer-Based Models: A Case Study on BERT
Speaker: Prakhar Ganesh

VPS4: Machine Translation and Multilinguality

19:00-20:00 (GatherTown)

Long: CipherDAug: Ciphertext based Data Augmentation for Neural Machine Translation
Speaker: Nishant Kambhatla

Findings: Meta-X$_{NLG}$: A Meta-Learning Approach Based on Language Clustering for Zero-Shot Cross-Lingual Transfer and Generation
Speaker: Kaushal Kumar Maurya

Findings: MR-P: A Parallel Decoding Algorithm for Iterative Refinement Non-Autoregressive Translation
Speaker: Hao Cheng

Findings: What Works and Doesn't Work, A Deep Decoder for Neural Machine Translation
Speaker: Zuchao Li

Findings: Fast Nearest Neighbor Machine Translation
Speaker: Yuxian Meng

Long: DEEP: DEnoising Entity Pre-training for Neural Machine Translation
Speaker: Junjie Hu

Long: Under the Morphosyntactic Lens: A Multifaceted Evaluation of Gender Bias in Speech Translation
Speaker: Beatrice Savoldi

Long: BiTIIMT: A Bilingual Text-infilling Method for Interactive Machine Translation
Speaker: Yanling Xiao

Long: Multilingual Mix: Example Interpolation Improves Multilingual Neural Machine Translation
Speaker: Yong Cheng

Long: Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder Translation Models
Speaker: Lorenzo Lupo

Findings: Detecting Various Types of Noise for Neural Machine Translation
Speaker: Christian Herold

Long: Can Synthetic Translations Improve Bitext Quality?
Speaker: Eleftheria Briakou

Long: Alternative Input Signals Ease Transfer in Multilingual Machine Translation
Speaker: Simeng Sun

Findings: Leveraging Knowledge in Multilingual Commonsense Reasoning
Speaker: Yuwei Fang

Short: Focus on the Target's Vocabulary: Masked Label Smoothing for Machine Translation
Speaker: Liang Chen

Findings: Rethinking Document-level Neural Machine Translation
Speaker: Zewei Sun

Findings: First the Worst: Finding Better Gender Translations During Beam Search
Speaker: Danielle Saunders

VPS4: NLP Applications

19:00-20:00 (GatherTown)

Long: Modeling U.S. State-Level Policies by Extracting Winners and Losers from Legislative Texts
Speaker: Maryam Davoodi

Long: TableFormer: Robust Transformer Modeling for Table-Text Encoding
Speaker: Jingfeng Yang

Long: It is AI's Turn to Ask Humans a Question: Question-Answer Pair Generation for Children's Story Books
Speaker: Dakuo Wang

Short: Automatic Detection of Entity-Manipulated Text using Factual Knowledge
Speaker: Ganesh Jawahar, Ganesh Jawahar

Findings: Constructing Open Cloze Tests Using Generation and Discrimination Capabilities of Transformers
Speaker: Mariano Felice

Long: Training Data is More Valuable than You Think: A Simple and Effective Method by Retrieving from Training Data
Speaker: Shuohang Wang

Findings: Question Generation for Reading Comprehension Assessment by Modeling How and What to Ask
Speaker: Bilal Ghanem

Long: A Neural Network Architecture for Program Understanding Inspired by Human Behaviors
Speaker: Renyu Zhu

Long: Improving Generalizability in Implicitly Abusive Language Detection with Concept Activation Vectors
Speaker: Isar Nejadgholi

Findings: How Can Cross-lingual Knowledge Contribute Better to Fine-Grained Entity Typing?
Speaker: Hailong Jin

Findings: Incremental Intent Detection for Medical Domain with Contrast Replay Networks
Speaker: Guirong Bai, Guirong Bai

Short: Adjusting the Precision-Recall Trade-Off with Align-and-Predict Decoding for Grammatical Error Correction
Speaker: Xin Sun

Long: Generating Scientific Definitions with Controllable Complexity
Speaker: Tal August

SRW: Improving Cross-domain, Cross-lingual and Multi-modal Deception Detection
Speaker: Subhadarshi Panda

SRW: Extraction of Diagnostic Reasoning Relations for Clinical Knowledge Graphs
Speaker: Vimig Socrates

VPS4: Phonology, Morphology and Word Segmentation

19:00-20:00 (GatherTown)

Short: Detecting Annotation Errors in Morphological Data with the Transformer
Speaker: Ling Liu, Ling Liu

Findings: Word Segmentation by Separation Inference for East Asian Languages
Speaker: Yu Tong

Findings: Unsupervised Chinese Word Segmentation with BERT Oriented Probing and Transformation
Speaker: Wei Li

VPS4: Question Answering

19:00-20:00 (GatherTown)

Findings: Plug-and-Play Adaptation for Continuously-updated QA
Speaker: Kyungjae Lee

Long: Answer-level Calibration for Free-form Multiple Choice Question Answering
Speaker: Sawan Kumar

Long: Modeling Multi-hop Question Answering as Single Sequence Prediction
Speaker: Semih Yavuz

Findings: Question Answering Infused Pre-training of General-Purpose Contextualized Representations
Speaker: Robin Jia

Long: Synthetic Question Value Estimation for Domain Adaptation of Question Answering
Speaker: Xiang Yue

Findings: Using Interactive Feedback to Improve the Accuracy and Explainability of Question Answering Systems Post-Deployment
Speaker: Zichao Li

Findings: Answer Uncertainty and Unanswerability in Multiple-Choice Machine Reading Comprehension
Speaker: Vatsal Raina

Short: Leveraging Explicit Lexico-logical Alignments in Text-to-SQL Parsing
Speaker: Runxin Sun

Findings: Two-Step Question Retrieval for Open-Domain QA
Speaker: yeon seonwoo

Findings: Hey AI, Can You Solve Complex Tasks by Talking to Agents?
Speaker: Tushar Khot

Short: C-MORE: Pretraining to Answer Open-Domain Questions by Consulting Millions of References
Speaker: Xiang Yue

Findings: Relevant CommonSense Subgraphs for "What if..." Procedural Reasoning
Speaker: Chen Zheng

Long: Deep Inductive Logic Reasoning for Multi-Hop Reading Comprehension
Speaker: Wenya Wang

Long: Simulating Bandit Learning from User Feedback for Extractive Question Answering
Speaker: Ge Gao

Long: MultiHiertt: Numerical Reasoning over Multi Hierarchical Tabular and Textual Data
Speaker: Yilun Zhao

Long: Improving Time Sensitivity for Question Answering over Temporal Knowledge Graphs
Speaker: Chao Shang

Outstanding Paper: Ditch the Gold Standard: Re-evaluating Conversational Question Answering
Speaker: Huihan Li

Long: Lite Unified Modeling for Discriminative Reading Comprehension
Speaker: Yilin Zhao

Long: Improving Machine Reading Comprehension with Contextualized Commonsense Knowledge
Speaker: Kai Sun

VPS4: Resources and Evaluation

19:00-20:00 (GatherTown)

Findings: Analyzing Dynamic Adversarial Training Data in the Limit
Speaker: Eric Wallace

Findings: RuCCoN: Clinical Concept Normalization in Russian
Speaker: Aleksandr Nesterov

Long: Fantastic Questions and Where to Find Them: FairytaleQA – An Authentic Dataset for Narrative Comprehension
Speaker: Ying Xu

Long: AraT5: Text-to-Text Transformers for Arabic Language Generation
Speaker: El Moatez Billah Nagoudi, El Moatez Billah Nagoudi

Long: Premise-based Multimodal Reasoning: Conditional Inference on Joint Textual and Visual Clues
Speaker: Qingxiu Dong

Long: Towards Robustness of Text-to-SQL Models Against Natural and Realistic Adversarial Table Perturbation
Speaker: Xinyu Pi

Long: Down and Across: Introducing Crossword-Solving as a New NLP Benchmark
Speaker: Saurabh Kulshreshtha

Short: To Find Waldo You Need Contextual Cues: Debiasing Who's Waldo
Speaker: Yiran Luo

Long: Understanding Iterative Revision from Human-Written Text
Speaker: Wanyu Du

Findings: ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning
Speaker: Ahmed Masry

Long: ePiC: Employing Proverbs in Context as a Benchmark for Abstract Language Understanding
Speaker: Sayan Ghosh

Long: Chart-to-Text: A Large-Scale Benchmark for Chart Summarization
Speaker: Shankar Kantharaj

Findings: XFUND: A Benchmark Dataset for Multilingual Visually Rich Form Understanding
Speaker: Yiheng Xu

Long: CLUES: A Benchmark for Learning Classifiers using Natural Language Explanations
Speaker: Rakesh Radhakrishnan Menon

Long: Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text
Speaker: Yao Dou

Short: CoDA21: Evaluating Language Understanding Capabilities of NLP Models With Context-Definition Alignment
Speaker: Lütfi Kerem Şenel

VPS4: Semantics

19:00-20:00 (GatherTown)

Findings: A Sentence is Worth 128 Pseudo Tokens: A Semantic-Aware Contrastive Learning Framework for Sentence Embeddings
Speaker: Haochen Tan

Long: FaiRR: Faithful and Robust Deductive Reasoning over Natural Language
Speaker: Soumya Sanyal

Long: LexSubCon: Integrating Knowledge from Lexical Resources into Contextual Embeddings for Lexical Substitution
Speaker: Georgios Michalopoulos

Long: WatClaimCheck: A new Dataset for Claim Entailment and Inference
Speaker: Kashif Khan

Findings: CoCoLM: Complex Commonsense Enhanced Language Model with Discourse Relations
Speaker: Changlong Yu

Long: Variational Graph Autoencoding as Cheap Supervision for AMR Coreference Resolution
Speaker: Irene Li, Irene Li

Findings: Dict-BERT: Enhancing Language Model Pre-training with Dictionary
Speaker: Wenhao Yu

Long: Zero-Shot Cross-lingual Semantic Parsing
Speaker: Tom Sherborne

Long: Nibbling at the Hard Core of Word Sense Disambiguation
Speaker: Marco Maru

Long: IMPLI: Investigating NLI Models' Performance on Figurative Language
Speaker: Kevin Stowe

Long: Bridging the Generalization Gap in Text-to-SQL Parsing with Schema Expansion
Speaker: Chen Zhao

Findings: HIE-SQL: History Information Enhanced Network for Context-Dependent Text-to-SQL Semantic Parsing
Speaker: Yanzhao Zheng

Findings: AMR-DA: Data Augmentation by Abstract Meaning Representation
Speaker: Ziyi Shou

Long: Few-Shot Learning with Siamese Networks and Label Tuning
Speaker: Thomas Müller

TACL: Is My Model Using The Right Evidence? Systematic Probes for Examining Evidence-Based Tabular Reasoning
Speaker: Vivek Gupta

SRW: AMR Alignment for Morphologically-rich and Pro-drop Languages
Speaker: K. Elif Oral

SRW: A large-scale computational study of content preservation measures for text style transfer and paraphrase generation
Speaker: Nikolay Babakov

SRW: Mining Logical Event Schemas From Pre-Trained Language Models
Speaker: Lane Lawley

VPS4: Sentiment Analysis, Stylistic Analysis, and Argument Mining

19:00-20:00 (GatherTown)

Findings: Efficient Argument Structure Extraction with Transfer Learning and Active Learning
Speaker: Xinyu Hua

Short: Pixie: Preference in Implicit and Explicit Comparisons
Speaker: Amanul Haque

Findings: Transfer Learning and Prediction Consistency for Detecting Offensive Spans of Text
Speaker: Amir Pouran Ben Veyseh

Long: Enhanced Multi-Channel Graph Convolutional Network for Aspect Sentiment Triplet Extraction
Speaker: Hao Chen

Findings: Seq2Path: Generating Sentiment Tuples as Paths of a Tree
Speaker: Yue Mao

Short: Direct parsing to sentiment graphs
Speaker: David Samuel

VPS4: Special Theme on Language Diversity: From Low Resource to Endangered

19:00-20:00 (GatherTown)

Findings: Pre-Trained Multilingual Sequence-to-Sequence Models: A Hope for Low-Resource Language Translation?
Speaker: En-Shiun Lee

Long: Expanding Pretrained Models to Thousands More Languages via Lexicon-based Adaptation
Speaker: Xinyi Wang

Long: Not always about you: Prioritizing community needs when developing endangered language technology
Speaker: Zoey Liu

Findings: Phoneme transcription of endangered languages: an evaluation of recent ASR architectures in the single speaker scenario
Speaker: Gilles Boulianne

Findings: Automatic Speech Recognition and Query By Example for Creole Languages Documentation
Speaker: Cécile Macaire

Short: Can a Transformer Pass the Wug Test? Tuning Copying Bias in Neural Morphological Inflection Models
Speaker: Ling Liu, Ling Liu

VPS4: Summarization

19:00-20:00 (GatherTown)

Long: Discriminative Marginalized Probabilistic Neural Method for Multi-Document Summarization of Medical Literature
Speaker: Gianluca Moro

Long: Modeling Hierarchical Syntax Structure with Triplet Position for Source Code Summarization
Speaker: Juncai Guo

Findings: End-to-End Segmentation-based News Summarization
Speaker: Yang Liu

Findings: Read Top News First: A Document Reordering Approach for Multi-Document News Summarization
Speaker: Chao Zhao

Long: Efficient Unsupervised Sentence Compression by Fine-tuning Transformers with Reinforcement Learning
Speaker: Demian Gholipour Ghalandari

Long: Efficient Unsupervised Sentence Compression by Fine-tuning Transformers with Reinforcement Learning
Speaker: Demian Gholipour Ghalandari

Long: DYLE: Dynamic Latent Extraction for Abstractive Long-Input Summarization
Speaker: Zimng Mao

Findings: HiStruct+: Improving Extractive Text Summarization with Hierarchical Structure Information
Speaker: Qian Ruan

Long: BRIO: Bringing Order to Abstractive Summarization
Speaker: Yixin Liu

Long: Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization
Speaker: Meng Cao

Long: MemSum: Extractive Summarization of Long Documents Using Multi-Step Episodic Markov Decision Processes
Speaker: Nianlong Gu

Findings: Focus on the Action: Learning to Highlight and Summarize Jointly for Email To-Do Items Summarization
Speaker: Kexun Zhang

Long: SummScreen: A Dataset for Abstractive Screenplay Summarization
Speaker: Mingda Chen

VPS4: Syntax: Tagging, Chunking and Parsing

19:00-20:00 (GatherTown)

Findings: Co-training an Unsupervised Constituency Parser with Weak Supervision
Speaker: Nickil Maveli

Best Paper: Learned Incremental Representations for Parsing
Speaker: Nikita Kitaev

Findings: Towards Few-shot Entity Recognition in Document Images: A Label-aware Sequence-to-Sequence Framework
Speaker: Zilong Wang

8 Workshops

Overview

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

Thursday, May 26, 2022

Friday, May 27, 2022

Thursday, May 26, 2022 - Friday, May 27, 2022

W1 - BioNLP 2022

Organizers:
Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, Jun-ichi Tsujii

https://aclweb.org/aclwiki/BioNLP_Workshop
Venue: Ecocem Room
Thursday, May 26, 2022

The BioNLP workshop associated with the ACL SIGBIOMED special interest group has established itself as the primary venue for presenting foundational research in language processing for the biological and medical domains. Despite, or maybe due to reaching maturity, the field of Biomedical NLP continues getting stronger. BioNLP welcomes and encourages inclusion and diversity. BioNLP truly encompasses the breadth of the domain and brings together researchers in bio- and clinical NLP from all over the world. The workshop will continue presenting work on a broad and interesting range of topics in NLP.

09:00 - 09:10	***Opening Remarks***
09:10 - 10:30	***Session 1: Question Answering, Discourse Structure and Clinical Applications***
09:10-09:30	*Explainable Assessment of Healthcare Articles with QA* Alodie Boissonnet, Marzieh Saeidi, Vassilis Plachouras and Andreas Vlachos
09:30-09:50	*A sequence-to-sequence approach for document-level relation extraction* John Giorgi, Gary Bader and Bo Wang
09:50-10:10	*Position-based Prompting for Health Outcome Generation* Micheal Abaho, Danushka Bollegala, Paula Williamson and Susanna Dodd
10:10-10:30	*How You Say It Matters: Measuring the Impact of Verbal Disfluency Tags on Automated Dementia Detection* Shahla Farzana, Ashwin Deshpande and Natalie Parde
10:30 - 11:00	***Coffee Break***
11:00 - 12:30	***Poster Session 1***
	Zero-Shot Aspect-Based Scientific Document Summarization using Self-Supervised Pre-training Amir Soleimani, Vassilina Nikoulina, Benoit Favre and Salah Ait Mokhtar
	Data Augmentation for Biomedical Factoid Question Answering Dimitris Pappas, Prodromos Malakasiotis and Ion Androutsopoulos
	Slot Filling for Biomedical Information Extraction Yannis Papanikolaou, Marlene Staib, Justin Joshua Grace and Francine Bennett
	Automatic Biomedical Term Clustering by Learning Fine-grained Term Representations Sihang Zeng, Zheng Yuan and Sheng Yu
	BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie and Sheng Yu
	Incorporating Medical Knowledge to Transformer-based Language Models for Medical Dialogue Generation Usman Naseem, Ajay Bandi, Shaina Raza, Junaid Rashid and Bharathi Raja Chakravarthi
	Memory-aligned Knowledge Graph for Clinically Accurate Radiology Image Report Generation Sixing Yan
	Simple Semantic-based Data Augmentation for Named Entity Recognition in Biomedical Texts Uyen Phan and Nhung Nguyen

Auxiliary Learning for Named Entity Recognition with Multiple Auxiliary Biomedical Training Data
Taiki Watanabe, Tomoya Ichikawa, Akihiro Tamura, Tomoya Iwakura, Chunpeng Ma and Tsuneo Kato

SNP2Vec: Scalable Self-Supervised Pre-Training for Genome-Wide Association Study
Samuel Cahyawijaya, Tiezheng Yu, Zihan Liu, Xiaopu Zhou, Tze Mak, Yuk Ip and Pascale Fung

Biomedical NER using Novel Schema and Distant Supervision
Anshita Khandelwal, Alok Kar, Veera Raghavendra Chikka and Kamalakar Karlapalem

Improving Supervised Drug-Protein Relation Extraction with Distantly Supervised Models
Naoki Iinuma, Makoto Miwa and Yutaka Sasaki

Named Entity Recognition for Cancer Immunology Research Using Distant Supervision
Hai-Long Trieu, Makoto Miwa and Sophia Ananiadou

Intra-Template Entity Compatibility based Slot-Filling for Clinical Trial Information Extraction
Christian Witte and Philipp Cimiano

Pretrained Biomedical Language Models for Clinical NLP in Spanish
Casimiro Pio Carrino, Joan Llop, Marc Pmies, Asier Gutirrez-Fandio, Jordi Armengol-Estap, Joaqun Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre and Marta Villegas

Few-Shot Cross-lingual Transfer for Coarse-grained De-identification of Code-Mixed Clinical Texts
Saadullah Amin, Noon Pokaratsiri Goldstein, Morgan Wixted, Alejandro Garcia-Rudolph, Catalina Martnez-Costa and Guenter Neumann

VPAI Lab at MedVidQA 2022: A Two Stage Cross-modal Fusion Method for Medical Instructional Video Classification
Bin Li, Yixuan Weng, Fei Xia, Bin Sun and Shutao Li

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

14:00 - 15:00 ***Session 2: Summarization and text mining***

14:00-14:20 *GenCompareSum: a hybrid unsupervised summarization method using salience*
Jennifer Bishop, Qianqian Xie and Sophia Ananiadou

14:20-14:40 *BioCite: A Deep Learning-based Citation Linkage Framework for Biomedical Research Articles*
Sudipta Roy and Robert Mercer

14:40-15:00 *Low Resource Causal Event Detection from Biomedical Literature*
Zhengzhong Liang, Enrique Noriega-Atala, Clayton Morrison and Mihai Surdeanu

15:00 - 15:30 ***Coffee Break***

15:30 - 17:00 ***Poster Session 2***

Overview of the MedVidQA 2022 Shared Task on Medical Video Question-Answering
Deepak Gupta and Dina Demner-Fushman

Inter-annotator agreement is not the ceiling of machine learning performance: Evidence from a comprehensive set of simulations
Russell Richie, Sachin Grover and Fuchiang Tsui

Conversational Bots for Psychotherapy: A Study of Generative Transformer Models Using Domain-specific Dialogues
Avisha Das, Salih Selek, Alia Warner, Xu Zuo, Yan Hu, Vipina Kuttichi Keloth, Jianfu Li, W. Jim Zheng and Hua Xu

Improving Romanian BioNER Using a Biologically Inspired System
Maria Mitrofan and Vasile Pais

BanglaBioMed: A Biomedical Named-Entity Annotated Corpus for Bangla (Bengali)
Salim Sazzed

BEEDS: Large-Scale Biomedical Event Extraction using Distant Supervision and Question Answering
Xing Wang, Ulf Leser and Leon Weber

Data Augmentation for Rare Symptoms in Vaccine Side-Effect Detection
Bosung Kim and Ndapa Nakashole

ICDBigBird: A Contextual Embedding Model for ICD Code Classification
George Michalopoulos, Michal Malyska, Nicola Sahar, Alexander Wong and Helen Chen

Doctor XAvIer: Explainable Diagnosis on Physician-Patient Dialogues and XAI Evaluation
Hillary Ngai and Frank Rudzicz

DISTANT-CTO: A Zero Cost, Distantly Supervised Approach to Improve Low-Resource Entity Extraction Using Clinical Trials Literature
Anjani Dhrangadhariya and Henning Mller

EchoGen: Generating Conclusions from Echocardiogram Notes
Liyan Tang, Shravan Kooragayalu, Yanshan Wang, Ying Ding, Greg Durrett, Justin F. Rousseau and Yifan Peng

Quantifying Clinical Outcome Measures in Patients with Epilepsy Using the Electronic Health Record
Kevin Xie, Brian Litt, Dan Roth and Colin Ellis

Comparing Encoder-Only and Encoder-Decoder Transformers for Relation Extraction from Biomedical Texts: An Empirical Study on Ten Benchmark Datasets
Mourad Sarrouti, Carson Tao and Yoann Randriamihaja

Utility Preservation of Clinical Text After De-Identification
Thomas Vakili and Hercules Dalianis

Horses to Zebras: Ontology-Guided Data Augmentation and Synthesis for ICD-9 Coding
Mat Falis, Hang Dong, Alexandra Birch and Beatrice Alex

Towards Automatic Curation of Antibiotic Resistance Genes via Statement Extraction from Scientific Papers: A Benchmark Dataset and Models
Sidhant Chandak, Liqing Zhang, Connor Brown and Lifu Huang

Model Distillation for Faithful Explanations of Medical Code Predictions
Zach Wood-Doughty, Isabel Cachola and Mark Dredze

Towards Generalizable Methods for Automating Risk Score Calculation
Jennifer Liang, Eric Lehman, Ananya Iyengar, Diwakar Mahajan, Preethi Raghavan, Cindy Chang and Peter Szolovits

DoSSIER at MedVidQA 2022: Text-based Approaches to Medical Video Answer Localization Problem
Wojciech Kusa, Georgios Peikos, Oscar Espitia, Allan Hanbury and Gabriella Pasi

17:00 - 17:10 ***Closing Remarks***

W2 - NLP Power! The First Workshop on Efficient Benchmarking in NLP

Organizers:
Tatiana Shavrina, Valentin Malykh, Ekaterina Artemova, Vladislav Mikhailov, Oleg Serikov, Vitaly Protasov

https://nlp-power.github.io/
Venue: Wicklow Hall 2a
Thursday, May 26, 2022

The main objectives of this workshop are to (1) create a space for critical reflection on current benchmarks and evaluation tools, (2) encourage the development of improved or new benchmarks and evaluation tools that resolve current challenges, (3) develop better approaches to model ranking, (4) rethink benchmarking strategies that best account for computational costs, energy and ethical considerations, out-of-domain language capabilities and meeting the end-user preferences. We welcome submissions on ongoing and finished research and hope to provide an opportunity for participants to present their work and exchange ideas.

09:00 - 09:10	***Opening Remarks***
09:10 - 10:30	***Benchmarking and evaluation*** - Chair: Tatiana Shavrina
09:10-09:30	*Checking HateCheck: a cross-functional analysis of behaviour-aware learning for hate speech detection* Pedro Henrique Luz de Araujo and Benjamin Roth
09:30-09:50	*Raison d'être of the benchmark dataset: A Survey of Current Practices of Benchmark Dataset Sharing Platforms* Jaihyun Park and Sullam Jeoung
09:50-10:10	*Beyond Static models and test sets: Benchmarking the potential of pre-trained models across tasks and languages* Kabir Ahuja, Sandipan Dandapat, Sunayana Sitaram and Monojit Choudhury
10:10-10:30	*Characterizing the Efficiency vs. Accuracy Trade-off for Long-Context NLP Models* Phyllis Ang, Bhuwan Dhingra and Lisa Wu Wills
10:30 - 11:00	***Coffee Break***
11:00 - 11:40	***Poster Session (Gathertown)***
11:40 - 12:00	***Oral Session*** - Chair: Tatiana Shavrina
11:40-12:00	*Why only Micro-F1? Class Weighting of Measures for Relation Classification* David Harbecke, Yuxuan Chen, Leonhard Hennig and Christoph Alt
12:00 - 12:20	***ACL Findings*** - Chair: Tatiana Shavrina
12:30 - 14:00	***Lunch Break***
14:00 - 15:00	***Invited Talk 1***
15:00 - 15:30	***Coffee Break***
15:30 - 16:10	***Invited Talk 2***
16:10 - 16:50	***Invited Talk 3***
16:50 - 17:50	***Round Table - Beyond GLUE***
17:50 - 18:00	***Closing remarks***

W3 - The Fifth Workshop on e-Commerce and NLP (ECNLP 5)

Organizers:
Shervin Malmasi, Eugene Agichtein, Surya Kallumadi, Nicola Ueffing, Oleg Rokhlenko, Ido Guy

https://sites.google.com/view/ecnlp
Venue: Wicklow Meeting Room 5
Thursday, May 26, 2022

NLP and IR have been closely linked to e-commerce applications since the early days of the fields. This close relationship between the two is evidenced by early publications as well as the growing number of recent publications at the intersection of the two areas. Today, NLP and IR play a significant role in e-commerce tasks, including product search, recommender systems, product question answering, sentiment analysis, product description and review summarization, chatbots and shopping assistants, and customer review processing, among many other tasks being investigated by researchers in the field. These methods play a key part in today's online retail and shopping landscape, and continue to evolve and further enhance the customer experience. The ECNLP workshop aims to provide a venue for the dissemination of NLP/IR research related to e-commerce and online shopping, bringing together researchers from both academia and industry.

09:00 - 18:00 ***Workshop Schedule will be announced on the website***

W4 - The Fifth Workshop on Fact Extraction and VERification (FEVER)

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

https://fever.ai/
Venue: Wicklow Hall 2b
Thursday, May 26, 2022

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

09:00 - 09:45	***Keynote Talk: Alon Halevy. Human Values in Recommender Systems: a Multi-Disciplinary Discussion***
09:45 - 10:30	***Keynote Talk: Alice Oh. Data collection, bias mitigation and hate speech detection in multiple languages***
10:30 - 11:00	***Coffee break***
11:00 - 11:45	***Keynote Talk: Carolina Scarton. Supporting professional fact-checking: how can NLP/AI help?***
11:45 - 12:30	***Contributed Talks***
11:45-12:00	*Neural Machine Translation for Fact-checking Temporal Claims* Marco Mori, Paolo Papotti, Luigi Bellomarini and Oliver Giudice
12:00-12:15	*Automatic Fake News Detection: Are current models "fact-checking" or"gut-checking"?* Ian Kelk, Benjamin Basseri, Wee Yi Lee, Richard Qiu and Chris Tanner
12:15-12:30	*Retrieval Data Augmentation Informed by Downstream Question Answering Performance* James Ferguson, Hannaneh Hajishirzi, Pradeep Dasigi and Tushar Khot
12:30 - 14:00	***Lunch Break***
14:00 - 14:30	***In-person poster session***
	XInfoTabS: Evaluating Multilingual Tabular Natural Language Inference Bhavnick Singh Minhas, Anant Shankhdhar, Vivek Gupta, Divyanshu Aggarwal and Shuo Zhang
	PHEMEPlus: Enriching Social Media Rumour Verification with External Evidence John Dougrez-Lewis, Elena Kochkina, Miguel Arana-Catania, Maria Liakata and Yulan He
	A Semantics-Aware Approach to Automated Claim Verification Blanca Calvo Figueras, Montse Cuadros Oller and Rodrigo Agerri
14:30 - 15:00	***Online poster session***
	Heterogeneous-Graph Reasoning and Fine-Grained Aggregation for Fact Checking Hongbin Lin and Xianghua Fu

	Distilling Salient Reviews with Zero Labels Chieh-Yang Huang, Jinfeng Li, Nikita Bhutani, Alexander Whedon, Estevam Hruschka and Yoshi Suhara
15:00 - 15:30	***Coffee break***
15:30 - 16:15	***Keynote Talk: Kiran Garimella. Content moderation on encrypted platforms***
16:15 - 17:00	***Keynote Talk: Tanu Mitra. Problematic Information on Social Media Platforms: Understanding and Countering***

W5 - The Second Workshop on Speech and Language Technologies for Dravidian Languages - (Dravidian LangTech-2022)

Organizers:
Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar M, Parameswari Krishnamurthy, Elizabeth Sherly, Sinnathamby Mahesan

https://dravidianlangtech.github.io/2022/
Venue: Liffey Meeting Room 4
Thursday, May 26, 2022

The development of technology increases our internet use, and most of the global languages have adapted themselves to the digital era. However, there are many regional, under-resourced languages that face challenges as they still lack developments in language technology. One such language family is the Dravidian family of languages. Dravidian languages are primarily spoken in south India and Sri Lanka. Pockets of speakers are found in Nepal, Pakistan, Malaysia, other parts of India and elsewhere in the world. The Dravidian languages, which are 4,500 years old and spoken by millions of speakers, are under-resourced in speech and natural language processing. The Dravidian languages are divided into four groups: South, South-Central, Central, and North groups. Dravidian morphology is agglutinating and exclusively suffixal. Syntactically, Dravidian languages are head-final and left-branching. They are free-constituent order languages. To improve access to and production of information for monolingual speakers of Dravidian languages, it is necessary to have speech and languages technologies. The aim of these workshops is to save the Dravidian languages from extinction in technology. This is the first workshop on speech and language technologies for Dravidian languages.

09:15 - 09:30	***Opening Remarks***
09:30 - 10:00	***Keynote***
10:00 - 11:00	***Multitask and Multimodal Learning in Dravidian Languages*** - Chair: Anand Kumar Madasamy
10:00-10:15	*Findings of the Shared Task on Multi-task Learning in Dravidian Languages* Bharathi Raja Chakravarthi, Ruba Priyadharshini, Subalalitha CN, Sangeetha S, Malliga Subramanian, Kogilavani Shanmugavadivel, Parameswari Krishnamurthy, Adeep Hande, Siddhanth U Hegde, Roshan Nayak and Swetha Valli
10:15-10:30	*MuCoT: Multilingual Contrastive Training for Question-Answering in Low-resource Languages* Gokul Karthik Kumar, Abhishek Singh Gehlot, Sahal Shaji Mullappilly and Karthik Nandakumar
10:30-10:45	*Findings of the Shared Task on Multimodal Sentiment Analysis and Troll Meme Classification in Dravidian Languages* Premjith B, Bharathi Raja Chakravarthi, Malliga Subramanian, Bharathi B, Soman KP, Dhanalakshmi V, Sreelakshmi K, Arunaggiri Pandian and Prasanna Kumar Kumaresan
10:45-11:00	*A Dataset for Detecting Humor in Telugu Social Media Text* Sriphani Vardhan Bellamkonda, Maithili Lohakare and Shaswat P Patel
11:00 - 11:30	***Break***
11:00 - 13:00	***Identifying Emotions, Troll, Abuse and Offensive Contents in Dravidian Languages*** - Chair: Parameswari Krishnamurthy

11:00-11:15	*Findings of the Shared Task on Offensive Span Identification fromCode-Mixed Tamil-English Comments* Manikandan Ravikiran, Bharathi Raja Chakravarthi, Anand Kumar Madasamy, Sangeetha S, Ratnavel Rajalakshmi, Sajeetha Thavareesan, Rahul Ponnusamy and Shankar Mahadevan
11:15-11:30	*Overview of the Shared Task on Machine Translation in Dravidian Languages* Anand Kumar Madasamy, Asha Hegde, Shubhanker Banerjee, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Hosahalli Lakshmaiah Shashirekha and John Philip McCrae
11:30-11:45	*BERT-Based Sequence Labelling Approach for Dependency Parsing in Tamil* C S Ayush Kumar, Advaith Das Maharana, Srinath Murali, Premjith B and Soman KP
11:45-12:00	*Zero-shot Code-Mixed Offensive Span Identification through Rationale Extraction* Manikandan Ravikiran and Bharathi Raja Chakravarthi
12:00-12:15	*Overview of Abusive Comment Detection in Tamil-ACL 2022* Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha CN, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde and Prasanna Kumar Kumaresan
12:15-12:30	*TamilATIS: Dataset for Task-Oriented Dialog in Tamil* Ramaneswaran S, Sanchit Vijay and Kathiravan Srinivasan
12:30-12:45	*Findings of the Shared Task on Emotion Analysis in Tamil* Anbukkarasi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Subalalitha CN, Kogilavani Shanmugavadivel, Sajeetha Thavareesan, Sathiyaraj Thangasamy, Parameswari Krishnamurthy, Adeep Hande, Sean Benhur, Kishore Kumar Ponnusamy and Santhiya Pandiyan
12:45-13:00	*Sentiment Analysis on Code-Switched Dravidian Languages with Kernel Based Extreme Learning Machines* Mithun Kumar S R, Lov Kumar and Aruna Malapati
13:00 - 14:00	***Break***
14:00 - 17:00	***Poster Session: Shared Task Papers*** - Chair: Manikandan Ravikiran
14:00-17:00	*hate-alert@DravidianLangTech-ACL2022: Ensembling Multi-Modalities for Tamil TrollMeme Classification* Mithun Das, Somnath Banerjee and Animesh Mukherjee
14:00-17:00	*TeamX@DravidianLangTech-ACL2022: A Comparative Analysis for Troll-Based Meme Classification* Rabindra Nath Nandi, Firoj Alam and Preslav Nakov
14:00-17:00	*Translation Techies @DravidianLangTech-ACL2022-Machine Translation in Dravidian Languages* Piyushi Goyal, Musica Supriya, Dinesh Acharya U and Ashalatha Nayak
14:00-17:00	*SSN_MLRG1@DravidianLangTech-ACL2022: Troll Meme Classification in Tamil using Transformer Models* Shruthi Hariprasad, Sarika Esackimuthu, Saritha Madhavan, Rajalakshmi Sivanaiah and Angel Deborah S
14:00-17:00	*BPHC@DravidianLangTech-ACL2022-A comparative analysis of classical and pre-trained models for troll meme classification in Tamil* Achyuta Krishna V, Mithun Kumar S R, Aruna Malapati and Lov Kumar
14:00-17:00	*CUET-NLP@DravidianLangTech-ACL2022: Investigating Deep Learning Techniques to Detect Multimodal Troll Memes* Md Maruf Hasan, Nusratul Jannat, Eftekhar Hossain, Omar Sharif and Mohammed Moshiul Hoque
14:00-17:00	*PICT@DravidianLangTech-ACL2022: Neural Machine Translation On Dravidian Languages* Aditya Vyawahare, Rahul Tangsali, Aditya Mandke, Onkar Rupesh Litake and Dipali Kadam
14:00-17:00	*CUET-NLP@DravidianLangTech-ACL2022: Exploiting Textual Features to Classify Sentiment of Multimodal Movie Reviews*

Nasehatul Mustakim, Nusratul Jannat, Md Maruf Hasan, Eftekhar Hossain, Omar Sharif and Mohammed Moshiul Hoque

14:00-17:00 *MUCIC@TamilNLP-ACL2022: Abusive Comment Detection in Tamil Language using 1D Conv-LSTM*
Fazlourrahman Balouchzahi, Anusha M D Gowda, Hosahalli Lakshmaiah Shashirekha and Grigori Sidorov

14:00-17:00 *NITK-IT_NLP@TamilNLP-ACL2022: Transformer based model for Toxic Span Identification in Tamil*
Hariharan RamakrishnaIyer LekshmiAmmal, Manikandan Ravikiran and Anand Kumar Madasamy

14:00-17:00 *GJG@TamilNLP-ACL2022: Using Transformers for Abusive Comment Classification in Tamil*
Gaurang Prasad, Janvi Prasad and Gunavathi C

14:00-17:00 *IIITDWD@TamilNLP-ACL2022: Transformer-based approach to classify abusive content in Dravidian Code-mixed text*
Shankar Biradar and Sunil Saumya

14:00-17:00 *PANDAS@Abusive Comment Detection in Tamil Code-Mixed Data Using Custom Embeddings with LaBSE*
Krithika Swaminathan, Divyasri K, Gayathri G L, Thenmozhi Durairaj and Bharathi B

14:00-17:00 *BpHigh@TamilNLP-ACL2022: Effects of Data Augmentation on Indic-Transformer based classifier for Abusive Comments Detection in Tamil*
Bhavish Pahwa

14:00-17:00 *SSNCSE NLP@TamilNLP-ACL2022: Transformer based approach for detection of abusive comment for Tamil language*
Bharathi B and Josephine Varsha

14:00-17:00 *DLRG@DravidianLangTech-ACL2022: Abusive Comment Detection in Tamil using Multilingual Transformer Models*
Ratnavel Rajalakshmi, Ankita Duraphe and Antonette Shibani

14:00-17:00 *Aanisha@TamilNLP-ACL2022:Abusive Detection in Tamil*
Aanisha Bhattacharyya

14:00-17:00 *COMBATANT@TamilNLP-ACL2022: Fine-grained Categorization of Abusive Comments using Logistic Regression*
Alamgir Hossain, Mahathir Mohammad Bishal, Eftekhar Hossain, Omar Sharif and Mohammed Moshiul Hoque

14:00-17:00 *Optimize_Prime@DravidianLangTech-ACL2022: Abusive Comment Detection in Tamil*
Shantanu Patankar, Omkar Bhushan Gokhale, Onkar Rupesh Litake, Aditya Mandke and Dipali Kadam

14:00-17:00 *DLRG@TamilNLP-ACL2022: Offensive Span Identification in Tamil usingBiLSTM-CRF approach*
Ratnavel Rajalakshmi, Mohit Madhukar More, Bhamatipati Naga Shrikriti, Gitansh Saharan, Hanchate Samyuktha and Sayantan Nandy

14:00-17:00 *DE-ABUSE@TamilNLP-ACL 2022: Transliteration as Data Augmentation for Abuse Detection in Tamil*
Vasanth Palanikumar, Sean Benhur, Adeep Hande and Bharathi Raja Chakravarthi

14:00-17:00 *UMUTeam@TamilNLP-ACL2022: Emotional Analysis in Tamil*
José Antonio García-Díaz, Miguel Ángel Rodríguez García and Rafael Valencia-García

14:00-17:00 *UMUTeam@TamilNLP-ACL2022: Abusive Detection in Tamil using Linguistic Features and Transformers*
José Antonio García-Díaz, Manuel Valencia-Garcia and Rafael Valencia-García

14:00-17:00 *JudithJeyafreedaAndrew@TamilNLP-ACL2022:CNN for Emotion Analysis in Tamil*
Judith Jeyafreeda Andrew

14:00-17:00 *CEN-Tamil@DravidianLangTech-ACL2022: Abusive Comment detection in Tamil using TF-IDF and Random Kitchen Sink Algorithm*

	Prasanth S N, R Aswin Raj, Adhithan P, Premjith B and Soman KP
14:00-17:00	*GJG@TamilNLP-ACL2022: Emotion Analysis and Classification in Tamil using Transformers* Janvi Prasad, Gaurang Prasad and Gunavathi C
14:00-17:00	*PANDAS@TamilNLP-ACL2022: Emotion Analysis in Tamil Text using Language Agnostic Embeddings* Divyasri K, Gayathri G L, Krithika Swaminathan, Thenmozhi Durairaj, Bharathi B and Senthil Kumar B
14:00-17:00	*SSNCSE_NLP@TamilNLP-ACL2022: Transformer based approach for Emotion analysis in Tamil language* Bharathi B and Josephine Varsha
14:00-17:00	*MUCS@DravidianLangTech@ACL2022: Ensemble of Logistic Regression Penalties to Identify Emotions in Tamil Text* Asha Hegde, Sharal Coelho and Hosahalli Lakshmaiah Shashirekha
14:00-17:00	*Varsini_and_Kirthanna@DravidianLangTech-ACL2022-Emotional Analysis in Tamil* Varsini S, Kirthanna Rajan, Angel Deborah S, Rajalakshmi Sivanaiah, Sakaya Milton Rajendram and Mirnalinee T T
14:00-17:00	*CUET-NLP@TamilNLP-ACL2022: Multi-Class Textual Emotion Detection from Social Media using Transformer* Nasehatul Mustakim, Rabeya Akter Rabu, Golam Sarwar Md. Mursalin, Eftekhar Hossain, Omar Sharif and Mohammed Moshiul Hoque
14:00-17:00	*Optimize_Prime@DravidianLangTech-ACL2022: Emotion Analysis in Tamil* Omkar Bhushan Gokhale, Shantanu Patankar, Onkar Rupesh Litake, Aditya Mandke and Dipali Kadam
17:00 - 17:15	***Meeting, Awards, Closing (TBD)***

W6 - The 19th International Conference on Spoken Language Translation (ACL-IWSLT 2022)

Organizers:
Marcello Federico, Alex Waibel, Marta R. Costa-jussà, Jan Niehues, Sebastian Stüker, Elizabeth Salesky

https://iwslt.org/2022/
Venue: Wicklow Meeting Room 3
Thursday, May 26, 2022 - Friday, May 27, 2022

The International Conference on Spoken Language Translation (IWSLT) is an annual scientific conference, associated with an open evaluation campaign on spoken language translation, where both scientific papers and system descriptions are presented.

09:00 - 09:10	***Welcome Remarks***
09:10 - 10:30	***Overview of the IWSLT 2022 Evaluation Campaign*** - Chair: Marcello Federico
09:00-10:30	*Findings of the IWSLT 2022 Evaluation Campaign* Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel M. Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang and Shinji Watanabe
10:30 - 11:00	***Coffee Break***
11:00 - 12:30	***Oral Session 1: Scientific Papers***
11:00-11:20	*SubER - A Metric for Automatic Evaluation of Subtitle Quality* Patrick Wilken, Panayota Georgakopoulou and Evgeny Matusov
11:20-11:40	*Improving Arabic Diacritization by Learning to Diacritize and Translate* Brian Thompson and Ali Alshehri
11:40-12:00	*Simultaneous Neural Machine Translation with Prefix Alignment* Yasumasa Kano, Katsuhito Sudoh and Satoshi Nakamura
12:00-12:20	*Locality-Sensitive Hashing for Long Context Neural Machine Translation* Frithjof Petrick, Jan Rosendahl, Christian Herold and Hermann Ney
12:30 - 14:00	***Lunch Break***
14:00 - 15:00	***Keynote***
15:00 - 15:30	***Coffee Break***
15:30 - 17:30	***Poster Session: System Papers***
15:30-17:30	*The YiTrans Speech Translation System for IWSLT 2022 Offline Shared Task* Ziqiang Zhang and Junyi Ao
15:30 - 17:30	***Poster Session: System Papers***
15:30-17:30	*Amazon Alexa AI's System for IWSLT 2022 Offline Speech Translation Shared Task* Akshaya Vishnu Kudlu Shanbhogue, Ran Xue, Ching-Yun Chang and Sarah Campbell
15:30-17:30	*Efficient yet Competitive Speech Translation: FBK@IWSLT2022*

	Marco Gaido, Sara Papi, Dennis Fucci, Giuseppe Fiameni, Matteo Negri and Marco Turchi
15:30-17:30	*Effective combination of pretrained models - KIT@IWSLT2022* Ngoc-Quan Pham, Tuan Nam Nguyen, Thai-Binh Nguyen, Danni Liu, Carlos Mullov, Jan Niehues and Alexander Waibel
15:30-17:30	*The USTC-NELSLIP Offline Speech Translation Systems for IWSLT 2022* Weitai Zhang, Zhongyi Ye, Haitao Tang, Xiaoxi Li, Xinyuan Zhou, Jing Yang, Jianwei Cui, Dan Liu, Junhua Liu and Lirong Dai
15:30-17:30	*The AISP-SJTU Simultaneous Translation System for IWSLT 2022* Qinpei Zhu, Renshou Wu, Guangfeng Liu, Xinyu Zhu, Xingyu Chen, Yang Zhou, Qingliang Miao, Rui Wang and Kai Yu
15:30-17:30	*The Xiaomi Text-to-Text Simultaneous Speech Translation System for IWSLT 2022* Bao Guo, Mengge Liu, Wen Zhang, Hexuan Chen, Chang Mu, Xiang Li, Jianwei Cui, Bin Wang and Yuhang Guo
15:30-17:30	*NVIDIA NeMo Offline Speech Translation Systems for IWSLT 2022* Oleksii Hrinchuk, Vahid Noroozi, Ashwinkumar Ganesan, Sarah Campbell, Sandeep Subramanian, Somshubra Majumdar and Oleksii Kuchaiev
15:30-17:30	*The NiuTrans's Submission to the IWSLT22 English-to-Chinese Offline Speech Translation Task* Yuhao Zhang, Canan Huang, Chen Xu, Xiaoqian Liu, Bei Li, Anxiang Ma, Tong Xiao and Jingbo Zhu
15:30-17:30	*The HW-TSC's Offline Speech Translation System for IWSLT 2022 Evaluation* Minghan Wang, Jiaxin Guo, Xiaosong Qiao, Yuxia Wang, Daimeng Wei, Chang Su, Yimeng Chen, Min Zhang, Shimin Tao, Hao Yang and Ying Qin
15:30-17:30	*The HW-TSC's Simultaneous Speech Translation System for IWSLT 2022 Evaluation* Minghan Wang, Jiaxin Guo, Yinglu Li, Xiaosong Qiao, Yuxia Wang, Zongyao Li, Chang Su, Yimeng Chen, Min Zhang, Shimin Tao, Hao Yang and Ying Qin
15:30 - 17:30	***Poster Session: System Papers***
15:30-17:30	*MLLP-VRAIN UPV systems for the IWSLT 2022 Simultaneous Speech Translation and Speech-to-Speech Translation tasks* Javier Iranzo-Sánchez, Javier Jorge Cano, Alejandro Pérez-González-de-Martos, Adrián Giménez Pastor, Gonçal V. Garcés Díaz-Munío, Pau Baquero-Arnal, Joan Albert Silvestre-Cerdà, Jorge Civera Saiz, Albert Sanchis and Alfons Juan
15:30-17:30	*Pretrained Speech Encoders and Efficient Fine-tuning Methods for Speech Translation: UPC at IWSLT 2022* Ioannis Tsiamas, Gerard I. Gállego, Carlos Escolano, José A. R. Fonollosa and Marta R. Costa-jussà
15:30-17:30	*CUNI-KIT System for Simultaneous Speech Translation Task at IWSLT 2022* Peter Polák, Ngoc-Quan Pham, Tuan Nam Nguyen, Danni Liu, Carlos Mullov, Jan Niehues, Ondřej Bojar and Alexander Waibel
15:30-17:30	*NAIST Simultaneous Speech-to-Text Translation System for IWSLT 2022* Ryo Fukuda, Yuka Ko, Yasumasa Kano, Kosuke Doi, Hirotaka Tokuyama, Sakriani Sakti, Katsuhito Sudoh and Satoshi Nakamura
15:30-17:30	*The HW-TSC's Speech to Speech Translation System for IWSLT 2022 Evaluation* Jiaxin Guo, Yinglu Li, Minghan Wang, Xiaosong Qiao, Yuxia Wang, Hengchao Shang, Chang Su, Yimeng Chen, Min Zhang, Shimin Tao, Hao Yang and Ying Qin
15:30-17:30	*CMU's IWSLT 2022 Dialect Speech Translation System* Brian Yan, Patrick Fernandes, Siddharth Dalmia, Jiatong Shi, Yifan Peng, Dan Berrebbi, Xinyi Wang, Graham Neubig and Shinji Watanabe
15:30-17:30	*ON-TRAC Consortium Systems for the IWSLT 2022 Dialect and Low-resource Speech Translation Tasks* Marcely Zanon Boito, John Ortega, Hugo Riguidel, Antoine Laurent, Loïc Barrault, Fethi Bougares, Firas Chaabani, Ha Nguyen, Florentin Barbier, Souhir Gahbiche and Yannick Estève
15:30-17:30	*JHU IWSLT 2022 Dialect Speech Translation System Description*

	Jinyi Yang, Amir Hussein, Matthew Wiesner and Sanjeev Khudanpur
15:30-17:30	*Controlling Translation Formality Using Pre-trained Multilingual Language Models* Elijah Rippeth, Sweta Agrawal and Marine Carpuat
15:30-17:30	*Controlling Formality in Low-Resource NMT with Domain Adaptation and Re-Ranking: SLT-CDT-UoS at IWSLT2022* Sebastian T. Vincent, Loïc Barrault and Carolina Scarton
15:30 - 17:30	***Poster Session: System Papers***
15:30-17:30	*Improving Machine Translation Formality Control with Weakly-Labelled Data Augmentation and Post Editing Strategies* Daniel Zhang, Jiang Yu, Pragati Verma, Ashwinkumar Ganesan and Sarah Campbell
15:30-17:30	*HW-TSC's Participation in the IWSLT 2022 Isometric Spoken Language Translation* Zongyao Li, Jiaxin Guo, Daimeng Wei, Hengchao Shang, Minghan Wang, Ting Zhu, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Lizhi Lei, Hao Yang and Ying Qin
15:30-17:30	*AppTek's Submission to the IWSLT 2022 Isometric Spoken Language Translation Task* Patrick Wilken and Evgeny Matusov
15:30-17:30	*Hierarchical Multi-task learning framework for Isometric-Speech Language Translation* Aakash Bhatnagar, Nidhir Bhavsar, Muskaan Singh and Petr Motlicek
09:00 - 10:30	***Oral Session 2: Scientific Papers*** - Chair: Marcello Federico
09:00-09:18	*Anticipation-Free Training for Simultaneous Machine Translation* Chih-Chiang Chang, Shun-Po Chuang and Hung-yi Lee
09:18-09:36	*Who Are We Talking About? Handling Person Names in Speech Translation* Marco Gaido, Matteo Negri and Marco Turchi
09:36-09:54	*Joint Generation of Captions and Subtitles with Dual Decoding* Jitao Xu, François Buet, Josep Crego, Elise Bertin-Lemée and François Yvon
09:54-10:12	*MirrorAlign: A Super Lightweight Unsupervised Word Alignment Model via Cross-Lingual Contrastive Learning* Di Wu, Liang Ding, Shuo Yang and Mingyang Li
10:12-10:30	*On the Impact of Noises in Crowd-Sourced Data for Speech Translation* Siqi Ouyang, Rong Ye and Lei Li
10:30 - 11:00	***Coffee Break***
11:00 - 12:30	***Discussion 1: Retrospective***
12:30 - 14:00	***Lunch Break***
14:00 - 15:00	***Discussion 2: Planning***
15:00 - 15:30	***Coffee Break***
15:30 - 16:30	***Discussion 3: Planning***
16:30 - 16:45	***Best Student Paper Award***
16:45 - 17:00	***Closing Remarks***

W7 - Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2022)

Organizers:
Emmanuele Chersoni, Nora Hollenstein, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, Enrico Santus

https://cmclorg.github.io/
Venue: Liffey Hall 2
Thursday, May 26, 2022

Cognitive Modeling and Computational Linguistics (CMCL) 2022 is a one-day workshop held in conjunction with the conference of the Association for Computational Linguistics (ACL). The goal of CMCL is providing a venue for computational research on cognitive theories of language processing, representation and acquisition. The 2022 workshop follows in the tradition of earlier meetings at ACL 2010, ACL 2011, NAACL-HLT 2012, ACL 2013, ACL 2014, NAACL 2015, EACL 2017, LSA 2018, NAACL 2019, EMNLP 2020 and NAACL 2021.

09:30 - 09:45	***Opening Remarks***
09:45 - 10:45	***Keynote Talk by Andrea E. Martin***
10:45 - 11:00	***Coffee Break***
11:00 - 12:30	***Session 1 (Oral Presentations)***
11:00-11:30	*Eye Gaze and Self-attention: How Humans and Transformers Attend Words in Sentences* Joshua Bensemann, Alex Yuxuan Peng, Diana Benavides Prado, Yang Chen, Neset Tan, Paul Michael Corballis, Patricia Riddle and Michael Witbrock
11:30-12:00	*Seeing the advantage: visually grounding word embeddings to better capture human semantic knowledge* Danny Merkx, Stefan Frank and Mirjam Ernestus
12:00-12:30	*Visually Grounded Interpretation of Noun-Noun Compounds in English* Inga Lang, Lonneke Van Der Plas, Malvina Nissim and Albert Gatt
12:30 - 13:30	***Lunch Break***
13:30 - 15:00	***Session 2 (Oral Presentations)***
13:30-14:00	*A Neural Model for Compositional Word Embeddings and Sentence Processing* Shalom Lappin and Jean-Philippe Bernardy
14:00-14:30	*Codenames as a Game of Co-occurrence Counting* Réka Cserháti, Istvan Kollath, András Kicsi and Gábor Berend
14:30-15:00	*About Time: Do Transformers Learn Temporal Verbal Aspect?* Eleni Metheniti, Tim Van De Cruys and Nabil Hathout
15:00 - 15:15	***Coffee Break***
15:15 - 15:30	***Shared Task Presentation***
15:15-15:30	*CMCL 2022 Shared Task on Multilingual and Crosslingual Prediction of Human Reading Behavior* Nora Hollenstein, Emmanuele Chersoni, Cassandra L Jacobs, Yohei Oseki, Laurent Prévot and Enrico Santus
15:30 - 17:00	***Poster Session***

15:30-17:00 *Estimating word co-occurrence probabilities from pretrained static embeddings using a log-bilinear model*
Richard Futrell

15:30-17:00 *Predicting scalar diversity with context-driven uncertainty over alternatives*
Jennifer Hu, Roger P. Levy and Sebastian Schuster

15:30-17:00 *Less Descriptive yet Discriminative: Quantifying the Properties of Multimodal Referring Utterances via CLIP*
Ece Takmaz, Sandro Pezzelle and Raquel Fernández

15:30-17:00 *Modeling the Relationship between Input Distributions and Learning Trajectories with the Tolerance Principle*
Jordan Kodner

15:30-17:00 *NU HLT at CMCL 2022 Shared Task: Multilingual and Crosslingual Prediction of Human Reading Behavior in Universal Language Space*
Joseph Marvin Imperial

15:30-17:00 *Team DMG at CMCL 2022 Shared Task: Transformer Adapters for the Multi- and Cross-Lingual Prediction of Human Reading Behavior*
Ece Takmaz

15:30-17:00 *Team ÚFAL at CMCL 2022 Shared Task: Figuring out the correct recipe for predicting Eye-Tracking features using Pretrained Language Models*
Sunit Bhattacharya, Rishu Kumar and Ondrej Bojar

15:30-17:00 *HkAmsters at CMCL 2022 Shared Task: Predicting Eye-Tracking Data from a Gradient Boosting Framework with Linguistic Features*
Lavinia Salicchi, Rong Xiang and Yu-Yin Hsu

15:30-17:00 *Poirot at CMCL 2022 Shared Task: Zero Shot Crosslingual Eye-Tracking Data Prediction using Multilingual Transformer Models*
Harshvardhan Srivastava

17:00 - 18:00 ***Keynote Talk by Vera Demberg***

18:00 - 18:15 ***Closing Remarks***

W8 - The 2nd DialDoc workshop on Document-grounded Dialogue and Conversational Question Answering

Organizers:
Song Feng, Chengguang Tang, Svitlana Vakulenko, Hui Wan, Zeqiu (Ellen) Wu, Caixia Yuan

https://doc2dial.github.io/workshop2022/
Venue: Liffey Meeting Room 2
Thursday, May 26, 2022

DialDoc Workshop focuses on document-grounded dialogue and conversational question answering. There is a vast amount of document content created every day by human writers to communicate with human readers for sharing knowledge, ranging from encyclopedias to social benefits. Making the document content accessible to users via conversational systems and scaling it to various domains could be a meaningful yet challenging task. There are significant individual research threads that show promise in handling heterogeous knowledge embedded in documents for building conversational systems, including (1) unstructured content, such as text passages; (2) semi-structured content, such as tables or lists; (3) multi-modal content, such as images and videos along with text descriptions, and so on. The purpose of this workshop is to invite researchers and practitioners to bring their individual perspectives on the subject of document-grounded dialogue and conversational question answering to advance the field in a joint effort.

09:00 - 09:05	***Opening Remark***
09:05 - 09:40	***Invited talk I by Siva Reddy***
09:40 - 10:25	***Paper presentation***
09:40-09:55	*Conversation- and Tree-Structure Losses for Dialogue Disentanglement* Tianda Li, Jia-Chen Gu, Zhen-Hua Ling and Quan Liu
09:55-10:10	*Construction of Hierarchical Structured Knowledge-based Recommendation Dialogue Dataset and Dialogue System* Takashi Kodama, Ribeka Tanaka and Sadao Kurohashi
10:10-10:25	*Retrieval-Free Knowledge-Grounded Dialogue Response Generation with Adapters* Yan Xu, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan SU and Pascale Fung
10:25 - 10:40	***Coffee break***
10:40 - 11:15	***Invited talk II by Jeff Dalton***
11:15 - 11:55	***Paper lightning talk I***
11:15-11:20	*TRUE: Re-evaluating Factual Consistency Evaluation* Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim and Yossi Matias
11:20-11:25	*Pseudo Ambiguous and Clarifying Questions Based on Sentence Structures Toward Clarifying Question Answering System* Yuya Nakano, Seiya Kawano, Koichiro Yoshino, Katsuhito Sudoh and Satoshi Nakamura
11:25-11:30	*Parameter-Efficient Abstractive Question Answering over Tables or Text* Vaishali Pal, Evangelos Kanoulas and Maarten de Rijke
11:30-11:35	*Conversational Search with Mixed-Initiative - Asking Good Clarification Questions backed-up by Passage Retrieval*

	Yosi Mass, Doron Cohen, Asaf Yehudai and David Konopnicki
11:35-11:40	*Graph-combined Coreference Resolution Methods on Conversational Machine Reading Comprehension with Pre-trained Language Model* Zhaodong Wang and Kazunori Komatani
11:40-11:45	*G4: Grounding-guided Goal-oriented Dialogues Generation with Multiple Documents* Shiwei Zhang, Yiyang Du, Guanzhong Liu, Zhao Yan and Yunbo Cao
11:45-11:50	*UniDS: A Unified Dialogue System for Chit-Chat and Task-oriented Dialogues* Xinyan Zhao, Bin He, Yasheng Wang, Yitong Li, Fei Mi, Yajiao Liu, Xin Jiang, Qun Liu and Huanhuan Chen
11:50-11:55	*MSAMSum: Towards Benchmarking Multi-lingual Dialogue Summarization* Xiachong Feng, Xiaocheng Feng and Bing Qin
11:55 - 12:55	***Poster session***
12:55 - 14:00	***Lunch break***
14:00 - 14:35	***Invited talk III by Zhou Yu***
15:05 - 14:35	***Paper presentation***
14:35-14:50	*Low-Resource Adaptation of Open-Domain Generative Chatbots* Greyson Gerhard-Young, Raviteja Anantha, Srinivas Chappidi and Bjorn Hoffmeister
15:05 - 15:20	***Coffee break***
15:20 - 15:35	***DialDoc 2022 Shared Task***
15:35 - 15:50	***Shared Task Prizes and Best Paper Awards presented by Luis Lastras***
15:50 - 16:25	***Invited talk IV by Michel Galley***
16:25 - 17:00	***Invited talk V by Mari Ostendorf***
17:00 - 17:05	***Ending Remark***

W9 - The 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2022)

Organizers:
Jeremy Barnes, Orphée De Clercq, Valentin Barriere, Shabnam Tafreshi, Sawsan Alqahtani, João Sedoc, Roman Klinger, Alexandra Balahur

https://wassa-workshop.github.io/
Venue: Wicklow Hall 1
Thursday, May 26, 2022

Starting with reviews on products on e-commerce sites and ending with the emotional effect presentin or intended by media coverage, research in automatic Subjectivity and Sentiment Analysis as wellas explicit and implicit Emotion Detection and Classification has flourished in the past years. Theimportance of the field has been proven by the high number of approaches proposed in research in thepast decade, as well as by the interest it generated in other disciplines, such as Economics, Sociology,Psychology, Marketing, Crisis Management Digital Humanities. Building on previous editions, the aim of WASSA 2022 is to bring together researchers workingon Subjectivity, Sentiment Analysis, Emotion Detection and Classification and their applications toother NLP or real-world tasks (e.g. public health messaging, fake news, media impact analysis, socialmedia mining, computational literary studies) and researchers working on interdisciplinary aspectsof affect computation from text.

09:00 - 09:10	***Opening Remarks***
09:10 - 10:30	***Session 1***
	Assessment of Massively Multilingual Sentiment Classifiers Krzysztof Rajda, Lukasz Augustyniak, Piotr Gramacki, Marcin Gruza, Szymon Woźniak and Tomasz Jan Kajdanowicz
	English-Malay Word Embeddings Alignment for Cross-lingual Emotion Classification with Hierarchical Attention Network Ying Hao Lim and Jasy Suet Yan Liew
	Uncertainty Regularized Multi-Task Learning Kourosh Meshgi, Maryam Sadat Mirzaei and Satoshi Sekine
	Improving Social Meaning Detection with Pragmatic Masking and Surrogate Fine-Tuning Chiyu Zhang and Muhammad Abdul-Mageed
10:30 - 11:00	***Coffee Break***
11:00 - 12:00	***Shared Task***
	WASSA 2022 Shared Task: Predicting Empathy, Emotion and Personality in Reaction to News Stories Valentin Barriere, Shabnam Tafreshi, João Sedoc and Sawsan Alqahtani
	IUCL at WASSA 2022 Shared Task: A Text-only Approach to Empathy and Emotion Detection Yue Chen, Yingnan Ju and Sandra Kübler
	Continuing Pre-trained Model with Multiple Training Strategies for Emotional Classification Bin Li, Yixuan Weng, Qiya Song, Bin Sun and Shutao Li
12:00 - 13:00	***Invited Talk 1 - Dirk Hovy***
13:00 - 14:00	***Lunch Break***

14:00 - 15:00	***Invited Talk 2 - Rada Mihalcea***
15:00 - 15:30	***Break***
15:30 - 16:15	***In-Person Poster Session***
	On the Complementarity of Images and Text for the Expression of Emotions in Social Media Anna Khlyzova, Carina Silberer and Roman Klinger
	Multiplex Anti-Asian Sentiment before and during the Pandemic: Introducing New Datasets from Twitter Mining Hao Lin, Pradeep Kumar Nalluri, Lantian Li, Yifan Sun and Yongjun Zhang
	"splink" is happy and "phrouth" is scary: Emotion Intensity Analysis for Nonsense Words Valentino Sabbatino, Enrica Troiano, Antje Schweitzer and Roman Klinger
	Irony Detection for Dutch: a Venture into the Implicit Aaron Maladry, Els Lefever, Cynthia Van Hee and Veronique Hoste
	Items from Psychometric Tests as Training Data for Personality Profiling Models of Twitter Users Anne Kreuter, Kai Sassenberg and Roman Klinger
16:15 - 17:15	***Session 2***
	Distinguishing In-Groups and Onlookers by Language Use Joshua R Minot, Milo Z Trujillo, Samuel F Rosenblatt, Guillermo De Anda-Jáuregui, Emily Moog, Allison M. Roth, Briane Paul Samson and Laurent Hébert-Dufresne
	"splink" is happy and "phrouth" is scary: Emotion Intensity Analysis for Nonsense Words Valentino Sabbatino, Enrica Troiano, Antje Schweitzer and Roman Klinger
	Can Emotion Carriers Explain Automatic Sentiment Prediction? A Study on Personal Narratives Seyed Mahed Mousavi, Gabriel Roccabruna, Aniruddha Tammewar, Steve Azzolin and Giuseppe Riccardi
17:15 - 18:00	***Virtual Poster Session***
	On the Complementarity of Images and Text for the Expression of Emotions in Social Media Anna Khlyzova, Carina Silberer and Roman Klinger
	Multiplex Anti-Asian Sentiment before and during the Pandemic: Introducing New Datasets from Twitter Mining Hao Lin, Pradeep Kumar Nalluri, Lantian Li, Yifan Sun and Yongjun Zhang
	Domain-Aware Contrastive Knowledge Transfer for Multi-domain Imbalanced Data Zixuan Ke, Mohammad Kachuee and Sungjin Lee
	Infusing Knowledge from Wikipedia to Enhance Stance Detection Zihao He, Negar Mokhberian and Kristina Lerman
	Understanding BERT's Mood: The Role of Contextual-Embeddings as User-Representations for Depression Assessment Matthew Matero, Albert Hung and H. Schwartz
	Emotion Analysis of Writers and Readers of Japanese Tweets on Vaccinations Patrick John Co Ramos, Kiki Ferawati, Kongmeng Liew, Eiji Aramaki and Shoko Wakamiya
	Opinion-based Relational Pivoting for Cross-domain Aspect Term Extraction Ayal Klein, Oren Pereg, Daniel Korat, Vasudev Lal, Moshe Wasserblat and Ido Dagan
	Pushing on Personality Detection from Verbal Behavior: A Transformer Meets Text Contours of Psycholinguistic Features Elma Kerz, Yu Qiao, Sourabh Zanwar and Daniel Wiechmann
	XLM-EMO: Multilingual Emotion Prediction in Social Media Text Federico Bianchi, Debora Nozza and Dirk Hovy
	Evaluating Content Features and Classification Methods for Helpfulness Prediction of Online Reviews: Establishing a Benchmark for Portuguese Rogério Figueredo Sousa and Thiago A. S. Pardo

W10 - The First Workshop on Learning with Natural Language Supervision

Organizers:
Jacob Andreas, Karthik Narasimhan, Aida Nematzadeh

https://sites.google.com/princeton.edu/nl-supervision
Venue: Wicklow Meeting Room 1
Thursday, May 26, 2022

While many approaches to language supervision are domain-general, and closely connected to "core" NLP research, there are currently no venues where researchers from across the field can meet to share ideas and draw connections between their disparate lines of research. Our workshop will offer a central meeting point for research on language-based supervision, enabling researchers within and beyond NLP to discuss how language processing models and algorithms can be brought to bear on problems beyond the textual realm (e.g. visual recognition, robotics, program synthesis, sequential decision making). In keeping with this interdisciplinary focus, our proposed format differs in two ways from a standard NLP workshop: first, with a special emphasis on speakers and attendees who would not typically attend NLP conferences; second, by replacing the standard panel discussion with a series of workshop-wide breakout sessions aimed at seeding cross-institutional collaborations around new tasks, datasets, and models.

09:00 - 09:30	***Opening Remarks***
09:30 - 10:00	***Invited talk: Hinrich Schuetze***
10:00 - 10:30	***Spotlight presentations***
10:30 - 11:00	***Coffee break***
11:00 - 12:00	***Breakout session 1 (datasets) and recap***
12:00 - 12:30	***Invited talk: Jeanette Bohg***
12:30 - 14:00	***Lunch break***
14:00 - 15:00	***Poster presentations***
15:00 - 15:30	***Coffee break***
15:30 - 16:00	***Invited talk: Justin Johnson***
16:00 - 16:30	***Invited talk: Anna Ivanova***
16:30 - 17:00	***Invited talk: Hannaneh Hajishirzi***
17:00 - 18:00	***Breakout session 2 (models) and recap***

W11 - The First Workshop on Intelligent and Interactive Writing Assistants (In2Writing)

Organizers:
Katy Ilonka Gero, Dongyeop Kang, Ting-Hao 'Kenneth' Huang, Mina Lee, Daniel Gissin, John Joon Young Chung, Vipul Raheja

https://in2writing.glitch.me/
Venue: Liffey Meeting Room 1
Thursday, May 26, 2022

The purpose of this interdisciplinary workshop is to bring together researchers from the natural language processing (NLP) and human-computer interaction (HCI) communities as well as industry practitioners and professional writers to discuss innovations in building, improving, and evaluating intelligent and interactive writing assistants. We plan to alternate our workshop venue between an NLP conference and a HCI conference every year to facilitate collaboration.

09:00 - 09:10	***Opening Remarks***
09:10 - 10:10	***Invited talk - Professional writer (Claire L. Evans)***
10:10 - 10:30	***Invited talk - Industry (Daniel Gissin)***
10:30 - 11:00	***Coffee break***
11:00 - 11:15	***Invited talk - Academia (Melissa Roemmele)***
11:15 - 11:30	***Invited talk - Academia (Elizabeth Clark)***
11:30 - 12:30	***Panel discussion 1 - Understanding the impact of writing assistants on ownership, authenticity, originality, and confidence with Jill Burstein, Courtney Napoles, Ekaterina Kochmar, and Dashiel Carrera.***
12:30 - 14:00	***Lunch break***
14:00 - 15:00	***Invited talk - Professional writer (Lillian-Yvonne Bertram)***
15:00 - 15:30	***Coffee break***
15:30 - 16:00	***Invited talk - Industry (Timo Mertens)***
16:00 - 16:15	***Invited talk - Best paper***
16:15 - 17:15	***Panel discussion 2 - Bridging NLP and HCI to design, build, and evaluate writing assistants with Melissa Roemmele, Courtney Napoles, Qian Yang, and Sherry Wu.***
17:15 - 18:00	***Demo and poster session***

W12 - The Third Workshop on Insights from Negative Results in NLP

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

https://insights-workshop.github.io/index
Venue: The Liffey A
Thursday, May 26, 2022

Publication of negative results is difficult in most fields, but in NLP the problem is exacerbated by the near-universal focus on improvements in benchmarks. This situation implicitly discourages hypothesis-driven research, and it turns creation and fine-tuning of NLP models into art rather than science. Furthermore, it increases the time, effort, and carbon emissions spent on developing and tuning models, as the researchers have no opportunity to learn what has already been tried and failed. This workshop invites both practical and theoretical unexpected or negative results that have important implications for future research, highlight methodological issues with existing approaches, and/or point out pervasive misunderstandings or bad practices. In particular, the most successful NLP models currently rely on different kinds of pretrained meaning representations (from word embeddings to Transformer-based models like BERT). To complement all the success stories, it would be insightful to see where and possibly why they fail. Any NLP tasks are welcome: sequence labeling, question answering, inference, dialogue, machine translation - you name it.

08:45 - 09:00	***Opening Remarks***
09:00 - 10:00	***Invited Talk: Barbara Plank***
10:30 - 11:00	***Coffee Break***
11:00 - 11:30	***Thematic Session 1: Linguistically Informed Analysis***
	Do Dependency Relations Help in the Task of Stance Detection? Alessandra Teresa Cignarella, Cristina Bosco and Paolo Rosso
	BPE beyond Word Boundary: How NOT to use Multi Word Expressions in Neural Machine Translation Dipesh Kumar and Avijit Thawani
	Challenges in including extra-linguistic context in pre-trained language models Ionut Teodor Sorodoc, Laura Aina and Gemma Boleda
11:30 - 12:00	***Thematic Session 2: Transformers***
	How Much Do Modifications to Transformer Language Models Affect Their Ability to Learn Linguistic Knowledge? Simeng Sun, Brian Dillon and Mohit Iyyer
	Pathologies of Pre-trained Language Models in Few-shot Fine-tuning Hanjie Chen, Guoqing Zheng, Ahmed Hassan Awadallah and Yangfeng Ji
	On Isotropy Calibration of Transformer Models Yue Ding, Karolis Martinkus, Damian Pascual, Simon Clematide and Roger Wattenhofer
12:00 - 12:30	***Thematic Session 3: Towards Better Data***

	Do Data-based Curricula Work? Maxim K. Surkov, Vladislav D. Mosin and Ivan P. Yamshchikov
	Clustering Examples in Multi-Dataset Benchmarks with Item Response Theory Pedro Rodriguez, Phu Mon Htut, John P. Lalor and João Sedoc
	On the Impact of Data Augmentation on Downstream Performance in Natural Language Processing Itsuki Okimura, Machel Reid, Makoto Kawano and Yutaka Matsuo
12:30 - 14:00	***Lunch***
14:00 - 15:00	***Panel Discussion***
15:00 - 15:30	***Coffee Break***
10:00 - 10:30	***Thematic Session 4: Improving Evaluation Practices***
	Replicability under Near-Perfect Conditions – A Case-Study from Automatic Summarization Margot Mieskes
	On the Limits of Evaluating Embodied Agent Model Generalization Using Validation Sets Hyounghun Kim, Aishwarya Padmakumar, Di Jin, Mohit Bansal and Dilek Hakkani-Tur
16:00 - 17:00	***Invited Talk: Tal Linzen***
17:00 - 18:00	***Poster Session***
18:00 - 18:10	***Closing Remarks***

W13 - The 7th Workshop on Representation Learning for NLP

Organizers:
Spandana Gella, He He, Burcu Can, Maximilian Mozes, Eleonora Giunchiglia, Sewon Min, Samuel Cahyawijaya, Xiang Lorraine Li and Bodhisattwa Prasad Majumder

https://sites.google.com/view/repl4nlp2022/
Venue: The Liffey B
Thursday, May 26, 2022

The workshop is being organised by Spandana Gella, He He, Burcu Can, Maximilian Mozes, Eleonora Giunchiglia, Sewon Min, Samuel Cahyawijaya, Xiang Lorraine Li and Bodhisattwa Prasad Majumder; and advised by Isabelle Augenstein, Anna Rogers, Kyunghyun Cho, Edward Grefenstette, Chris Dyer and Laura Rimell. The workshop is organised by the ACL Special Interest Group on Representation Learning (SIGREP). The 7th Workshop on Representation Learning for NLP aims to continue the success of the Repl4NLP workshop series, with the 1st Workshop on Representation Learning for NLP having received about 50 submissions and over 250 attendees – the second most attended collocated event at ACL'16 after WMT. The workshop was introduced as a synthesis of several years of independent *CL workshops focusing on vector space models of meaning, compositionality, and the application of deep neural networks and spectral methods to NLP. It provides a forum for discussing recent advances on these topics, as well as future research directions in linguistically motivated vector-based models in NLP. The workshop will take place in a hybrid setting, and, as in previous years, feature interdisciplinary keynotes, paper presentations, posters, as well as a panel discussion.

08:50 - 09:00	***Opening remark***
09:00 - 09:45	***Invited talk 1":" Sebastian Riedel***
09:45 - 10:30	***Invited talk 2":" Monojit Choudhury***
10:30 - 10:45	***Outstanding Papers Spotlight Presentations***
10:45 - 11:00	***Coffee break***
11:00 - 12:00	***Poster session I (Virtual and in-person)***
	ANNA":" Enhanced Language Representation for Question Answering Changwook Jun, Hansol Jang, Myoseop Sim, Hyun Kim, Jooyoung Choi, Kyungkoo Min and Kyunghoon Bae
	Binary Encoded Word Mover's Distance Christian Johnson
	Clozer":" Adaptable Data Augmentation for Cloze-style Reading Comprehension Holy Lovenia, Bryan Wilie, Willy Chung, Zeng Min, Samuel Cahyawijaya, Dan Su and Pascale Fung
	Q-Learning Scheduler for Multi Task Learning Through the use of Histogram of Task Uncertainty Kourosh Meshgi, Maryam Sadat Mirzaei and Satoshi Sekine
	TRAttack":" Text Rewriting Attack Against Text Retrieval Junshuai Song, Jiangshan Zhang, Jifeng Zhu, Mengyun Tang and Yong Yang
	From Hyperbolic Geometry Back to Word Embeddings

Zhenisbek Assylbekov, Sultan Nurmukhamedov, Arsen Sheverdin and Thomas Mach

On Target Representation in Continuous-output Neural Machine Translation
Evgeniia Tokarchuk and Vlad Niculae

Detecting Textual Adversarial Examples Based on Distributional Characteristics of Data Representations
Na Liu, Mark Dras and Wei Emma Zhang

A Comparative Study of Pre-trained Encoders for Low-Resource Named Entity Recognition
Yuxuan Chen, Jonas Mikkelsen, Arne Binder, Christoph Alt and Leonhard Hennig

Towards Improving Selective Prediction Ability of NLP Systems
Neeraj Varshney, Swaroop Mishra and Chitta Baral

When does CLIP generalize better than unimodal models? When judging human-centric concepts
Romain Bielawski, Benjamin Devillers, Tim Van De Cruys and Rufin Vanrullen

A Vocabulary-Free Multilingual Neural Tokenizer for End-to-End Task Learning
Md Mofijul Islam, Gustavo Aguilar, Pragaash Ponnusamy, Clint Solomon Mathialagan, Chengyuan Ma and Chenlei Guo

PALBERT":" Teaching ALBERT to Ponder
Daniil Gavrilov and Nikita Balagansky

Same Author or Just Same Topic? Towards Content-Independent Style Representations
Anna Wegmann, Marijn Schraagen and Dong Nguyen

12:00 - 13:30 ***Lunch break***

13:30 - 14:15 ***Invited talk 3":" Emma Strubell***

14:15 - 15:15 ***Poster session II (Virtual and in-person)***

Zero-shot Cross-lingual Transfer is Under-specified Optimization
Shijie Wu, Benjamin Van Durme and Mark Dredze

Identifying the Limits of Cross-Domain Knowledge Transfer for Pretrained Models
Zhengxuan Wu, Nelson F. Liu and Christopher Potts

Detecting Word-Level Adversarial Text Attacks via SHapley Additive exPlanations
Edoardo Mosca, Lukas Huber, Marc Alexander Kühn and Georg Groh

Isomorphic Cross-lingual Embeddings for Low-Resource Languages
Sonal Sannigrahi and Jesse Read

Unsupervised Geometric and Topological Approaches for Cross-Lingual Sentence Representation and Comparison
Shaked Haim Meirom and Omer Bobrowski

On the Geometry of Concreteness
Christian Wartena

Temporal Knowledge Graph Reasoning with Low-rank and Model-agnostic Representations
Ioannis Dikeoulias, Saadullah Amin and Günter Neumann

PARADISE":" Exploiting Parallel Data for Multilingual Sequence-to-Sequence Pretraining
Machel Reid and Mikel Artetxe

Distributionally Robust Recurrent Decoders with Random Network Distillation
Antonio Valerio Miceli Barone, Alexandra Birch and Rico Sennrich

Video Language Co-Attention with Multimodal Fast-Learning Feature Fusion for VideoQA
Adnen Abdessaied, Ekta Sood and Andreas Bulling

WeaNF":" Weak Supervision with Normalizing Flows
Andreas Stephan and Benjamin Roth

A Study on Entity Linking Across Domains":" Which Data is Best for Fine-Tuning?
Hassan Soliman, Heike Adel, Mohamed H. Gad-Elrab, Dragan Milchevski and Jannik Strötgen

Analyzing Gender Representation in Multilingual Models

	Hila Gonen, Shauli Ravfogel and Yoav Goldberg
	Lacking the embedding of a word? Look it up into a traditional dictionary Elena Sofia Ruzzetti, Leonardo Ranaldi, Michele Mastromattei, Francesca Fallucchi, Noemi Scarpato and Fabio Massimo Zanzotto
15:15 - 15:30	***Coffee break***
15:30 - 16:15	***Invited talk 4":" Been Kim***
16:15 - 17:00	***Panel discussion***
17:00 - 17:45	***Invited talk 5":" Percy Liang***
17:45 - 17:50	***Closing remark***

W14 - The Second Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI-2022)

Organizers:
Bharathi Raja Chakravarthi, Bharathi B, John P. McCrae, Manel Zarrouk, Kalika Bali, Paul Buitelaar

https://sites.google.com/view/lt-edi-2022/home
Venue: Liffey Meeting Room 3
Friday, May 27, 2022

Equality, Diversity and Inclusion (EDI) is an important agenda across every field throughout the world. Language as a major part of communication should be inclusive and treat everyone with equality. Today's large internet community uses language technology (LT) and has a direct impact on people across the globe. EDI is crucial to ensure everyone is valued and included, so it is necessary to build LT that serves this purpose. Recent results have shown that big data and deep learning are entrenching existing biases and that some algorithms are even naturally biased due to problems such as 'regression to the mode'. Our focus is on creating LT that will be more inclusive of gender, racial, sexual orientation, persons with disability. The workshop will focus on creating speech and language technology to address EDI not only in English, but also in less resourced languages.

09:00 - 09:15	***Opening Remarks***
09:15 - 10:30	***Session 1*** - Chair: Paul Buitelaar, National University of Ireland Galway, Ireland
09:15-09:30	*Mind the data gap(s): Investigating power in speech and language datasets* Nina Markl
09:45-10:00	*Detoxifying Language Models with a Toxic Corpus* Yoona Park and Frank Rudzicz
10:00-10:15	*Inferring Gender: A Scalable Methodology for Gender Detection with Online Lexical Databases* Marion Bartl and Susan Leavy
10:15-10:30	*Debiasing Pre-Trained Language Models via Efficient Fine-Tuning* Michael Gira, Ruisu Zhang and Kangwook Lee
10:30 - 11:00	***Coffee Break***
11:00 - 11:30	***Invited talk***
11:00 - 12:30	***Session 2*** - Chair: John P. McCrae, National University of Ireland Galway, Ireland
11:30-11:45	*Disambiguation of morpho-syntactic features of African American English – the case of habitual be* Harrison Santiago, Joshua Martin, Sarah Moeller and Kevin Tang
11:45-12:00	*Behind the Mask: Demographic bias in name detection for PII masking* Courtney Mansfield, Amandalynne Paullada and Kristen Howell
12:00-12:15	*Mapping the Multilingual Margins: Intersectional Biases of Sentiment Analysis Systems in English, Spanish, and Arabic* António Câmara, Nina Taneja, Tamjeed Azad, Emily Allaway and Richard Zemel
12:15-12:30	*Monte Carlo Tree Search for Interpreting Stress in Natural Language* Kyle Swanson, Joy Hsu and Mirac Suzgun
12:30 - 14:00	***Lunch Break***

14:00 - 15:00 ***Session 3*** - Chair: B. Bharathi, SSN College of Engineering, India

14:00-14:15 *The Best of both Worlds: Dual Channel Language modeling for Hope Speech Detection in low-resourced Kannada*
Adeep Hande, Siddhanth U Hegde, Sangeetha S, Ruba Priyadharshini and Bharathi Raja Chakravarthi

14:15-14:25 *Findings of the Shared Task on Detecting Signs of Depression from Social Media*
Kayalvizhi S, Thenmozhi Durairaj, Bharathi Raja Chakravarthi and Jerin Mahibha C

14:25-14:35 *Findings of the Shared Task on Speech Recognition for Vulnerable Individuals in Tamil*
Bharathi B, Bharathi Raja Chakravarthi, Subalalitha CN, Sripriya N, Arunaggiri Pandian and Swetha Valli

14:35-14:45 *Overview of The Shared Task on Homophobia and Transphobia Detection in Social Media Comments*
Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John Philip McCrae, Paul Buitelaar, Prasanna Kumar Kumaresan and Rahul Ponnusamy

14:45-15:00 *Overview of the Shared Task on Hope Speech Detection for Equality, Diversity, and Inclusion*
Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, Subalalitha CN, John Philip McCrae, Miguel Ángel García, Salud María Jiménez-Zafra, Rafael Valencia-García, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Daniel García-Baena and José Antonio García-Díaz

15:00 - 15:30 ***Coffee Break***

15:30 - 17:30 ***Posters*** - Chair: Thenmozhi Durairaj, SSN College of Engineering, India

15:30-17:30 *Regex in a Time of Deep Learning: The Role of an Old Technology in Age Discrimination Detection in Job Advertisements*
Anna Pillar, Kyrill Poelmans and Martha Larson

15:30-17:30 *Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals*
Debora Nozza, Federico Bianchi, Anne Lauscher and Dirk Hovy

15:30-17:30 *Using BERT Embeddings to Model Word Importance in Conversational Transcripts for Deaf and Hard of Hearing Users*
Akhter Al Amin, Saad Hassan, Cecilia Alm and Matt Huenerfauth

15:30-17:30 *IIITSurat@LT-EDI-ACL2022: Hope Speech Detection using Machine Learning*
Pradeep Kumar Roy, Snehaan Bhawal, Abhinav Kumar and Bharathi Raja Chakravarthi

15:30-17:30 *NYCU_TWD@LT-EDI-ACL2022: Ensemble Models with VADER and Contrastive Learning for Detecting Signs of Depression from Social Media*
Wei-Yao Wang, Yu-Chien Tang, Wei-Wei Du and Wen-Chih Peng

15:30-17:30 *UMUTeam@LT-EDI-ACL2022: Detecting homophobic and transphobic comments in Tamil*
José Antonio García-Díaz, Camilo Caparros-Laiz and Rafael Valencia-García

15:30-17:30 *UMUTeam@LT-EDI-ACL2022: Detecting Signs of Depression from text*
José Antonio García-Díaz and Rafael Valencia-García

15:30-17:30 *bitsa_nlp@LT-EDI-ACL2022: Leveraging Pretrained Language Models for Detecting Homophobia and Transphobia in Social Media Comments*
Vitthal Bhandari and Poonam Goyal

15:30-17:30 *ABLIMET @LT-EDI-ACL2022: A Roberta based Approach for Homophobia/Transphobia Detection in Social Media*
Abulimiti Maimaitituoheti

15:30-17:30 *MUCIC@LT-EDI-ACL2022: Hope Speech Detection using Data Re-Sampling and 1D Conv-LSTM*
Anusha M D Gowda, Fazlourrahman Balouchzahi, Hosahalli Lakshmaiah Shashirekha and Grigori Sidorov

15:30-17:30 *DeepBlues@LT-EDI-ACL2022: Depression level detection modelling through domain specific BERT and short text Depression classifiers*
Nawshad Farruque, Osmar Zaiane, Randy Goebel and Sudhakar Sivapalan

15:30-17:30 *SSN_ARMM@ LT-EDI -ACL2022: Hope Speech Detection for Equality, Diversity, and Inclusion Using ALBERT model*
Praveenkumar Vijayakumar, Prathyush S, Aravind P, Angel Deborah S, Rajalakshmi Sivanaiah, Sakaya Milton Rajendram and Mirnalinee T T

15:30-17:30 *SUH_ASR@LT-EDI-ACL2022: Transformer based Approach for Speech Recognition for Vulnerable Individuals in Tamil*
Suhasini S and Bharathi B

15:30-17:30 *LPS@LT-EDI-ACL2022:An Ensemble Approach about Hope Speech Detection*
Yue Ying Zhu

15:30-17:30 *CURAJ_IIITDWD@LT-EDI-ACL 2022: Hope Speech Detection in English YouTube Comments using Deep Learning Techniques*
Vanshita Jha, Ankit Kumar Mishra and Sunil Saumya

15:30-17:30 *SSN_MLRG3 @LT-EDI-ACL2022-Depression Detection System from Social Media Text using Transformer Models*
Sarika Esackimuthu, Shruthi Hariprasad, Rajalakshmi Sivanaiah, Angel Deborah S, Sakaya Milton Rajendram and Mirnalinee T T

15:30-17:30 *BERT 4EVER@LT-EDI-ACL2022-Detecting signs of Depression from Social Media Detecting Depression in Social Media using Prompt-Learning and Word-Emotion Cluster*
Xiaotian Lin, Yingwen Fu, Ziyu Yang, Nankai Lin and Shengyi Jiang

15:30-17:30 *CIC@LT-EDI-ACL2022: Are transformers the only hope? Hope speech detection for Spanish and English comments*
Fazlourrahman Balouchzahi, Sabur Butt, Grigori Sidorov and Alexander Gelbukh

15:30-17:30 *scubeMSEC@LT-EDI-ACL2022: Detection of Depression using Transformer Models*
Sivamanikandan S, Santhosh V, Sanjaykumar N, Jerin Mahibha C and Thenmozhi Durairaj

15:30-17:30 *SSNCSE_NLP@LT-EDI-ACL2022:Hope Speech Detection for Equality, Diversity and Inclusion using sentence transformers*
Bharathi B, Dhanya Srinivasan, Josephine Varsha, Thenmozhi Durairaj and Senthil Kumar B

15:30-17:30 *SOA_NLP@LT-EDI-ACL2022: An Ensemble Model for Hope Speech Detection from YouTube Comments*
Abhinav Kumar, Sunil Saumya and Pradeep Kumar Roy

15:30-17:30 *IIT Dhanbad @LT-EDI-ACL2022- Hope Speech Detection for Equality, Diversity, and Inclusion*
Vishesh Gupta, Ritesh Kumar and Rajendra Pamula

15:30-17:30 *IISERB@LT-EDI-ACL2022: A Bag of Words and Document Embeddings Based Framework to Identify Severity of Depression Over Social Media*
Tanmay Basu

15:30-17:30 *SSNCSE_NLP@LT-EDI-ACL2022: Homophobia/Transphobia Detection in Multiple Languages using SVM Classifiers and BERT-based Transformers*
Krithika Swaminathan, Bharathi B, Gayathri G L and Hrishik Sampath

15:30-17:30 *KUCST@LT-EDI-ACL2022: Detecting Signs of Depression from Social Media Text*
Manex Agirrezabal and Janek Amann

15:30-17:30 *E8-IJS@LT-EDI-ACL2022 - BERT, AutoML and Knowledge-graph backed Detection of Depression*
Ilija Tavchioski, Boshko Koloski, Blaž Škrlj and Senja Pollak

15:30-17:30 *Nozza@LT-EDI-ACL2022: Ensemble Modeling for Homophobia and Transphobia Detection*
Debora Nozza

15:30-17:30 *KADO@LT-EDI-ACL2022: BERT-based Ensembles for Detecting Signs of Depression from Social Media Text*
Morteza Janatdoust, Fatemeh Ehsani-Besheli and Hossein Zeinali

15:30-17:30 *Sammaan@LT-EDI-ACL2022: Ensembled Transformers Against Homophobia and Transphobia*
Ishan Sanjeev Upadhyay, Kv Aditya Srivatsa and Radhika Mamidi

15:30-17:30 *OPI@LT-EDI-ACL2022: Detecting Signs of Depression from Social Media Text using RoBERTa Pre-trained Language Models*
Rafał Poświata and Michał Wiktor Perełkiewicz

15:30-17:30 *FilipN@LT-EDI-ACL2022-Detecting signs of Depression from Social Media: Examining the use of summarization methods as data augmentation for text classification*
Filip Nilsson and György Kovács

15:30-17:30 *NAYEL @LT-EDI-ACL2022: Homophobia/Transphobia Detection for Equality, Diversity, and Inclusion using SVM*
Nsrin Ashraf, Mohamed Taha, Ahmed Taha Abd Elfattah and Hamada Nayel

15:30-17:30 *giniUs @LT-EDI-ACL2022: Aasha: Transformers based Hope-EDI*
Harshul Raj Surana and Basavraj Chinagundi

15:30-17:30 *SSN_MLRG1@LT-EDI-ACL2022: Multi-Class Classification using BERT models for Detecting Depression Signs from Social Media Text*
Karun Anantharaman, Angel Deborah S, Rajalakshmi Sivanaiah, Saritha Madhavan and Sakaya Milton Rajendram

15:30-17:30 *DepressionOne@LT-EDI-ACL2022: Using Machine Learning with SMOTE and Random UnderSampling to Detect Signs of Depression on Social Media Text.*
Suman Dowlagar and Radhika Mamidi

15:30-17:30 *LeaningTower@LT-EDI-ACL2022: When Hope and Hate Collide*
Arianna Muti, Marta Marchiori Manerba, Katerina Korre and Alberto Barrón-Cedeño

15:30-17:30 *MUCS@Text-LT-EDI@ACL 2022: Detecting Sign of Depression from Social Media Text using Supervised Learning Approach*
Asha Hegde, Sharal Coelho, Ahmad Elyas Dashti and Hosahalli Lakshmaiah Shashirekha

15:30-17:30 *SSNCSE_NLP@LT-EDI-ACL2022: Speech Recognition for Vulnerable Individuals in Tamil using pre-trained XLSR models*
Dhanya Srinivasan, Bharathi B, Thenmozhi Durairaj and Senthil Kumar B

15:30-17:30 *IDIAP_TIET@LT-EDI-ACL2022 : Hope Speech Detection in Social Media using Contextualized BERT with Attention Mechanism*
Deepanshu Khanna, Muskaan Singh and Petr Motlicek

15:30-17:30 *SSN@LT-EDI-ACL2022: Transfer Learning using BERT for Detecting Signs of Depression from Social Media Texts*
Adarsh S and Betina Antony

15:30-17:30 *DLRG@LT-EDI-ACL2022:Detecting signs of Depression from Social Media using XGBoost Method*
Herbert Goldwin Sharen and Ratnavel Rajalakshmi

15:30-17:30 *IDIAP Submission@LT-EDI-ACL2022 : Hope Speech Detection for Equality, Diversity and Inclusion*
Muskaan Singh and Petr Motlicek

15:30-17:30 *IDIAP Submission@LT-EDI-ACL2022: Homophobia/Transphobia Detection in social media comments*
Muskaan Singh and Petr Motlicek

15:30-17:30 *IDIAP Submission@LT-EDI-ACL2022: Detecting Signs of Depression from Social Media Text*
Muskaan Singh and Petr Motlicek

17:30 - 18:00 ***Closing Remarks***

W15 - The Second Workshop on Human Evaluation of NLP Systems (HumEval 2022)

Organizers:
Anya Belz, Maja Popović, Ehud Reiter, Anastasia Shimorina

https://humeval.github.io/
Venue: Wicklow Hall 2b
Friday, May 27, 2022

With this workshop, we wish to create a forum for current human evaluation research and future directions, a space for researchers working with human evaluations to exchange ideas and begin to address the issues human evaluation in NLP faces from many points of view, including experimental design, meta-evaluation and reproducibility.

09:00 - 10:00	***Invited talk by Markus Freitag***
10:00 - 10:30	***Session 1***
10:00-10:10	*A Methodology for the Comparison of Human Judgments With Metrics for Coreference Resolution* Mariya Borovikova, Loïc Grobol, Anaïs Lefeuvre Halftermeyer and Sylvie Billot
10:10-10:20	*Perceptual Quality Dimensions of Machine-Generated Text with a Focus on Machine Translation* Vivien Macketanz, Babak Naderi, Steven Schmidt and Sebastian Möller
10:20-10:30	*Towards Human Evaluation of Mutual Understanding in Human-Computer Spontaneous Conversation: An Empirical Study of Word Sense Disambiguation for Naturalistic Social Dialogs in American English* Alex Luu
10:30 - 11:00	***Coffee Break***
11:00 - 12:20	***Session 2***
11:00-11:20	*A Study on Manual and Automatic Evaluation for Text Style Transfer: The Case of Detoxification* Varvara Logacheva, Daryna Dementieva, Irina Krotova, Alena Fenogenova, Irina Nikishina, Tatiana Shavrina and Alexander Panchenko
11:20-11:40	*Beyond calories: evaluating how tailored communication reduces emotional load in diet-coaching* Simone Balloccu and Ehud Reiter
11:40-12:00	*Human Judgement as a Compass to Navigate Automatic Metrics for Formality Transfer* Huiyuan Lai, Jiali Mao, Antonio Toral and Malvina Nissim
12:00-12:20	*The Human Evaluation Datasheet: A Template for Recording Details of Human Evaluation Experiments in NLP* Anastasia Shimorina and Anya Belz
12:20 - 14:00	***Lunch***
14:00 - 15:00	***Session 3***
14:00-14:20	*Human evaluation of web-crawled parallel corpora for machine translation* Gema Ramírez-Sánchez, Marta Bañón, Jaume Zaragoza-Bernabeu and Sergio Ortiz Rojas
14:20-14:40	*Toward More Effective Human Evaluation for Machine Translation* Belén C Saldías Fuentes, George Foster, Markus Freitag and Qijun Tan
14:40-15:00	*Vacillating Human Correlation of SacreBLEU in Unprotected Languages*

	Ahrii Kim and Jinhyeon Kim
15:00 - 15:30	***Coffee Break***
15:30 - 16:30	***Invited talk by Samira Shaikh***
16:30 - 17:00	***General discussion and wrap-up***

W16 - Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures

Organizers:
Eneko Agirre, Marianna Apidianaki, Ivan Vulić

https://sites.google.com/view/deelio-ws/
Venue: Wicklow Hall 1
Friday, May 27, 2022

DeeLIO 2022 will be the Third Workshop on Knowledge Extraction and Integration for Deep Learning Architectures. DeeLIO aims to bring together the knowledge interpretation, extraction and integration lines of research in deep learning, and cover the area in between. DeeLIO will follow ACL 2022 and will be a hybrid or online event, on May 27.

09:20 - 09:30	***Opening Remarks***
09:30 - 10:30	***Invited Talk 1: Tal Linzen***
10:30 - 11:00	***Coffee Break***
11:00 - 12:30	***On-Site Poster Session***
	Cross-lingual Semantic Role Labelling with the ValPaL Database Knowledge Chinmay Choudhary and Colm O'Riordan
	How Do Transformer-Architecture Models Address Polysemy of Korean Adverbial Postpositions? Seongmin Mun and Guillaume Desagulier
	Trans-KBLSTM: An External Knowledge Enhanced Transformer BiLSTM Model for Tabular Reasoning Yerram Varun, Aayush Sharma and Vivek Gupta
	On Masked Language Models for Contextual Link Prediction Angus Brayne, Maciej Wiatrak and Dane Corneil
12:30 - 14:00	***Lunch***
14:00 - 15:00	***Virtual Poster Session***
	Cross-lingual Semantic Role Labelling with the ValPaL Database Knowledge Chinmay Choudhary and Colm O'Riordan
	How Do Transformer-Architecture Models Address Polysemy of Korean Adverbial Postpositions? Seongmin Mun and Guillaume Desagulier
	Query Generation with External Knowledge for Dense Retrieval Sukmin Cho, Soyeong Jeong, Wonsuk Yang and Jong C. Park
	Uncovering Values: Detecting Latent Moral Content from Natural Language with Explainable and Non-Trained Methods Luigi Asprino, Luana Bulla, Stefano De Giorgis, Aldo Gangemi, Ludovica Marinucci and Misael Mongiovi
	Jointly Identifying and Fixing Inconsistent Readings from Information Extraction Systems Ankur Padia, Francis Ferraro and Tim Finin
	KIQA: Knowledge-Infused Question Answering Model for Financial Table-Text Data

Rungsiman Nararatwong, Natthawut Kertkeidkachorn and Ryutaro Ichise

Trans-KBLSTM: An External Knowledge Enhanced Transformer BiLSTM Model for Tabular Reasoning
Yerram Varun, Aayush Sharma and Vivek Gupta

Fast Few-shot Debugging for NLU Test Suites
Christopher Malon, Kai Li and Erik Kruus

On Masked Language Models for Contextual Link Prediction
Angus Brayne, Maciej Wiatrak and Dane Corneil

What Makes Good In-Context Examples for GPT-3?
Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin and Weizhu Chen

15:00 - 15:30 ***Coffee Break***

15:30 - 16:30 ***Invited Talk 2: Yejin Choi***

16:30 - 16:40 ***Short Break***

16:40 - 17:40 ***Invited Talk 3: Allyson Ettinger***

17:40 - 17:50 ***Closing Remarks***

W17 - Workshop on Challenges & Perspectives in Creating Large Language Models

Organizers:
Angela Fan, Matthias Gallé, Suzana Ilić, Thomas Wolf

https://bigscience.huggingface.co/acl-2022
Venue: Wicklow Hall 2a
Friday, May 27, 2022

Two years after the appearance of GPT-3, large language models seem to have taken over NLP. Their capabilities, limitations, societal impact and the potential new applications they unlocked have been discussed and debated at length. A handful of replication studies have been published since then, confirming some of the initial findings and discovering new limitations. This workshop aims to gather researchers and practitioners involved in the creation of these models.

11:00 - 12:30	***Poster Session***
14:00 - 15:00	***BigScience***
15:00 - 15:20	***Data Governance***
15:20 - 15:40	***Data***
15:40 - 16:00	***Modeling***
16:00 - 16:20	***Prompt Engineering***
16:20 - 16:40	***Evaluation***

W18 - Speech and Language Processing for Assistive Technologies (SLPAT 2022)

Organizers:
Emily Prud'hommeaux, Sarah Ebling, Preethi Vaidyanathan

http://www.slpat.org/slpat2022/
Venue: Wicklow Meeting Room 5
Friday, May 27, 2022

This workshop will bring together researchers from areas such as natural language processing, speech signal processing, (special) education, rehabilitation sciences, computer science, HCI, communication, psychology, psycholinguistics, computer vision, and computer graphics with domain experts like clinicians, therapists, caretakers, and interpreters, as well as users to share their findings, to discuss present and future challenges, and to explore possibilities for collaboration.

09:00 - 09:30	***Opening Remarks***
09:30 - 10:30	***Keynote 1 (Annalu Waller)***
10:30 - 11:00	***Break***
11:00 - 12:30	***Session 1***
11:00-11:30	*Applying the Stereotype Content Model to assess disability bias in popular pre-trained NLP models underlying AI-based assistive technologies* Brienna Herold, James Waller and Raja Kushalnagar
11:30-12:00	*A comparison study on patient-psychologist voice diarization* Rachid Riad, Hadrien Titeux, Laurie Lemoine, Justine Montillot, Agnes Sliwinski, Jennifer Bagnou, Xuan Cao, Anne-Catherine Bachoud-Levi and Emmanuel Dupoux
12:00-12:30	*ColorCode: A Bayesian Approach to Augmentative and Alternative Communication with Two Buttons* Matthew Daly
12:30 - 14:00	***Break***
14:00 - 14:15	***Poster pitches***
14:15 - 15:15	***Poster session***
15:15 - 16:00	***Break***
16:00 - 16:30	***Session 2***
16:00-16:30	*On the Ethical Considerations of Text Simplification* Sian Gooding
16:30 - 17:30	***Keynote 2 (Raja Kushalnagar)***
17:30 - 18:00	***Closing Remarks***

W19 - The Second Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations (CONSTRAINT)

Organizers:
Tanmoy Chakraborty, Md. Shad Akhtar, Kai Shu, H. Russell Bernard, Maria Liakata, Preslav Nakov

https://lcs2.iiitd.edu.in/CONSTRAINT-2022/
Venue: Wicklow Meeting Room 4
Friday, May 27, 2022

The increasing accessibility of the Internet has dramatically changed the way we consume information. The ease of social media usage not only encourages individuals to freely express their opinion (freedom of speech) but also provides content polluters with ecosystems to spread hostile posts (hate speech, fake news, cyberbullying, propaganda, etc.). Such hostile activities are expected to increase manifold during emergencies such as the presidential election and COVID-19 pandemic spreading. Most of such hostile posts are written in regional languages, and therefore can easily evade online surveillance engines that are majority trained on the posts written in resource-rich languages such as English and Chinese. Therefore, regions such as Asia, Africa, South America, where low-resource regional languages are used for day-to-day communication, suffer due to the lack of tools, benchmark datasets and learning techniques. Other developing countries such as Italy, Spain, where the used languages (pseudo-low-resource) are not as equipped with sophisticated computational resources as English, might also be facing the same issues. Following the success of the first edition of CONSTRAINT (collocated with AAAI-21), the second edition will encourage researchers from interdisciplinary domains working on multilingual social media analytics to think beyond the conventional way of combating online hostile posts.

09:00 - 09:10	***Opening Remarks***
09:10 - 10:10	***Keynote 1: Isabelle Augenstein | Automatically Detecting Scientific Misinformation***
10:10 - 10:30	***Regular Paper Session - I***
10:10-10:30	*M-BAD: A Multilabel Dataset for Detecting Aggressive Texts and Their Targets* Omar Sharif, Eftekhar Hossain and Mohammed Moshiul Hoque
10:30 - 11:00	***Coffee break***
11:00 - 12:00	***Keynote 2: Andreas Vlachos | Fact-Checking Using Structured and Unstructured Information***
13:00 - 12:00	***Regular Paper Session - II***
12:00-12:20	*How does fake news use a thumbnail? CLIP-based Multimodal Detection on the Unrepresentative News Image* Hyewon Choi, Yejun Yoon, Seunghyun Yoon and Kunwoo Park
12:20-12:40	*Detecting False Claims in Low-Resource Regions: A Case Study of Caribbean Islands* Jason Lucas, Limeng Cui, Thai Le and Dongwon Lee
12:40-13:00	*Document Retrieval and Claim Verification to Mitigate COVID-19 Misinformation* Megha Sundriyal, Ganeshan Malhotra, Md Shad Akhtar, Shubhashis Sengupta, Andrew Fano and Tanmoy Chakraborty
13:00 - 14:00	***Lunch Break***

14:00 - 15:00	***Keynote 3: Smaranda Muresan	The Role of Text Generation in Fighting Hostile Posts***
15:00 - 15:30	***Coffee Break***
15:30 - 17:15	***Shared Task Session***
15:30-15:50	*Findings of the CONSTRAINT 2022 Shared Task on Detecting the Hero, the Villain, and the Victim in Memes* Shivam Sharma, Tharun Suresh, Atharva Kulkarni, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar and Tanmoy Chakraborty
15:50-16:00	*DD-TIG at Constraint@ACL2022: Multimodal Understanding and Reasoning for Role Labeling of Entities in Hateful Memes* Ziming Zhou, Han Zhao, Jingjing Dong, Jun Gao and Xiaolong Liu
16:00-16:10	*Are you a hero or a villain? A semantic role labelling approach for detecting harmful memes.* Shaik Fharook, Syed Sufyan Ahmed, Gurram Rithika, Sumith Sai Budde, Sunil Saumya and Shankar Biradar
16:10-16:20	*Logically at the Constraint 2022: Multimodal role labelling* Ludovic Kun, Jayesh Bankoti and David Kiskovski
16:20-16:30	*Combining Language Models and Linguistic Information to Label Entities in Memes* Pranaydeep Singh, Aaron Maladry and Els Lefever
16:30-16:40	*Detecting the Role of an Entity in Harmful Memes: Techniques and their Limitations* Rabindra Nath Nandi, Firoj Alam and Preslav Nakov
16:40-16:50	*Fine-tuning and Sampling Strategies for Multimodal Role Labeling of Entities under Class Imbalance* Syrielle Montariol, Étienne Simon, Arij Riabi and Djamé Seddah
16:50 - 17:15	***Closing***

W20 - Semiparametric Methods in NLP: Decoupling Logic from Knowledge

Organizers:
Dung Thai, Manzil Zaheer, Patrick Lewis, Rajarshi Das, Sewon Min

http://www.semiparametric.ml/
Venue: Liffey Meeting Room 2
Friday, May 27, 2022

Large parametric language models have achieved dramatic empirical success across many applications. However, these models lack several desirable properties such as explainability (providing provenance), privacy (ability to remove knowledge from the model), robust controllability, and debuggability. On the other hand, nonparametric models provide many of these features by design such as provenance, ability to incorporate/remove information. However, these models often suffer from weaker empirical performance as compared to deep parametric models. Recently, many works have independently proposed a middle ground that combines a parametric model (that encodes logic) with a nonparametric model (that retrieves knowledge) in various areas from question answering over natural languages to complex reasoning over knowledge bases to even protein structure predictions. Given the increasingly promising results on various tasks of such semiparametric model, we believe this area is ripe for targeted investigation on understanding efficiency, generalization, limitations, widening its applicability, etc. As a result, we want to host a workshop on this topic.

09:00 - 18:00 ***Workshop Schedule will be announced on the website***

W21 - Workshop on Commonsense Representation and Reasoning

Organizers:
Antoine Bosselut, Xiang Lorraine Li, Bill Yuchen Lin, Vered Shwartz, Bodhisattwa Prasad Majumder, Yash Kumar Lal, Rachel Rudinger, Xiang Ren, Niket Tandon, Vilém Zouhar

https://csrr-workshop.github.io/
Venue: The Liffey A
Friday, May 27, 2022

We organize this workshop to encourage discussion of current progress on building machines with commonsense knowledge and reasoning abilities. We aim to bring together researchers from different areas (e.g., NLP, computer vision, computational neuroscience, psychology) to communicate promising working directions in the area of commonsense reasoning.

09:00 - 09:10	***Opening Remarks***
09:10 - 09:55	***Invited Speaker 1 - Mor Geva***
09:55 - 10:25	***Lightning Talks 1***
10:25 - 10:30	***Best Paper Talk - Cloze Evaluation for Deeper Understanding of Commonsense Stories in Indonesian***
10:30 - 10:45	***Break***
10:45 - 11:30	***Invited Speaker 2 - Marie-Francine Moens***
11:30 - 12:00	***Lightning Talks 2***
12:00 - 13:30	***Lunch Break***
13:30 - 14:15	***Invited Speaker 3 - Evelina Fedorenko***
14:15 - 15:00	***Invited Speaker 4 - Greg Durrett***
15:00 - 15:30	***Break***
15:30 - 16:15	***Invited Speaker 5 - Prithviraj Ammanabrolu***
16:15 - 17:00	***Invited Speaker 6 - Tobias Gerstenberg***
17:00 - 17:10	***Closing Remarks***

W22 - Workshop on Federated Learning for Natural Language Processing (FL4NLP 2022)

Organizers:
Bill Yuchen Lin, Chaoyang He, Ninareh Mehrabi, Tian Li, Chulin Xie,
Fatemehsadat Mireshghallah, Mahdi Soltanolkotabi, Xiang Ren

https://fl4nlp.github.io/
Venue: Liffey Meeting Room 1
Friday, May 27, 2022

Due to increasing concerns and regulations about data privacy (e.g., General Data Protection Regulation), coupled with the growing computational power of edge devices, emerging data from realistic users have become much more fragmented, forming distributed private datasets across different clients (i.e., organizations or personal devices). Respecting users' privacy and restricted by these regulations, we have to assume that users' data in a client are not allowed to transfer to a centralized server or other clients. For example, a hospital does not want to share its private data (e.g., conversations, questions asked on its website/app) with other hospitals. This is despite the fact that models trained by a centralized dataset (i.e., combining data from all clients) usually enjoy better performance on downstream tasks (e.g., dialogue, question answering). Therefore, it is of vital importance to study NLP problems in such a scenario, where data are distributed across different isolated organizations or remote devices and cannot be shared for privacy concerns.

09:00 - 09:10	***Opening Remarks***
09:10 - 10:05	***Invited Speaker 1***
10:05 - 11:00	***Invited Speaker 2***
11:05 - 12:00	***Invited Speaker 3***
12:00 - 12:30	***Lunch Break***
12:35 - 12:50	***Paper Presentation 1***
12:50 - 13:05	***Paper Presentation 2***
13:05 - 13:20	***Paper Presentation 3***
13:20 - 13:35	***Paper Presentation 4***
13:35 - 14:30	***Invited Speaker 4***
14:30 - 15:25	***Invited Speaker 5***
15:25 - 16:20	***Invited Speaker 6***
16:20 - 16:35	***Paper Presentation 5***
16:35 - 16:50	***Paper Presentation 6***
16:50 - 17:05	***Paper Presentation 7***
17:05 - 17:20	***Paper Presentation 8***
17:20 - 17:35	***Paper Presentation 9***
17:35 - 18:30	***Panel Discussion (TBA)***
18:40 - 18:30	***Closing Remarks***

W23 - The 4th Workshop on NLP for Conversational AI

Organizers:
Bing Liu, Alexandros Papangelis, Stefan Ultes, Abhinav Rastogi, Yun-Nung (Vivian) Chen, Georgios Spithourakis, Elnaz Nouri, Weiyan Shi

https://sites.google.com/view/4thnlp4convai/
Venue: The Liffey B
Friday, May 27, 2022

Over the past decades, mathematicians, linguists, and computer scientists have dedicated their efforts towards empowering human-machine communication in natural language. While in recent years the emergence of virtual personal assistants such as Siri, Alexa, Google Assistant, and Cortana has pushed the field forward, the development of such conversational agents remains difficult with numerous unanswered questions and challenges. Following the success of the 3rd NLP for Conversational AI workshop at EMNLP, "The 4th NLP4ConvAI" will be a one-day workshop, co-located with ACL 2022 in Dublin. The goal of this workshop is to bring together researchers and practitioners to discuss impactful research problems in this area, share findings from real-world applications, and generate ideas for future research directions.

09:00 - 18:00 ***Workshop Schedule will be announced on the website***

W24 - The 2nd Workshop on Deriving Insights from User-Generated Text

Organizers:
Estevam Hruschka, Tom Mitchell, Marko Grobelnik, Dunja Mladenic, Nikita Bhutani

https://megagon.ai/2nd-workshop-on-deriving-insights-from-user-generated-text-wit/

Venue: Wicklow Meeting Room 1

Friday, May 27, 2022

Recent progress in natural language processing, machine learning, knowledge bases and database management have demonstrated promising results and far-reaching uses of text. However, there is tremendous untapped potential in exploring and exploiting advanced AI/ML/NLP techniques on user-generated text, which is rich in user insights and experiences. The goal of this workshop is to bring together researchers and practitioners in this area, to clarify impactful research problems, share findings from adaptation of existing approaches to user-generated data, and generate new ideas for future research. We seek papers that address challenges in harnessing user-generated data.

09:00 - 18:00 ***Workshop Schedule will be announced on the website***

W25 - The 6th Workshop on Structured Prediction for NLP

Organizers:
Andreas Vlachos, Priyanka Agrawal, André Martins, Gerasimos Lampouras, Chunchuan Lyu

https://structuredprediction.github.io/SPNLP22
Venue: Ecocem Room
Friday, May 27, 2022

From the NLP perspective, syntax and semantics of natural language are clearly structured and advances in this area will enable researchers to understand the linguistic structure of data. From the ML perspective, the large amount of available text and graph/relational data and complex linguistic structures bring challenges to the learning community. Designing expressive yet tractable models and studying efficient learning and inference algorithms become important issues. This workshop follows the four previous successful editions in 2020, 2019, 2017 and 2016 on Structured Prediction for NLP, as well as the closely related ICML 17 Workshop on Deep Structured Prediction. It is very timely, as there has been a renewed interest in structured prediction among NLP researchers due to recent advances in methods using continuous representations, able to learn with task-level supervision, or modeling latent linguistic structure.

09:00 - 09:10	***Opening Remarks***
09:10 - 09:50	***Invited Talk 1 - "Can we learn more explicit relationships between languages in multilingual machine translation?" - Angela Fan***
09:50 - 10:30	***Invited Talk 2 - "Decoding is deciding under uncertainty — the case of NMT" - Wilker Aziz***
10:30 - 11:00	***Coffee break***
11:00 - 11:45	***Contributed talks***
11:00-11:15	*Neural String Edit Distance* Jindřich Libovický and Alexander Fraser
11:15-11:30	*A Joint Learning Approach for Semi-supervised Neural Topic Modeling* Jeffrey Chiu, Rajat Mittal, Neehal Tumma, abhisheksharma@g.harvard.edu abhisheksharma@g.harvard.edu and Finale Doshi-Velez
11:30-11:45	*Predicting Attention Sparsity in Transformers* Marcos Vinicius Treviso, António Góis, Patrick Fernandes, Erick Rocha Fonseca and Andre Martins
11:45 - 12:30	***Online poster session***
12:30 - 14:00	***Lunch break***
14:00 - 14:45	***Invited Talk 3 - "Autoregressive Retrieval" - Sebastian Riedel***
14:45 - 15:30	***In-person poster session***
15:00 - 15:30	***Coffee break***
15:30 - 16:15	***Invited Talk 4 - "Do we still need inductive biases after Transformer language models?" - Siva Reddy***
16:15 - 17:00	***Invited Talk 5 - "Efficiently Modeling Long Sequences with Structured State Spaces" - Albert Gu***
17:00 - 17:10	***Closing remarks***

W26 - Workshop on Multilingual Multimodal Learning

Organizers:
Emanuele Bugliarello, Kai-Wei Chang, Desmond Elliott, Spandana Gella, Aishwarya Kamath, Liunian Harold Li, Fangyu Liu, Jonas Pfeiffer, Edoardo M. Ponti, Krishna Srinivasan, Ivan Vulić, Yinfei Yang, Da Yin

https://mml-workshop.github.io/

Venue: Liffey Hall 2

Friday, May 27, 2022

Multilingual multimodal research focuses on collecting resources, developing models, and evaluating systems that need to jointly reason over multilingual text and multimodal inputs, including images, videos, texts, and knowledge bases. Multilingual multimodal NLP presents new and unique challenges. First, it is one of the areas that suffer the most from language imbalance issues. Texts in most multimodal datasets are usually only available in high-resource languages. Second, multilingual multimodal research provides opportunities to investigate culture-related phenomena. On top of the language imbalance issue in text-based corpora and models, the data of additional modalities (e.g. images or videos) are mostly collected from North American and Western European sources (and their worldviews). As a result, multimodal models do not capture our world's multicultural diversity and do not generalise to out-of-distribution data from minority cultures. The interplay of the two issues leads to extremely poor performance of multilingual multimodal systems in real-life scenarios. This workshop encourages and promotes research efforts towards more inclusive multimodal technologies and tools to assess them.

09:20 - 09:30	***Opening Remarks***
09:30 - 10:30	***Invited Talk 1: David Ifeoluwa Adelaini***
10:30 - 11:00	***Coffee Break***
11:00 - 12:00	***Invited Talk 2: Lei Ji***
12:00 - 12:30	***Findings from the MaRVL Shared Task***
12:30 - 14:00	***Lunch***
14:00 - 15:00	***Invited Talk 3: Lisa Anne Hendricks***
15:00 - 15:45	***Workshop Papers: Archival and Non-Archival***
15:45 - 16:00	***Short Break***
16:00 - 17:00	***Invited Talk 4: Preethi Jyothi***
17:00 - 17:10	***Concluding Remarks***

W27 - 3rd International Workshop on Computational Approaches to Historical Language Change (LChange'22)

Organizers:
Nina Tahmasebi, Lars Borin, Simon Hengchen, Syrielle Montariol, Haim Dubossarsky, Andrey Kutuzov

https://languagechange.org/events/2022-acl-lchange/
Venue: Wicklow Meeting Room 2
Thursday, May 26, 2022 - Friday, May 27, 2022

This workshop explores state-of-the-art computational methodologies, theories and digital text resources on exploring the time-varying nature of human language. The aim of this workshop is three-fold. First, we want to provide pioneering researchers who work on computational methods, evaluation, and large-scale modelling of language change an outlet for disseminating cutting-edge research on topics concerning language change. We want to utilize this proposed workshop as a platform for sharing state-of-the-art research progress in this fundamental domain of natural language research. Second, in doing so we want to bring together domain experts across disciplines. by connecting researchers in historical linguistics with those that develop and test computational methods for detecting semantic change and laws of semantic change; and those that need knowledge (of the occurrence and shape) of language change, for example, in digital humanities and computational social sciences where text mining is applied to diachronic corpora subject to e.g., lexical semantic change. Third, the detection and modelling of language change using diachronic text and text mining raise fundamental theoretical and methodological challenges for future research.

09:00 - 09:15	***Introduction***
09:15 - 10:35	***Session 1*** - Chair: Mario Giulianelli
09:15-09:40	*Low Saxon dialect distances at the orthographic and syntactic level* Janine Siewert, Yves Scherrer and Martijn Wieling
09:40-10:05	*A New Framework for Fast Automated Phonological Reconstruction Using Trimmed Alignments and Sound Correspondence Patterns* Johann-Mattis List, Robert Forkel and Nathan Hill
10:05-10:35	*What is Done is Done: an Incremental Approach to Semantic Shift Detection* Francesco Periti, Alfio Ferrara, Stefano Montanelli and Martin Ruskov
10:35 - 11:05	***BREAK***
11:05 - 12:30	***Session 2*** - Chair: Andrey Kutuzov
11:05-11:30	*Lexicon of Changes: Towards the Evaluation of Diachronic Semantic Shift in Chinese* Jing Chen, Emmanuele Chersoni and Chu-ren Huang
11:30-12:00	*Deconstructing destruction: A Cognitive Linguistics perspective on a computational analysis of diachronic change* Karlien Franco, Mariana Montes and Kris Heylen
12:00-12:30	*Using Cross-Lingual Part of Speech Tagging for Partially Reconstructing the Classic Language Family Tree Model* Anat Samohi, Daniel Weisberg Mitelman and Kfir Bar
12:30 - 14:00	***LUNCH / BREAK***
14:00 - 15:00	***Keynote 1 - Dirk Geeraerts*** - Chair: Nina Tahmasebi

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

Do Not Fire the Linguist: Grammatical Profiles Help Language Models Detect Semantic Change
Mario Giulianelli, Andrey Kutuzov and Lidia Pivovarova

15:30 - 17:00 ***Poster session + COFFEE***

From qualifiers to quantifiers: semantic shift at the paradigm level
Quentin Feltgen

Explainable Publication Year Prediction of Eighteenth Century Texts with the BERT Model
Iiro Rastas, Yann Ciarán Ryan, Iiro Tiihonen, Mohammadreza Qaraei, Liina Repo, Rohit Babbar, Eetu Mäkelä, Mikko Tolonen and Filip Ginter

Caveats of Measuring Semantic Change of Cognates and Borrowings using Multilingual Word Embeddings
Clémentine Fourrier and Syrielle Montariol

"Vaderland", "Volk" and "Natie": Semantic Change Related to Nationalism in Dutch Literature Between 1700 and 1880 Captured with Dynamic Bernoulli Word Embeddings
Marije Timmermans, Eva Vanmassenhove and Dimitar Shterionov

Language Acquisition, Neutral Change, and Diachronic Trends in Noun Classifiers
Aniket Kali and Jordan Kodner

A Multilingual Benchmark to Capture Olfactory Situations over Time
Stefano Menini, Teresa Paccosi, Sara Tonelli, Marieke Van Erp, Inger Leemans, Pasquale Lisena, Raphael Troncy, William Tullett, Ali Hürriyetoğlu, Ger Dijkstra, Femke Gordijn, Elias Jürgens, Josephine Koopman, Aron Ouwerkerk, Sanne Steen, Inna Novalija, Janez Brank, Dunja Mladenic and Anja Zidar

Using neural topic models to track context shifts of words: a case study of COVID-related terms before and after the lockdown in April 2020
Olga Kellert and Md Mahmud Uz Zaman

Roadblocks in Gender Bias Measurement for Diachronic Corpora
Saied Alshahrani, Esma Wali, Abdullah R Alshamsan, Yan Chen and Jeanna Matthews

[LSCDiscovery shared task] *UAlberta at LSCDiscovery: Lexical Semantic Change Detection via Word Sense Disambiguation*
Daniela Teodorescu, Spencer von der Ohe and Grzegorz Kondrak

09:30 - 09:45 ***Introduction day 2***

09:40 - 10:40 ***Keynote 2 - Dominik Schlechtweg*** - Chair: Syrielle Montariol

10:40 - 11:05 ***BREAK***

11:05 - 11:25 ***Task description paper***

[LSCDiscovery shared task] *LSCDiscovery: A shared task on semantic change discovery and detection in Spanish*
Frank D. Zamora-Reina, Felipe Bravo-Marquez and Dominik Schlechtweg

11:25 - 12:45 ***Best task paper 1***

[LSCDiscovery shared task] *DeepMistake at LSCDiscovery: Can a Multilingual Word-in-Context Model Replace Human Annotators?*
Daniil Homskiy and Nikolay Arefyev

11:45 - 12:05 ***Best task paper 2***

[LSCDiscovery shared task] *GlossReader at LSCDiscovery: Train to Select a Proper Gloss in English – Discover Lexical Semantic Change in Spanish*
Maxim Rachinskiy and Nikolay Arefyev

12:05 - 13:30 ***LUNCH / BREAK***

13:30 - 15:00 ***Virtual poster session + COFFEE*** - Chair: TBD

[LSCDiscovery shared task] *BOS at LSCDiscovery: Lexical Substitution for Interpretable Lexical Semantic Change Detection*

Artem Kudisov and Nikolay Arefyev

[LSCDiscovery shared task] *UAlberta at LSCDiscovery: Lexical Semantic Change Detection via Word Sense Disambiguation*
Daniela Teodorescu, Spencer von der Ohe and Grzegorz Kondrak

[LSCDiscovery shared task] *CoToHiLi at LSCDiscovery: the Role of Linguistic Features in Predicting Semantic Change*
Ana Sabina Uban, Alina Maria Cristea, Anca Daniela Dinu, Liviu P Dinu, Simona Georgescu and Laurentiu Zoicas

[LSCDiscovery shared task] *HSE at LSCDiscovery in Spanish: Clustering and Profiling for Lexical Semantic Change Discovery*
Kseniia Kashleva, Alexander Shein, Elizaveta Tukhtina and Svetlana Vydrina

15:00 - 16:00 ***Mentoring***

16:00 - 16:30 ***Closing***

W28 - The Fifth Workshop on Computational Methods for Endangered Languages (ComputEL-5)

Organizers:
Sarah Moeller, Antonios Anastasopoulos, Antti Arppe, Aditi Chaudhary, Atticus Harrigan, Josh Holden, Jordan Lachler, Alexis Palmer, Shruti Rijhwani, Lane Schwartz,

https://computel-workshop.org/computel-5/
Venue: Liffey Hall 1
Thursday, May 26, 2022 - Friday, May 27, 2022

The ComputEL-5 workshop will focus on the use of computational methods in the study, support, and revitalization of endangered languages. The primary aim of the workshop is to continue narrowing the gap between computational linguists interested in working on methods for endangered languages, field linguists working on documenting these languages, and the language communities who are striving to maintain their languages. We take seriously the goal of reaching all relevant communities. To support this goal, ComputEL-5 aims to alternate between co-location with computational linguistics conferences and with language documentation conferences.

09:00 - 09:30	***Day-1 Welcome + Opening Remarks***
09:30 - 10:30	***Session A***
09:30-10:00	*Learning Through Transcription* Mat Bettinson and Steven Bird
10:00-10:30	*G_i2P_i Rule-based, index-preserving grapheme-to-phoneme transformations* Aidan Pine, Patrick William Littell, Eric Joanis, David Huggins-Daines, Christopher D Cox, Fineen Davis, Eddie Antonio Santos, Shankhalika Srikanth, Delasie Torkornoo and Sabrina Yu
10:30 - 11:00	***Day-1 Break***
11:00 - 12:30	***Session B***
11:00-11:30	*One Wug, Two Wug+s Transformer Inflection Models Hallucinate Affixes* Farhan Samir and Miikka Silfverberg
11:30-12:00	*Using Graph-Based Methods to Augment Online Dictionaries of Endangered Languages* Khalid Alnajjar, Mika Hämäläinen, Niko Tapio Partanen and Jack Rueter
12:00-12:30	*A Word-and-Paradigm Workflow for Fieldwork Annotation* Maria Copot, Sara Court, Noah Diewald, Stephanie Antetomaso and Micha Elsner
12:30 - 14:00	***Day-1 Lunch***
14:00 - 15:00	***Session C***
14:00-14:30	*Closing the NLP Gap Documentary Linguistics and NLP Need a Shared Software Infrastructure* Luke Gessler
14:30-15:00	*Automated speech tools for helping communities process restricted-access corpora for language revival efforts* Nay San, Martijn Bartelds, Tolulope Ogunremi, Alison Mount, Ruben Thompson, Michael Higgins, Roy Barker, Jane Helen Simpson and Dan Jurafsky
15:00 - 15:30	***Day-1 Break***
15:30 - 17:00	***Day-1 Poster***

09:00 - 10:30	***Session E***
09:00-09:30	*Developing a Part-Of-Speech tagger for te reo Māori* Aoife Finn, Peter-Lucas Jones, Keoni Mahelona, Suzanne Duncan and Gianna Leoni
09:30-10:00	*CLD² Language Documentation Meets Natural Language Processing for Revitalising Endangered Languages* Roberto Zariquiey, Arturo Oncevay and Javier Vera
10:00-10:30	*Corpus Development of Kiswahili Speech Recognition Test and Evaluation sets, Preemptively Mitigating Demographic Bias Through Collaboration with Linguists* Kathleen Siminyu, Kibibi Mohamed Amran, Abdulrahman Ndegwa Karatu, Mnata Resani, Mwimbi Makobo Junior, Rebecca Ryakitimbo and Britone Mwasaru
10:30 - 11:00	***Day-2 Break***
11:00 - 12:00	***Session F***
11:00-11:30	*Challenges and Perspectives for Innu-Aimun within Indigenous Language Technologies* Antoine Cadotte, Tan Le Ngoc, Mathieu Boivin and Fatiha Sadat
11:30-12:00	*Fine-tuning pre-trained models for Automatic Speech Recognition, experiments on a fieldwork corpus of Japhug (Trans-Himalayan family)* Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyên and Maxime Fily
12:00 - 13:30	***Day-2 Lunch***
13:30 - 15:00	***Day-2 Special Sesson A***
15:00 - 15:30	***Day-2 Break***
15:30 - 17:00	***Day-2 Special Sesson B***
15:00 - 15:15	***Day-2 Closing Remarks***

9

Local Guide

Conference Venue

The Convention Centre Dublin was developed to provide a world-class conference venue in the heart of Ireland's capital city. Located just 15 minutes from the airport in Dublin's Docklands, the finance and technology hub of the city, The CCD is ideally positioned to entice the international business tourism market.

Since The CCD opened in September 2010, they have hosted over 1,500 events. As business tourists spend money on hotels, taxis, dining out, entertainment, shopping, cultural experiences and tourist attractions, events held at The CCD help to generate significant revenue for the Irish economy.

We look forward to welcoming Association for Computational Linguistics Conference delegates to this stunning building in the wonderful city of Dublin!

About Ireland

The perfect escape? The city break you've been dreaming of? The cliff-path walk you'll remember forever? Well, Ireland has you covered. This magical island is just waiting to thrill you with its stunning windswept scenery along the Wild Atlantic Way, to capture you with its history of ancient sites such as Newgrange and to entice you with its traditional music. Wild, astounding, inspirational – there are so many words to describe the awesome majesty of the Irish Coast. But one thing's for sure: this place is pure magic. From Ballycastle in the north of Country Antrim to Allahies in the west of County Cork, this stunning country is renowned for its sylvan beauty, rich shades of green and individual character. You'll find quiet villages, rugged mountain roads, tranquil rivers and mythical tales.

Discover Ireland: www.discoverireland.ie
Discover Northern Ireland: www.discovernorthernireland.com
Wild Atlantic Way: www.wildatlanticway.com

Whether you're looking for stories of old, dreaming of fantastical castles, or just want to visit some of the most romantic settings imaginable there is something to suit every taste. From charming coastal villages to spectacular natural wonders, taking to the road on the island of Ireland reveals surprises at every turn. The most westerly island in Europe, Ireland is 450 km long and 300 km wide. With a population of 6.5million Ireland enjoys a rich diversity of ethnic groups and cultures.

About Dublin

Dublin is Ireland's capital city and was founded by the Vikings in 841. The city is steeped in history and buzzing with energy. Medieval, Georgian and modern architecture provides a backdrop to a friendly cosmopolitan city. Dublin is a thriving centre for culture and is home to a great musical and literary tradition, its native sons include Shaw, Yeats, Joyce, Wilde and Beckett The city's attractions include castles, museums, art galleries, pubs and cafes. Within half an hour of the city are mountain walks, stately homes and gardens, numerous golf courses, sandy beaches and fishing villages. A bustling city with a population of over 1.7 million and home to over 100 different nationalities all of whom contribute to the fabric of Dublin. While it has a genuine cosmopolitan feel, Dublin has still managed to retain its own distinct culture which is expressed in a love of literature, drama, traditional music and sport. The quintessential

Dublin Pub provides the focal point of Dublin's social life, illuminating the vibrant hues of Dubliners and their culture. Conversation flows freely unleashing the unique atmosphere that defines the city.

Dublin is one of the oldest cities in Europe and with ancient churches, grand buildings and fine museums, cultural riches abound. From the ancient to the avant-garde, from history, architecture, literature, art and archaeology to the performing arts Dublin has it, with the real advantage to the visitor being that everything is contained within a small area. When the conference business is over, there is a wealth of activities and culture for you to explore.

Further information is available by visiting the below websites:
Visit Dublin: www.visitdublin.com
Top 10 Attractions: www.visitdublin.com/top-10-dublin-attractions

Enjoying Ireland

Top Visitor Attractions

Trinity College

Trinity was founded in 1592 by Queen Elizabeth 1st on grounds confiscated from an Augustinian priory and is the oldest University in Ireland. The Campanile, erected in 1852, was built on what is believed to be the centre of the monastery. Built to further the education of the ruling Anglo-Irish families, restrictions were imposed to prevent Catholic from attending courses. These restrictions were not fully lifted until the 1970's. Trinity however admitted women in 1902, earlier than most British universities. Most of the main buildings off the main square were built during the Georgian period, some of which replaced older buildings. Within its walls, you will be able to admire Parliament Square and its 18th Century edifices. Trinity College has had many famous students such as Jonathan Swift and Samuel Beckett who later became a lecturer in French at the university. The Inter-denominational Church is very much worth a visit, should it be open during your visit.

Book of Kells at Trinity College

The Book of Kells dates back to the 9th century and is one of the most famous medieval manuscripts in the world. It is one of the main attractions in Dublin. Thomas Burgh built the Old Library building in the 18th century. Today it houses one of Irelands most illustrious books, the 9th century "Book of Kells". Before viewing the famous book visitors pass through an excellent exhibition based on the book of Kells and other important books written in monasteries around Ireland from the 9th century. After viewing the book of Kells visitors are invited to visit the long room built in 1745. Once the principal library of the University, it now contains over twenty thousand books and manuscripts of the Trinity's oldest volumes. Brian Boru's harp said to be the " oldest harp in Ireland" and a copy of the 1916 proclamation, one of the

most important documents relating to Irish history are also on display in the long room.

Guinness Storehouse

The Guinness Brewery in Dublin is Europe's largest stout producing brewery and home to the Guinness Storehouse. Opened in 1904 The Storehouse was an operational plant for fermenting and storing GUINNESS. Today it houses a very fine exhibition dedicated to the Guinness story. Visitors will discover what goes into the making a pint of GUINNESS - the ingredients, the brewing process, the time, the craft and the passion. Finish the tour with a complimentary pint of Guinness in the Gravity Bar with astonishing view of Dublin city. Visitors will also have the opportunity to spend some time in the Guinness souvenir shop.

St. Patrick's Cathedral

St Patrick's is the National Cathedral of Ireland and is built on the site where St Patrick preached. There was a small church on the site which was still in existence when the when the Anglo-Normans arrived. This church was replaced with a stone church in 1191 and it was further remodeled in 1225 to the same design as Salisbury Cathedral. Ireland's first university was founded at St. Patrick's in 1320 and intermittently operated for 200 years. St Patrick's is Gothic in style and it's splendid interior, is adorned with funeral monuments, such as The Boyle Family Memorial and the grave of Dean Jonathan Swift. Swift was dean here until his death in 1745. The Chancel has ornate stained-glass windows, and spectacular choir stalls, once used by the knights of St Patrick adjoin the Altar. The massive west towers, houses a large peal of bells whose ringing tones are so much part of the character of Dublin.

Kilmainham Gaol

Built in 1796, Kilmainham Jail has witnessed many of the events leading to Irish independence. It has housed many patriots taken prisoner during the many rebellions witnessed in Ireland from the United Irish Rebellion of 1796 to those prisoners taken during the Irish Civil War. While opened initially as a jail for all offences, it became intrinsically linked with Irish Nationalism. The jail has two main areas of cells, and several exercise yards, one of which was used for executing the leaders of the 1916 Easter Rising. The original wing dating from the opening of the jail is incredibly dark and oppressive. The later Victorian wing with its wide walkways and toplit main hall paved the way for new thinking in designs of jails in the 19th century. Touching in so many ways on the people and forces that shaped modern Ireland, Kilmainham Jail offers a panoramic insight into some of the most profound, disturbing and inspirational themes of modern Irish history. A must for visitors interested in Irish history.

Jameson Distillery

The Old Jameson Distillery Smithfield Village is located in the heart of Old Dublin. This old barley storehouse, once the centre of Distilling in Dublin, is today a museum where all the secrets of Irish whiskey's distillation will be revealed. A 15-minute promotional film and a visit around the museum will reveal all the secrets in the distillation of good Irish whiskey. The visit terminates in the Jameson Bar where all are invited to enjoy a glass of Irish.

Glasnevin Cemetery Tour

Glasnevin Museum is the world's first cemetery museum. Visitors can explore the history and the lives of more than 1.5 million people that are buried here Glasnevin Cemetery is Ireland's most important – it's the final resting place of many of the big names in Irish history. The cemetery hosts daily tours that will bring the stories of its residents to fascinating life!

Escorted Tours

Dublin Literary Pub Crawl

The Dublin Literary Pub Crawl is a tour of Dublin's historic pubs in the company of two actors who introduce the writers and perform scenes from their works: The Pub, the Poet, the Pint! It lasts approximately 2

hours. Famous writers featured include: Joyce, Beckett, Behan, Mary Lavin, Oscar Wilde, Eavan Boland, Paula Meehan, Seamus Heaney, Michael Longley. Four pubs are visited each night – and there is always a stop in Trinity College to talk about Oscar Wilde and some of the writers from the 1960s to the present day.

Sandeman's New Dublin Tour

This free, three-hour walking tour of the city departs Dame St every day at 11 am and 2 pm. The guides at Sandeman's are informed, energetic and lots of fun – you can tip them if you like, but it's strictly optional.

Pat Liddy's Walks

Visit Dublin with renowned Dublin historian Pat Liddy as your personal guide with this series of downloadable audio walking tours, which cover a range of themes from Georgian Dublin to the story of the 1916 Rising.

City Highlights Tour

Our City Highlights tour is your essential guide to Dublin. The charisma of our Sight Walking Dublin Tour guides and their valuable local insights make it the most memorable tour in our city! Through amusing historical anecdotes and of the moment entertaining stories, your friendly Dubliner guide will bring to life the city's finest squares and liveliest quarters while pointing out Ireland's best art Galleries, Viking urban design and heritage and where Dublin's music and nightlife comes alive. Of course we will show you Dublin's most important landmarks such as our Government buildings, Trinity College and Dublin Castle, all the time learning about a city stepped in history but with a modern friendly and vibrant outlook.

Historic Dublin Tour

This is a fascinating historical journey of Dublin through the ages. Our friendly local guides showcase their pride in their city and trace its development from its earliest Celtic roots. During our tour you will learn about Viking Dublin, medieval times in our city, the episodes that brought about our emergence as an independent nation and the famed inhabitants that shaped our past and cover a multitude of anecdotes and scandals. From bullet holes on buildings to city walls and edifices which have stood the test of time – walk with us and discover Dublin's living history which runs side by side with the modern city. Booking is required.

Dublin Sightseeing Jogging Tour

For the more active minded people the local knowledge tips from your friendly Dubliner guide will give you an enhanced and exciting experience of our city on this 6.5km jog. You'll get insider tips on the best bars and restaurants, the most interesting museums and where's good from music to comedy venues.

Outside Dublin's Surroundings

South Coast - Glendalough& Powerscourt Gardens Tour

Located in County Wicklow also known at the Garden or Ireland The Glendalough & Powerscourt Gardens Tour offers you an incredible day out with visits to two of Ireland's most beautiful destinations.The tour begins with a trip along the South coast of Dublin, passing Sandymount Strand, made famous in James Joyce's Ulysses. You'll enjoy stunning views of Dublin Bay as you continue on towards Dún Laoghaire harbour, and then turn inland towards the Wicklow Mountains. Stop one is Powerscourt Gardens which was voted No.3 in the Top 10 Gardens of the World by National Geographic in 2014. The gardens are located to the rear of Powerscourt House, a Palladian Mansion which is now home to fine Café's, exhibition spaces and design / craft shops. There are formal walks laid out throughout the gardens, which showcase an extensive varieties of trees, shrubs and flowers, as well as fine statuary and ironworks collected from across Europe, with plenty of hidden treasures along the way. The next stop is just a short distance away in spectacular Glendalough. A favourite destination for Dubliners as well as visitors, Glendalough is steeped in history and you'll visit an ancient monastic settlement and enjoy stunning views of the lakes and mountain. Your Fáilte Ireland accredited guide and driver will inform and entertain throughout the trip with

fascinating historical and cultural insights into sights you'll see.

North Coast & Malahide Castle

If you're in Ireland to see castles, allow us to introduce you to the Dublin North Coast and Castle Tour. Want to escape the bustling centre of Dublin? Then why not jump aboard this enchanting half day tour that is bursting with splendid views, amazing visits and incredible history. From the magnificent stories of Malahide Castle, one of Ireland's oldest castles, to the unparalleled scenery at Howth Harbour; our Coast and Castle tour will not disappoint. The tour includes a visit to Malahide Castle where tour guides will take you on a journey through the history of the Talbot family and the fascinating legacy they left behind. The castle itself dates back to the 12th century and is set amidst 250 acres of expansive parkland and gardens. The castle also plays host to the delectable delicacies of Avoca Café and food hall where you can enjoy a cake after browsing the fashion from Ireland's top designers at the onsite Avoca Store. Then you will be whisked away to the magnificent fishing village of Howth, known as one of the world's great seascapes. This working harbour is a hub of activity with plenty to see to do from shopping in the beautiful local boutiques to people watching from the quaint cafés and tea rooms. Whether you want to stroll along the pier, spot the seals lazing in the crystal waters or simply munch on some fish and chips while sitting on the harbour wall; Howth has it all.

Wild Atlantic Way

See Ireland as never before, while you travel along the beautiful west coast, taking in incredible scenery and amazing experiences. The Wild Atlantic Way is the world's longest defined coastal touring route. It's inspiring, renewing, relaxing and invigorating. It's yours to experience however you choose. Wild Atlantic Way is the world's longest defined coastal touring route. It's inspiring, renewing, relaxing and invigorating. It's yours to experience however you choose. www.wildatlanticway.com

Discover the Lakelands

The Lakelands is a well-loved Irish destination, and it's easy to see why. Stunning countryside around the lake shores and an abundance of picturesque towns like Killaloe and Ballina makes this area perfect for walking, cycling, horse riding and other activities. Browse the four different route sections of the Lakelands - Lough Derg, Lough Ree and Mid Shannon, Upper Lakelands and Lough Erne all have their own truly unique character. Visit Ireland's Lakelands for an unforgettable break. www.discoverireland.ie

Discover Northern Ireland

Discover Northern Ireland - Inspirational locations, stunning landscapes and the friendliest of welcomes. Learn about our unique stories from Saint Patrick to Titanic Belfast, from the Giant's Causeway to the Mountains of Mourne and the Walled City of Derry. www.discovernorthernireland.com

Useful Information

Electricity

220 / 240 volts . 3 Pin Plug.

Driving in Ireland

Traffic in Ireland drives on the left.

Insurance

The Conference Organising Committee or its agents will not be responsible for any medical expenses, loss or accidents incurred during the conference. Delegates are strongly advised to arrange their own personal insurance to cover medical and other expenses including accident or loss. Where a delegate has to cancel for medical reasons, the normal cancellation policy will apply. It is recommended that citizens from EU

countries bring with them a current European Health Insurance Card (EHIC) card.

Language
The main languages are English and Irish and most signposts in the Republic are bilingual. English is spoken by everyone while Irish is generally confined to pockets of the south-west, west and north-western coastal areas.

Money
The Euro is the currency in the Republic of Ireland. The Euro has 100 cents in the euro with coins in denominations of 1, 2, 5, 10, 20 & 50 cents and 1 and 2 euros. Euro notes come in denominations of 5, 10, 20, 50, 100, 200 and 500 euro. Foreign exchange bureaus are available in most banks, post offices, Tourist Information Offices, airports, some shops and accommodation. Bureau de Change kiosks are also situated in many towns and most cities. There are no exchange controls in Ireland. Any sums of money in any currency can be freely brought into or taken out of the country without disclosure or other formalities.

Smoking
Under current legislation, smoking is banned in all public areas and work places, including restaurants, pubs and bars. Smoking is still permitted in hotel bedrooms which are designated as smoking bedrooms by the hotel. Smoking in bedrooms in guest houses and bed and breakfast accommodation is at the discretion of the owner. There are substantial penalties in place for those found to be in breach of these regulations.

Tax
Refunds Value Added Tax (VAT) is charged at 23% on most goods. Cash back is the simplest and most widely used VAT refund service that issues cash refunds on departure for a handling fee. Ask for cash back form when you make your purchase.

Time
From November until February, Ireland operates on GMT 0 hour Greenwich Mean Time. From March to October, Ireland operates on GMT Greenwich Mean Time + 1 hour.

Shopping
Dublin has a busy city centre shopping area around Grafton Street and Henry Street. There is a huge range of products to bring home – from traditional Irish hand-made crafts to international designer labels. Shopping hours in general are from 9.00am to 6.00pm Monday to Saturday, with shops open until 8.00pm on Thursdays, and many shops open from 2.00pm – 6.00pm on Sunday. Dundrum Town Centre is a large shopping centre located in South Dublin. The LUAS Green Line serves Dundrum Town Centre from St. Stephens Green to Brides Glen. The Dundrum and Balally stops are only a few minutes-walk from the centre.

Tipping
Hotels and restaurants often add 10-15% to the bill especially for large parties. This is not mandatory in the Republic of Ireland. Tip cabs 10% and porters 60c per bag.

Weather
Ireland enjoys a temperate climate, with mild winters and relatively cool summers. The daily temperature in May in Dublin is on average 15 degrees Celsius. Dublin enjoys reasonable sunshine and rain belts reaching the east coast are frequently light and generally clear within a few hours. It is always wise when travelling to Ireland to pack rain gear or an umbrella.

Visa & Passport

Everyone entering Ireland must have a valid passport, or in the case of European Union Member States, a national identity card. Visitors are advised to check what form of ID is required by their airline carrier before travelling. Some airlines within Europe will only accept passport identification.

Visas are required by delegates travelling from some countries. We would encourage all delegates to check with their local Irish representation and also with your travel operator if your originating country requires a visa. The Department of Justice, Equality and Law Reform has primary responsibility for Ireland's immigration and visa policy. For more information on visa requirements for Ireland please see their website www.dfa.ie

Delegates who require a visa will require a letter of booking confirmation from the conference office. Once your registration has been completed and paid in full we would be delighted to issue this letter for you. Please email the conference office at acl2022@abbey.ie Please note that we cannot issue a letter until your registration and payment is completed in full. It is advisable that you have booked and paid for accommodation for the duration of your stay. This should be included with your visa application.

Note that VISA applications can take 6+ weeks for Ireland. We would encourage all delegates to ensure that you have left ample time.

Do I need a visa?

Please consult with the relevant Irish Embassy/Consulate/Visa Office or visit the official government website:
https://www.dfa.ie/travel/visas/visas-for-ireland/

Covid-19 Safety

The Convention Centre Dublin (The CCD) aims to provide a safe and healthy work environment for all staff, clients, service providers, sub-contractors and other third parties, including delegates, exhibitors and visitors.

Staff will adhere to all recommendations put forward by the Irish government regarding Covid-19 and commit to following all government advice and guidelines. The CCD team have implemented and maintained number of measures to help to protect delegates and to prevent the spread of Covid-19.

Measures in place

- A specific Covid-19 policy in place.
- Dedicated Covid-19 management team and response plan in place outlining all controls put in place to manage risk
- Air conditioning is a full fresh-air system with fresh air coming into the building and being discharged externally.
- Hand sanitiser is available at all entry points and at additional points during live events.
- Hot water and soap is available throughout the building for handwashing.
- Increased sanitisation has been put in place by the cleaning contractor. A schedule of all areas that are being sanitised is signed off daily.
- Technical equipment and microphones are sanitised before and after use.

- Site emergency preparedness plans have been reviewed to take into consideration restricted or changed routes, first aid and isolation areas.
- Procedures are in place for dealing with suspected Covid-19 cases, including isolation areas both front and back of house.

The latest information on Covid-19 measurements will be sent out just before the conference.

Staying safe

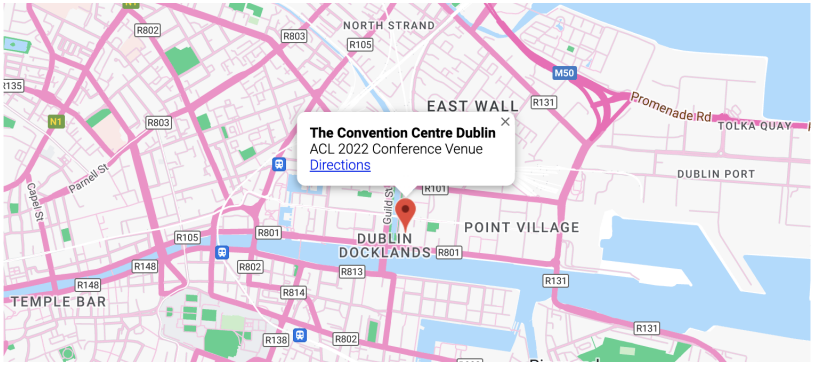
This is essential information for anyone who will be present at our conference. Learn about our increased hygiene measures and medical safeguarding. (Event health and safety precautionary guidelines are subject to change as directed by the local authorities.)

We will add more information closer to the time of the Meeting, to ensure it is up to date.

Before you leave, please ensure that you have checked with your airline to understand their policies regarding Covid-19 and have adequate travel insurance in place. You can find this information here. https://www.gov.ie/en/campaigns/c36c85-covid-19-coronavirus/.

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.

Travel to the Conference Venue



By Foot

The CCD's city centre location makes it quick and easy to reach by foot from hotels or local accommodation.

By Bike

There are parking facilities in the CCD car park, which delegates can use to secure their bikes. They are accessible by entering the East door to the building and taking the elevator to the B1 or B2 car park levels. There are over 100 dublinbike stations distributed throughout the city centre, many of which are within

close proximity to The CCD.

By Rail - DART

Visitors to the capital are encouraged to use the DART, (Dublin Area Rapid Transport), to enjoy stunning views of Dublin Bay and travel out of the city to the numerous fishing villages and towns that are dotted along the coastline. The city center DART stations are Connolly, Pearse and Tara.

By Rail - LUAS

Dublin now has its own state-of-the-art light rail system which services the city and its outskirts. This is a hassle-free way to travel to and from the city center.

By Air

Dublin Airport is located 10km north of Dublin city centre and is just 15 minutes from The CCD via the Port Tunnel. Dublin Bus operates a 24-hour service from the city centre to the airport. Dublin Express also offers a direct service to and from the airport, stopping at Custom House, which is a short walk from The CCD. Taxi and car hire services are also easily available.

The Visitor Leap Card

Leap Cards are quicker and cheaper than paying for single fares with cash and they can be used to pay-as-you-go on Airlink, Dublin Bus, LUAS and DART. You can purchase them in the arrivals hall at Dublin Airport. The Leap Card takes you anywhere you want to go in the city and suburbs for the following time periods; 24 hours, 3 days, 7 days.

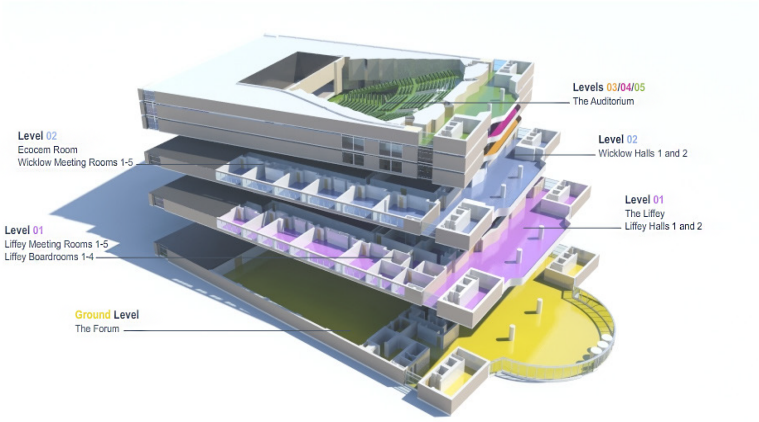
By Car

The CCD is only minutes from the Port Tunnel with connections to the upgraded M50 motorway. There is a public car park located below The CCD with 320 spaces, 8 disabled spaces and 2 e-car spaces, managed by Euro Car Parks. The opening times are 7am to midnight, 7 days a week. Discounted parking rates are available when pre-booked through Euro Car Parks. Both cash and credit card payments are accepted. CCD parking | Euro Car Parks

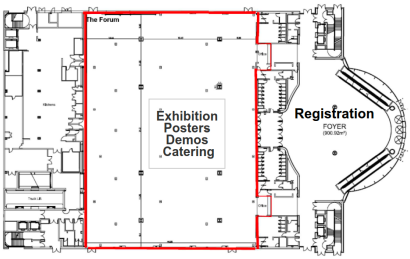
10 Venue Map

The Convention Centre Dublin Floor Plan

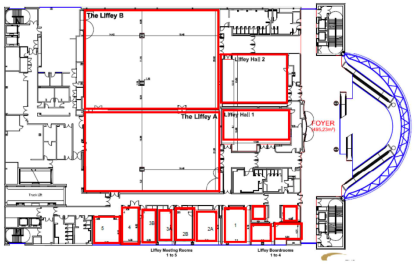
The Convention Centre Dublin offers extremely flexible and multi-purpose meeting spaces, which can be tailored to your individual requirements. This technically sophisticated venue is designed to the very latest standards. All the halls and meeting rooms are Wi-Fi enabled and have the most advanced audio-visual equipment and lighting systems. What's more, at The CCD we offer a level of service that is second to none. From our professional Sales Team to our meticulous Event Managers, our Technical Production experts to our friendly CCD Hosts, our mission is to make your event a successful and memorable one for you, your clients, delegates and guests.



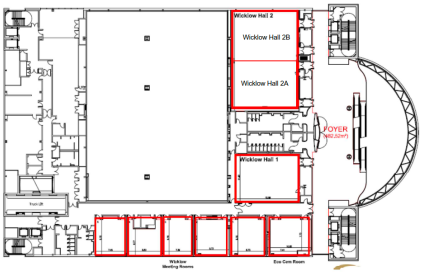
Ground Floor



Level 1

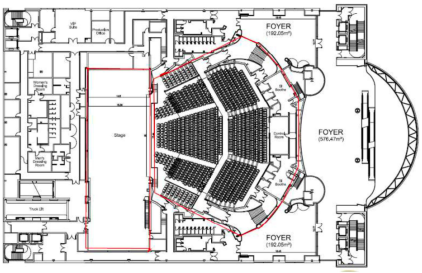


Level 2



Level 3

Auditorium



List of Exhibitors

1. Defined.ai
2. Spotify
3. Flitto
5. IBM
6. Grammarly
7. Meta
8. LivePerson
9. DeepMind
10. Google
11. Bloomberg
12. Amazon
13. Relativity
14. Bosch
15. Cohere.ai
16. G-Research
17. An unmanned Springer Book booth
18. aiXplain

Author Index

amazon | science

Amazon at ACL 2022

Learn more about research at Amazon Science

Amazon Science gives you insight into our approach to customer-obsessed scientific innovation. Our scientists continue to publish, teach, and engage with the academic community, in addition to utilizing our working backwards method to enrich the way we live and work.

To learn more visit:
https://www.amazon.science/

Academics@Amazon

Academics@Amazon are programs aimed at enabling university professors to work on large-scale and high-impact technical challenges at Amazon on a part-time basis without leaving their academic institutions.

Amazon Science Internships

We hire PhD and Masters interns throughout the year across a wide variety of teams, locations, and domains.

Visit: 2022amazonscienceinternships.splashthat.com

Diversity @ Amazon

The ongoing struggles with diversity and inclusion have profoundly influenced the lives and work of scientists and academics around the world. Read their stories, and find out what Amazon's science community is doing to help address these issues.

Job Opportunities

Want to learn more about NLP, speech, ML and other opportunities at @Amazon? Check out our career opportunities today at
http://www.amazon.jobs

Global research locations

Aachen
Atlanta
Austin
Barcelona
Bellevue
Bengaluru
Berkeley
Berlin
Boston
Cambridge
Chennai
Culver City
Cupertino
Dublin
Edinburgh
Gdansk
Graz
Haifa
Hyderabad
Irvine
London
Luxembourg
Arlington
Manhattan Beach
New York
Newark
North Reading
Palo Alto
Pasadena
Pittsburgh
San Diego
San Francisco
Santa Clara
Santa Monica
Seattle
Shanghai
Sunnyvale
Sydney
Tel Aviv
Tokyo
Toronto
Tübingen
Turin
Vancouver
Westborought

Email us directly at acl-2022@amazon.com to learn more.


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

LIVEPERSON

OUR INDUSTRY-LEADING

Conversational Cloud

Say hello to LivePerson's Conversational Cloud — creating closer connections between brands and customers — all under one roof.

AI and automation capabilities

UNDERSTAND intent from text to automate dynamic actions or routing, conversation automation, chatbots, automated processing, and more.

RESPOND to trends in the marketplace and benchmark against other brands in their vertical via out-of-the-box analytics.

CONNECT contextual data with consumer intent to engage with the right consumer at the right time, for hyper-personalized experiences.

Conversational AI experiences

The tech behind the best brands

We're evolving the tools needed to maximize the performance of machine learning technology so we can get to the future of self-learning Conversational AI chatbots.

Curiously Human™ dialogue

Focused on consumer effort and intent to develop a Curiously Human dialogue, our machine learning Meaningful Automated Conversation Score (MACS) algorithm recognizes when and where the bot fails in the conversation. This provides an additional foundation of self-learning automation to recognize when, where, and how AI chatbots fail, helping improve performance.

High-quality annotation

All of these machine learning tools require annotation, using humans to teach AI models. LivePerson's tools make annotation as easy and scalable as possible — and our annotation team provides the expertise critical to success in solving complex language problems.

Powered by insights and intents from nearly **1 billion conversational interactions every month**, LivePerson's Conversational Cloud delivers exceptional understanding, connection, and business outcome.

VISIT LIVEPERSON.COM TO LEARN MORE

Meta

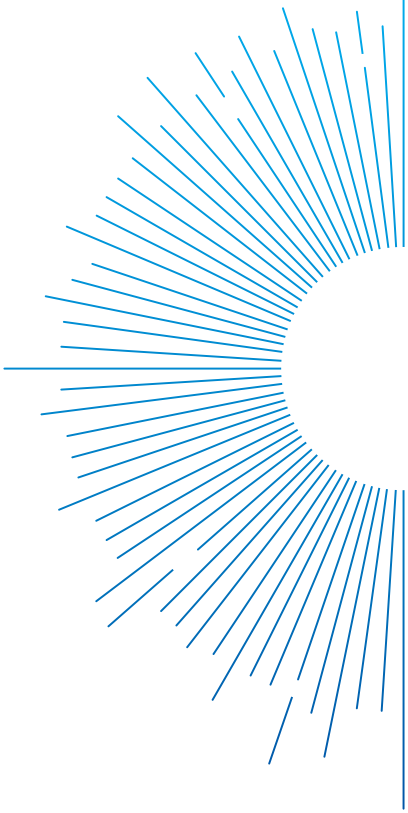
Realizing the potential of AI today and creating the experiences of tomorrow.

Help us pioneer the future of AI:
www.metacareers.com

BAIDU NLP

BAIDU NATURAL LANGUAGE PROCESSING

On a mission to enable machines to understand language and acquire intelligence so as to make the world better, Baidu NLP is dedicated to core NLP technologies, leading technology platforms and innovative products that are set to serve users across the globe and make the complex world simpler.



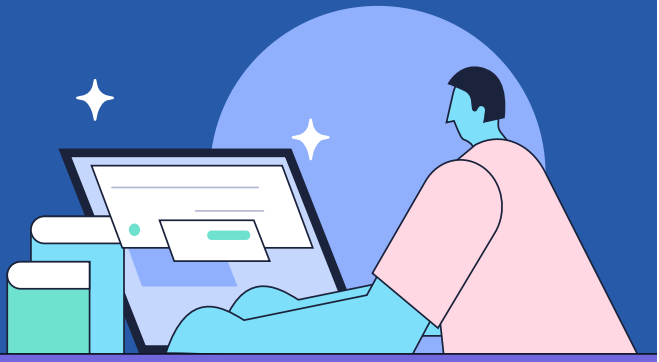
Baidu is the leading Chinese language internet search provider. Baidu aims to make the complicated world simpler through technology.

Email: nlp@baidu.com
Web: ai.baidu.com

grammarly

- Our AI-powered writing assistance scales across multiple platforms and devices, helping to empower users worldwide wherever they communicate.
- We use innovative approaches—including advanced machine learning and deep learning—to develop our writing assistance.
- Grammarly helps 30 million people and 30,000 professional teams write more clearly and effectively every day.
- We are a values-driven team of more than 700, and we're growing. Join us!

grammarly.com/jobs



What's Next

At IBM Research, we're inventing what's next in AI. This involves unleashing AI on the language of industry and the world of code, moving AI beyond simple chatbots and building trusted, bias-free AI that takes the burden out of business decision-making. We're working on general-purpose fluid intelligence, making fair and explainable AI systems, and developing new, more energy-efficient AI hardware.

IBM

IBM **Research**

Conference papers, office hours, and featured openings:
ibm.biz/ACL2022

Connect with us:
ibm.biz/connectwithus

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 and Glassdoor. Our mission is to empower people with better information to make their best decision.

The areas we focus are Natural Language Processing, Data Management, Data Integration, Machine Learning, and Human-Computer Interaction.

For more information about our lab and hiring, please visit www.megagon.ai!

Megagon Labs @ ACL 2022

Findings of ACL

Comparative Opinion Summarization via Collaborative Decoding

May 23-25, 2022
Poster Session

2nd WIT

Workshop On Deriving Insights from User-Generated Text

May 27, 2022

Microsoft Research is where leading scientists and engineers have the freedom and support to propel discovery and innovation. Here, they pursue and publish curiosity-driven research in a range of scientific and technical disciplines that can be translated into products. With access to vast computing power, global multi-disciplinary teams tackle complex problems that drive breakthrough technologies and improve lives.

Careers

Imagine having the freedom and resources to pursue and publish curiosity-driven research that tackles complex problems to improve lives. **aka.ms/msrcareers**

Events

Connect with our researchers at conferences and Microsoft Research events around the world. **aka.ms/msrevent**

Microsoft Research Blog

Read in-depth technical and notable articles from our researchers, scientists, and engineers. **aka.ms/msrblog**

Microsoft Research Podcast

Listen in on conversations that bring you closer to the cutting-edge of technology research and the scientists behind it. **aka.ms/msrpod**

Programs

Further your research with fellowships, grants, and opportunities. **aka.ms/msrprog**

Connect with us:

MicrosoftResearch
@MSFTResearch
microsoftresearch
Microsoft Research Group
@msft_research
#msftresearch

Microsoft

A community with a unique mission

We're a dedicated scientific community, committed to "solving intelligence" and ensuring our technology is used for widespread public benefit.

Our pioneering and collaborative culture is made up of people from unusually diverse backgrounds. Together, we build computer systems that learn how to solve problems and advance scientific discovery for all.

deepmind.com/careers

Language Technology Lab of Alibaba DAMO Academy

About US

The Language Technology Lab is committed to the R&D of natural language processing (NLP) technologies. As part of DAMO Academy, the Language Technology Lab provides the Alibaba economy with basic NLP technologies, dialog intelligence, applied algorithms, machine translation, and content search and recommendation. The lab has published over 200 papers in leading AI/NLP conferences or journals. These techniques power a myriad of applications that are used by thousands of business partners across the Alibaba economy, helping them translate trillions of words on a daily basis. The lab provides services to thousands of partners and external customers from key sectors such as e-commerce, justice, healthcare, and telecommunications.

Accepted Main Conference Papers

UniTE: Unified Translation Evaluation
Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek F. Wong, Lidia S. Chao

Learning to generalize to More: Continuous Semantic Augmentation for Neural Machine Translation
Xiangpeng Wei, Heng Yu, Yue Hu, Rongxiang Weng, Weihua Luo, Jun Xie, Rong Jin

Efficient Cluster-based k-Nearest-Neighbor Machine Translation
Dexin Wang, Kai Fan, Boxing Chen, Deyi Xiong

A Comprehensive and Large-Scale Dataset for Integrated Argument Mining Tasks
Liying Cheng, Lidong Bing, Ruidan He, Qian Yu, Yan Zhang, Luo Si

MELM: Data Augmentation with Masked Entity Language Modeling for Low-Resource NER
Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, Chunyan Miao

GlobalWoZ: Globalizing MultiWoZ to Develop Multilingual Task-Oriented Dialogue Systems
Bosheng Ding, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si, Chunyan Miao

WikiDiverse: A Multimodal Entity Linking Dataset with Diversified Contextual Topics and Entity Types
Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, Yanghua Xiao

Identifying Chinese Opinion Expressions with Extremely-Noisy Crowdsourcing Annotations
Xin Zhang, Guangwei Xu, Yueheng Sun, Meishan Zhang, Xiaobin Wang, Min Zhang

Code Synonyms Do Matter: Multiple Synonyms Matching Network for Automatic ICD Coding
Zheng Yuan, Chuanqi Tan, Songfang Huang

Parallel Instance Query Network for Named Entity Recognition
Yongliang Shen, Xiaobin Wang, Zeqi Tan, Guangwei Xu, Pengjun Xie, Fei Huang, Weiming Lu, Yueting Zhuang

SPONSORS

DIAMOND

amazon | science

LIVEPERSON

Meta

Google Research

Bloomberg Engineering

PLATINUM

Baidu 百度

grammarly

IBM

ByteDance

DeepMind

GTCOM

Megagon Labs

Microsoft

GOLD

co:here

Alibaba Group 阿里巴巴集团

Relativity

RESEARCH

BOSCH Invented for life

servicenow

SILVER

duolingo

ASAPP

NAVER

BRONZE

Adobe

Babelscape

Spotify