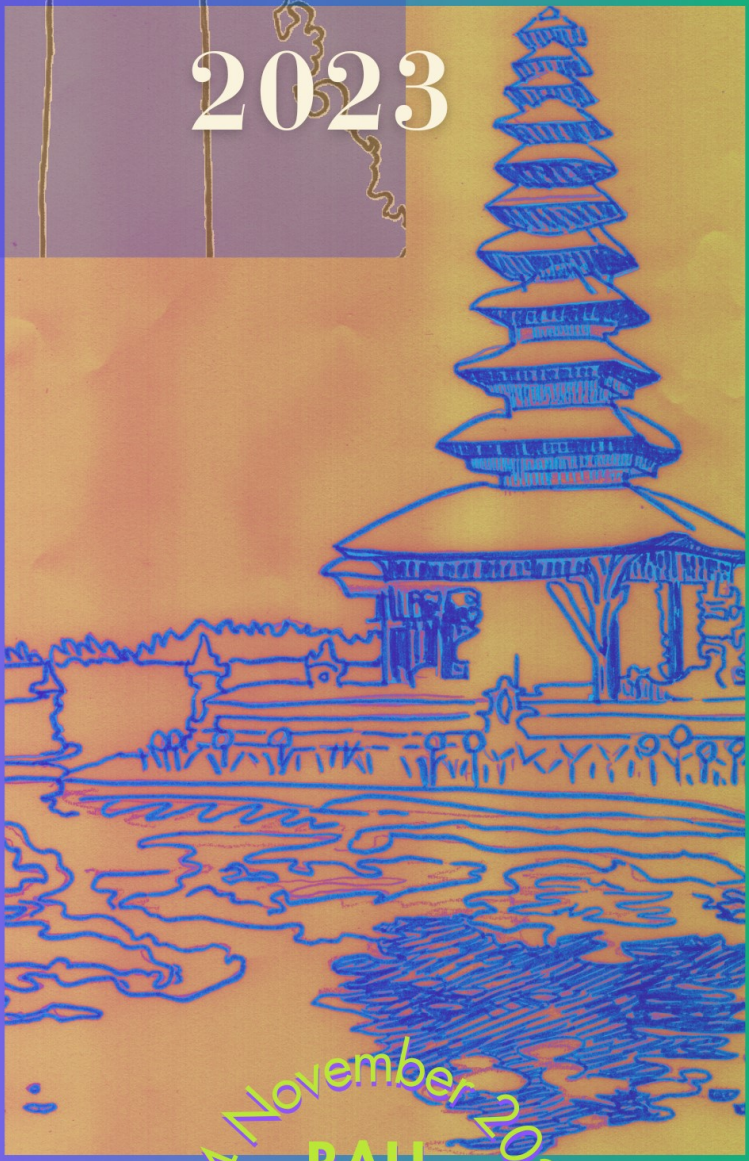
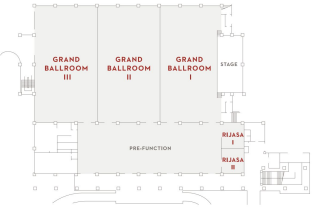
IJCNLP-AACL 2023



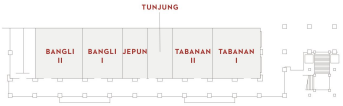
1-4 November 2023

BALI

INDONESIA

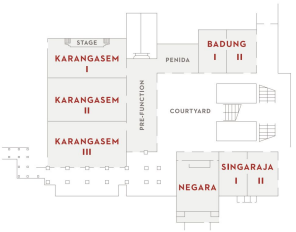


Grand Ballroom
Lower Level

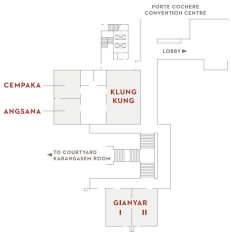


Grand Ballroom
Upper Level

Grand Ballroom
Karangasem Ballroom



Karangasem Ballroom
Lower Level



Karangasem Ballroom
Upper Level

Cover design by Sonoko Miura
Handbook assembled by Yasuhide Miura and Yunita Sari

Contents

1

Conference Information

Message from the General Chair

Welcome to IJCNLP-AACL 2023, the 13th International Joint Conference on Natural Language Processsing (IJCNLP) and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (AACL). The conference will be held in Nusa Dua, Bali, Indonesia, during November 1-4, 2023. The joint conferences of IJCNLP and AACL are organized by alternating leadership in the Asia-Pacific region, in odd years by Asian Federation of Natural Language Processing (AFNLP) and in even years by AACL, except when the annual meetings of ACL are offered in the region every three years, in which case the conferences will be offered and organized solely by ACL. This year, the conference is organized by AFNLP. As President of AFNLP, I have formed the Conference Coordination Committee (CCC) last year with executive members of AFNLP: Sadao Kurohashi (Vice President), Min Zhang (Secretary), Liang-Chih Yu (Honorary Treasurer), and Minghui Dong (Web Master). Xuanjing Huang has participated in the meetings of the committee as AACL Representative. I am truly thankful for their dedication in the initial stages as CCC members, in particular during the selection of the Local Chairs together with the conference venue and the nomination and appointment of General Chair and Program Committee Chairs, and, after it is decided that I also serve as General Chair of the conference, for their service as respective chairs throughout the conference preparation stages. Their continued support has been of tremendous help and is much appreciated. I would also like to thank deeply the entire organizing committee of the conference for their hardwork and superb handling of difficult situations, most of all due to the much limited time to address their respective tasks.

- Program Committe Chairs: Yuki Arase, Baotian Hu and Wei Lu have managed the most critical task of establishing the main program of the conference, which includes, among many others, forming and coordinating the Program Committee, constructing the call for papers with a timely theme track, overseeing the review process on both Softconf and ARR, making decisions on paper acceptance, selecting keynote speak-

ers and putting together the program schedule, and most of all communicating with the reviewers, authors, and other committee members over numerous times during the entire process. This is a formidable task, and their efforts for leading to its successful completion are so much appreciated.

- Tutorial Chairs: Yun-Nung (Vivian) Chen and Sadao Kurohashi have gone through the tutorial proposals to select six tutorials that are well balanced both topically and geographically, actively fostering diversity and inclusion.

- Workshop Chairs: Kehai Chen and Lun-Wei Ku have gone over the workshop proposals to select eight strong workshops, NLPMC, WIESP, Eval4NLP, SocialNLP, SEA, FinNLP, NLInt, and ArtOfSafety.

- Demo Chairs: Sriparna Saha and Herry Sujaini have overseen the process of selecting papers among those submitted to the system demonstration session independently of the main program.

- Publication Chairs: Minghui Dong and Kotaro Funakoshi have put together all the documents and articles to construct the conference proceedings expertly, cordially and in a timely manner.

- Sponsorship Chairs: Min Zhang, Satoshi Sekine, Haofeng Wang, and Zhongqing Wang have worked hard to solicit funding to support the conference, and successfully, despite the short time to operate. We are all much grateful to the funding organizations for their generous help.

- Finance Chair: Liang-Chih Yu has always been the person to consult whenever any financial issues arise. Things are not yet fully over at the time of preparing for this message, but are believed to come to self-sustainment, with some room for financial aid as well, thanks to the valiant efforts of all the people involved.

- Website & App Chair: Juntao Li has worked single-handedly and dependably to reflect all the requests to post important news at our conference homepage.

- Publicity & Social Media Chairs: Koustava Goswami and JinYeong Bak have worked to disseminate information over social media as widely as possible whenever there are notable events for the conference.

- Handbook Chairs: Yasuhide Miura and Yunita Sari have worked very hard to meet the deadline for the conference handbook when there is really not much time left. The help of Sonoko Miura as Conference Handbook Designer is also gratefully acknowledged.

- Faculty Advisors to Student Research Workshop (SRW): Hyeju Jang, Yugo Murawaki, and Derek Fai Wong have worked to oversee the SRW, in particular communicating with its chairs, Dongfang Li, Rahmad Mahendra and Zilu Peter Tang. The efforts of the faculty advisors and SRW chairs are much appreciated.

- Diversity & Inclusion Chair: Hwanhee Lee has worked to resolve requests for financial aid, mostly in the form of registration waiver, and as amicably as possible, under a strict limit on granting such requests.

- Registration Chair: Dessi Puji Lestari has worked to oversee the process of r egistration, together with the kind help of Minghui Dong as the manager of the registration site.

- Local Chairs: Derry Wijaya, Ayu Purwarianti and Adila Alfa Krisnadhi have worked to resolve practically numerous issues, many times beyond the duties of local chairs, after successfully bidding to host the conference in Nusa Dua, Bali, Indonesia. It is with much gratitude and respect that I commend their efforts to the fullest.

In summary, I can only say that their sense of duty and responsibility has been critical to meeting and resolving so many difficulties that we faced along the way. Many, many thanks to all of them. The mode of the conference has been determined early on to be hybrid, considering the fact that we are not fully over the coronavirus pandemic. However, it has not been quite clear until very recently whether we would need, and can afford, the top-notch professional service of an on-demand video platform such as Underline for a hybrid conference. I am very grateful to all the chairs for the remarkable patience and very happy with the concerted decision to go for the service of Underline, albeit in a limited scope. I wish to thank kindly Damira Mrsic and Sol Rosenberg, representatives of Underline, for the patience and expert help. I also thank Richard Gerber for the help with our use of Softconf. The Local Chairs made a proposal to host the conference in Nusa Dua early on, but it took all of us a long time to finally come to a decision to choose Grand Hyatt Bali, Nusa Dua, Bali, Indonesia for its exceptional location in the region as the conference venue. Its daytime temperatures are expected to reach 31°C, with the nighttime temperatures of around 24°C. I wish a very pleasant time for all the off-line participants of the conference and an equally memorable time for all the on-line participants.

Welcome to the conference and Nusa Dua!

IJCNLP-AACL 2023 General Chair
Jong C. Park
Korea Advanced Institute of Science and Technology (KAIST)

Message from the Program Chair

Welcome to IJCNLP-AACL 2023, the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics. This year, we have organized the conference in a hybrid format to accommodate both in-person and virtual participation, making it accessible to everyone regardless of travel constraints. The physical conference takes place in Bali; we are looking forward to discussions on present and future research in NLP with all of you.

This conference has been made possible through the dedicated voluntary efforts of hundreds of individuals. We sincerely thank everyone involved, in particular:

- The incredible team of 30 SACs, 72 ACs, and 427 reviewers who carefully handled submissions.
- Our Best Paper Award Committee, Danushka Bollegala, Jing Jiang, Yang Liu (Tsinghua University), Nanyun (Violet) Peng, Xiaojun Wan, Taro Watanabe, and Koichiro Yoshino, who assessed the candidate papers under the tight schedule.
- Computational Linguistics Editor-in-Chief, Hwee Tou Ng, for exploring the presentation possibility of CL papers with us.
- The ARR technical team, Jonathan Kummerfeld, Nils Dycke, Yoshitomo Matsubara, Dhruv Naik, and Amanda Stent, for supporting IJCNLP-AACL 2023 to have submissions through ARR.
- Rich Gerber at Softconf, who always addressed our inquiries and issues quickly.
- The Program Chairs of conferences going on the same period of time, including Houda Bouamor, Juan Pino, Kalika Bali (EMNLP 2023), Maria Keet, Hung-Yi Lee, Sina Zarrieß (INLG 2023), Svetlana Stoyanchev, Shafiq Joty (SIGDIAL 2023), and Jennifer Dy, Sriraam Natarajan, Kristian Kersting, Sameer Singh, Prasad Tadepalli, Angela Yao (AAAI 2024), whom we cooperated with to preclude multiple submissions.
- The ACL Executive Committee, especially Yusuke Miyao (Conference Officer), for supporting IJCNLP-AACL 2023 with various advice on conference organization.
- The ACL Ethics Committee, Min-Yen Kan, for advising us to conduct ethics reviews.
- The past $*$ACL conference program chairs, in particular Jordan Boyd-Graber, Naoaki Okazaki, Anna Rogers (ACL 2023) for sharing their experiences that were crucial to avoid pitfalls and practically sharing tools to handle a large volume of submissions.
- ACL Anthology director, Matt Post, for supporting us to publish the proceedings.
- Underline team, Jernej Masnec, Luka Simic, Sol Rosenberg, Borna Bevanda, and Damira Mrsic (Underline Science, Inc.), for their professional services and supports to realize the online part of the conference.

The conference owes its existence to the collaborative efforts of multiple chairs of IJCNLP-AACL 2023. We are deeply grateful to:

- General Chair, Jong C. Park, who was responsible for and led the whole process.
- Publication Chairs, Minghui Dong and Kotaro Funakoshi, for compiling the proceedings and ensuring the correct formatting.
- Handbook Chairs, Yunita Sari and Yasuhide Miura, for creating the conference handbooks.
- Web Chair, Juntao Li, and Publicity & Social Media chairs, JinYeong Bak and Koustava Goswami, for distributing the important information on the web, which allowed efficient and timely communication.
- Local Chairs, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, who did everything for realizing the physical part of the conference.
- Our assistants, Jindi Yu, Senbao Shi, and Jiaxi Li for their help and support throughout the conference preparation period.

Last but not least, we extend our heartfelt gratitude to all the authors for their invaluable scientific contributions, which are at the very heart of IJCNLP-AACL 2023. We hope everyone will enjoy IJCNLP-AACL 2023 either on-site or online.

Submission and Acceptance

IJCNLP-AACL 2023 used two submission platforms: direct submission through SoftConf and submissions through ACL Rolling Review (ARR). We received 338 direct submissions (229 long and 109 short) and 25 commitments from ARR (19 long and 6 short), in total 363 valid papers went through the review process. The submissions came from 33 countries and regions of all over the world: 49.2% from Asia-Pacific, 26.7% from North America, 18.7% from Europe, 4.1% from Middle East, and 1.3% from Africa.

The direct submissions were thoroughly reviewed by 427 reviewers under 72 ACs and 30 SACs in 19 tracks. The ARR committed papers were handled by the SACs based on review and meta-review comments gained through ARR. Finally, we have accepted 94 submissions (72 long and 22 short) as the main conference papers accompanied by 37 Findings papers (23 long and 14 short), which results in the acceptance rates of 25.9% for the main conference (29.0% long and 19.1% short) and 36.1% when including the Findings papers (38.3% long and 31.3% short).

In adherence to the ACL code of ethics, we have conducted ethics reviews. Given the limited number of papers that raised ethical concerns, the program chairs assumed the responsibilities of ethics reviewers. Authors of papers that received ethics review comments were requested to address these concerns in their final versions.

The papers accepted to the main conference will be presented in the form of either oral and poster presentations. The presentation modes were determined solely based on the nature of the work, irrespective of review scores. There is no distinction in the proceedings regarding

the presentation modes. The Findings papers will have optional lightning talks to introduce highlights of their work.

Design of Tracks

IJCNLP-AACL 2023 has incorporated 18 general areas reflecting the trends of the field. Most of these areas are consistent with the recent ∗ACL conferences. In addition, recognizing the conference centers on Asia-Pacific regions, which boast a variety of regional languages with limited language resources, we have introduced a special theme track this year: Large Language Models and Regional/Low-Resource Languages. The most popular areas were "NLP Applications," "Resources and Evaluation," and "Large Language Models and Regional/Low-Resource Languages."

- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Generation and Summarization
- Information Extraction
- Information Retrieval and Text Mining
- Interpretability and Analysis of Models for NLP
- Language Modeling and Analysis
- Linguistic Theories, Cognitive Modeling, and Psycholinguistics
- Machine Learning for NLP
- Machine Translation and Multilingualism
- Multimodality: Speech, Vision, Robotics, and Beyond
- NLP Applications
- Question Answering
- Resources and Evaluation
- Semantics: Lexical, Sentence-level Semantics, Textual Inference, etc.
- Sentiment Analysis, Stylistic Analysis, and Argument Mining
- Society and NLP
- Syntax: Tagging, Chunking, Parsing, etc.
- Theme Track: Large Language Models and Regional/Low-Resource Languages

Limitations Section and Responsible NLP Checklist

Following ACL 2023, EMNLP 2022, and ARR, we made it mandatory to discuss limitations of a study and submit the responsible NLP checklist to foster the open and honest scientific discussions and promote consideration of responsibility in NLP research. We also encouraged authors to include Ethics Statements whenever there might be potential concerns. The Limitations Section and Ethics Statements did not count towards the page limit.

Best Paper Awards

Following the new ACL conference awards policy, we organized the Best Paper Award Committee to select the Best Paper Awards followed by Outstanding Paper Awards, together with two special awards, namely Social Impact Award and Resource Award. The committee assessed 15 papers nominated by reviewers. In addition, SACs of each track selected at most a paper for Area Chair's Awards. The award winners will be announced in a dedicated plenary session on November 4, 2023.

Yuki Arase (Osaka University)
Baotian Hu (Harbin Institute of Technology (Shenzhen))
Wei Lu (Singapore University of Technology and Design)
IJCNLP-AACL 2023 Program Comittee Co-Chairs

Meal Info

The following meals are provided as part of your registration fee:

- **Break:** Coffee, tea, infused water and light snacks are provided late morning (approximately 10:30 AM) and midafternoon (approximately 15:30 PM) on all conference days.

- **Lunch:** Lunch will be provided each day starting at 12:00 PM alternating between Indonesian (Wednesday and Friday) and International (Thursday and Saturday) menus on the 4 days.
  Venue:
  Day 1: Watercourt
  Day 2 : Karangasem III
  Day 3 : Karangasem III
  Day 4 : Karangasem III

- **Dinner:** Buffet Dinner with Indonesian and International menus are provided on 4th November 2023 at South Beach starting at 06:00 PM onwards.

- **Welcome reception:** Welcome reception will be held on 1st November 2023 at Beach Bale starting at 06:00 PM onwards. Classic selection of beverages and assorted crackers and peanuts will be provided.

Welcome Reception/Dinner tickets are included as part of the Full Conference Registration. No admission without an entry ticket.

2 Anti-harassment policy

The following anti-harassment policy is quoted in verbatim, in deference to the policy as set by ACL for all ACL affiliated conferences, though the present conference is organized by Asian Federation of Natural Language Processing (AFNLP), a different organization. While AFNLP does not presently have a similar statement, related issues should be brought up to either the committee mentioned below or to the AFNLP executive officers as listed.

IJCNLP-AACL 2023 follows the ACL Anti-Harassment Policy. The open exchange of ideas, the freedom of thought and expression, and respectful scientific debate are central to the aims and goals of the ACL. 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 all the members, as well as participants at our events and in our programs.

Harassment and hostile behavior are unwelcome at any ACL conference, associated event, or in ACL-affiliated on-line discussions. This includes: speech or behavior that intimidates, creates discomfort, or interferes with a person's participation or opportunity for participation in a conference or an event. We aim for ACL-related activities 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, appearance, 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 policy is not intended to inhibit challenging scientific debate, but rather to promote it through ensuring that all are welcome to participate in shared spirit of scientific inquiry. Vexatious complaints and willful misuse of this procedure will render the complainant subject to the same sanctions as a violation of the anti-harassment policy.

It is the responsibility of the community as a whole to promote an inclusive and positive environment for our scholarly activities. In addition, anyone who experiences harassment or hostile behavior may contact any current member of the ACL Executive Committee, who is usually available at the registration desk during ACL conferences. Members of the executive committee will be instructed to keep any such contact in strict confidence, and those who approach the committee will be consulted before any actions are taken.

The full policy and its implementation is defined at:
https://aclweb.org/adminwiki/index.php/Anti-Harassment_Policy

The AFNLP executive officers are listed at:
http://www.afnlp.org/wp/?page_id=176

3 Keynotes and Featured Plenary Talks

Glot500: Creating and Evaluating a Language Model for 500 Languages

Hinrich Schuetze
University of Munich

Date/time: Friday, November 3, 2023, 14:00–15:00

Abstract: Most work on large language models (LLMs) has focused on what we call "vertical" scaling: making LLMs even better for a relatively small number of high-resource languages. We address "horizontal" scaling instead: extending LLMs to a large subset of the world's languages, focusing on low-resource languages. Our Glot500-m model is trained on more than 500 languages, many of which are not covered by any other language model. But how do we know that the model has actually learned these 500 languages? Broad low-

resource evaluation turns out to be a difficult problem in itself and we tried to innovate in several ways. One issue we were not able to solve is that parts of our evaluation standard cannot be distributed due to copyright restrictions. We also find that attributing good/bad performance to the so-called curse of multilinguality is naive and there is in fact also a "boon of multilinguality". We have released Glot500-m and are in the process of making our training corpus Glot500-c publicly available.

Biography: Hinrich Schuetze is Professor at the Center for Information and Language Processing at LMU Munich. His lab is engaged in research on multilinguality, representation learning and linguistic analysis of NLP models. His research has been funded by NSF, the German National Science Foundation and the European Research Council (ERC Advanced Grant), inter alia. Hinrich is coauthor of two well-known textbooks (Foundations of Statistical Natural Language Processing and Introduction to Information Retrieval), a fellow of HessianAI, ELLIS (the European Laboratory for Learning and Intelligent Systems) and ACL (Association for Computational Linguistics) and (co-)awardee of several best paper awards and the ACL 2023 25-year test of time award.

How new tasks and datasets have enabled progress in NLP

Julia Hockenmaier
University of Illinois at Urbana-Champaign

Date/time: Thursday, November 2, 2023, 9:00–10:00

Abstract: Much of the current progress in our field seems to be driven by a small number of industry labs that have access to computing resources, data, and financial backing on a scale that would have seemed unimaginable only a few years ago (and that remain unattainable for the majority of researchers in academia or industry). However, many of the datasets that have defined novel tasks and that have catalyzed research on new problems came out of much more modest efforts by academic researchers. I will focus on the work that my own group has done on image description and on grounded instruction giving/following in Minecraft as examples of how this kind of work has helped drive scientific progress.

Biography: Julia Hockenmaier is a professor of computer science at the University of Illinois at Urbana-Champaign. She is a recipient of an NSF CAREER award and of the JCAI-JAIR Best Paper Prize 2018. She has also served as a co-chair of CoNLL 2013, program chair of EMNLP 2018, and as member and chair (2018-2019) of the NAACL board.

Multimodal Generative LLMs: Unification, Interpretability, Evaluation

Mohit Bansal
University of North Carolina at Chapel Hill

Date/time: Saturday, November 4, 2023, 9:00–10:00

Abstract: In this talk, I will present our journey of large-scale multimodal pretrained (generative) models across various modalities (text, images, videos, audio, layouts, etc.) and enhancing important aspects such as unification, efficiency, interpretability, and evaluation. We will start by discussing early cross-modal vision-and-language pretraining models (LXMERT). We will then look at early unified models (VL-T5) to combine several multimodal tasks (such as visual QA, referring expression comprehension, visual entailment, visual commonsense reasoning, captioning, and multimodal translation) by treating all tasks as text generation. We will also look at recent advanced unified models (with joint objectives and architecture, as well as newer unified modalities during encoding and decoding) such as textless video-audio transformers (TVLT), vision-text-layout transformers for universal document processing (UDOP), and composable any-to-any text-audio-image-video multimodal generation (CoDi). Second, we will discuss interpretable and controllable multimodal generation via LLM-based planning and programming, such as layout-controllable image generation via visual programming (VPGen), consistent multi-scene video generation via LLM-guided planning (VideoDirectorGPT), and open-domain, open-platform diagram generation (DiagrammerGPT). I will conclude with important evaluation aspects of multimodal generation models, based on fine-grained skill and social bias evaluation (DALL-Eval), as well as interpretable and explainable visual programs (VPEval).

Biography: Dr. Mohit Bansal is the John R. & Louise S. Parker Professor and the Director of the MURGe-Lab (UNC-NLP Group) in the Computer Science department at UNC Chapel Hill. He received his PhD from UC Berkeley in 2013 and his BTech from IIT Kanpur in 2008. His research expertise is in natural language processing and multimodal machine learning, with a particular focus on multimodal generative models, grounded and embodied semantics, language generation and Q&A/dialogue, and interpretable and generalizable deep learning. He is a recipient of IIT Kanpur Young Alumnus Award, DARPA Director's Fellowship, NSF CAREER Award, Google Focused Research Award, Microsoft Inves-

tigator Fellowship, Army Young Investigator Award (YIP), DARPA Young Faculty Award (YFA), and outstanding paper awards at ACL, CVPR, EACL, COLING, and CoNLL. He has been a keynote speaker for the AACL 2023 and INLG 2022 conferences. His service includes ACL Executive Committee, ACM Doctoral Dissertation Award Committee, CoNLL Program Co-Chair, ACL Americas Sponsorship Co-Chair, and Associate/Action Editor for TACL, CL, IEEE/ACM TASLP, and CSL journals. Webpage: `https://www.cs.unc.edu/~mbansal/`

Can We Hear You? Models for Listening in the Animal Kingdom and Scientific Communities

Katherine Zacarian
Earth Species Project

Date/time: Friday, November 3, 2023, 9:00–9:30

Abstract: Katie Zacarian is the CEO and co-founder of Earth Species Project. Earth Species Project is a non-profit organization with the mission of decoding animal communication. To achieve this mission, ESP collaborates broadly across many scientific disciplines, bringing together AI researchers, biological scientists, ethicists, philosophers, conservationists, and more. With so many disciplines of research partnering with the organization, ESP's listening to animals hinges almost entirely on its ability to listen to all of the various collaborators contributing to the work. In this talk, Zacarian talks about the challenges, and joys of listening deeply to the science and expertise outside of our own disciplines–and eventually–outside of our species.

Biography: Katie Zacarian is the CEO and Co-Founder of Earth Species Project, a 501(c)3 organization devoted to decoding animal communication. Prior to her work leading AI researchers and biological scientists in solving one of the oldest questions of our species, Katie led efforts to build tracking measurement technologies for researchers in the Sumatran jungle. Her work focuses on these front lines of conservation research, asking, "how can technology help solve problems for the park rangers, wildlife veterinarians, and biologists working on the ground in conservation research?" In 2019, she was honored with the Next Generation Conservationist Award by The International Seakeepers Society for her work in conversation technology and public education. Prior to her leadership roles in conservation technology, Katie developed and launched products at Facebook (now Meta), working on some of the site's most foundational features still in use today. Katie holds a BA from Harvard University and lives in San Francisco, CA where she is an avid surfer and whale-watcher.

Natural Language Processing Research and Product Development in Indonesia

Ayu Purwarianti
Bandung Institute of Technology (ITB)

Date/time: Friday, November 3, 2023, 9:30–10:00

Abstract: In this talk, I will present the efforts of Natural Language Processing (NLP) research and product development in Indonesia. In about two decades ago, the NLP researches in Indonesia was not well developed, whether in method or NLP data sources. The number of NLP researchers itself is quite low, compared to the population number of Indonesian people. There was even no available research publication for any regional language in Indonesia. Since then, there have been several attempts to push NLP research in Indonesia, including the forming of Indonesian Association for Computational Linguistics, the participation of Indonesia researchers as part of several international conference committees, and NLP courses in universities which motivate students to conduct NLP researches. Some obstacles are still exist until now including the research fund support from the government; but the number of Indonesian NLP researchers are now increasing, and so is the NLP related start ups in Indonesia.

Biography: Ayu Purwarianti was graduated from PhD program at Toyohashi University of Technology in December 2007 with dissertation title of "Cross Lingual Question Answering System (Indonesian Monolingual QA, Indonesian-English CLQA, Indonesian-Japanese CLQA)". Since then, she has worked as a lecturer at ITB (Bandung Institute of Technology). Other than teaching and doing research, her other activity is in Indonesian Association for Computational Linguistics where she was elected as the chair for 2016-2018; and she was also the chair of IEEE Education chapter of Indonesian section for 2017-2019. She has joined IABEE since 2015 until now. She also founded a start-up named Prosa.ai in 2018. She is now the Chair of Artificial Intelligence Center at ITB since August 2019.

4

Tutorials: Wednesday, November 1

Overview: Tutorials

In parallel with *the workshops* on the same day (p.58).

9:00 – 12:30 **Morning Tutorials**

T1: Language and Robotics: Toward Building Robots Coexisting with Human Society Using Language Interface Conversations

Tabanan

T2: Current Status of NLP in South East Asia with Insights from Multilingualism and Language Diversity

Gianyar 1

T3: Practical Tools from Domain Adaptation for Designing Inclusive, Equitable, and Robust Generative AI

Bangli 1

10:30 – 11:00 **Coffee break**

12:30 – 14:00 **Lunch break**

14:00 – 17:30 **Afternoon Tutorials**

T4: Editing Large Language Models

Tabanan

T5: Learning WHO Saying WHAT to WHOM in Multi-Party Conversations

Gianyar 1

T6: Developing State-Of-The-Art Massively Multilingual Machine Translation Systems for Related Languages

Bangli 1

15:30 – 16:00 **Coffee break**

18:00 – 20:00 **Welcome Reception**

Beach Bale

Tutorial 1

Language and Robotics: Toward Building Robots Coexisting with Human Society Using Language Interface

Yutaka Nakamura, **Shuhei Kurita**, **Koichiro Yoshino**

Wednesday, November 1, 2023, 9:00–12:30

Tabanan

Robots are one of the archetypes of AI systems we imagine, and the realization of such robots operating in the real world with language interfaces has long been a dream of us. This introductory tutorial aims to help researchers who will start language and robotics, (LangRobo) research in the future by summarizing three points: awareness of the community's issues, recent approaches for these issues, and remaining problems. We arrange this tutorial involving not only NLP researchers but also robotics researchers in order to raise issues that are relevant to actual robotics problems. There are several difficulties in connecting NLP and robotics, but the following three are particularly problematic:
- The great difference in granularity between language and robot behavior.
- Robotics tasks involving real-world control often do not allow for language ambiguity.
- Language expressions themselves are often ambiguous and require background knowledge or commonsense reasoning to understand them correctly.
Many recent works have suggested that deep learning or LLMs can provide solutions. This tutorial summarizes the recent approaches to the language and robotics problem using such learning-based approaches. The goal of this tutorial is to share the discussion on how these problems can be solved in the future.

Yutaka Nakamura is a Team Leader at the Institute of Physical and Chemical Research (RIKEN) and an Affiliate Professor at Osaka University. He received his degree, Dr. Eng., from Nara Institute of Science and Technology (NAIST) in 2004. He worked at Osaka University as an assistant professor and an associate professor. Since 2020, he has been working at Guardian Robot Project (GRP) of RIKEN as the team leader of Behavior Learning research team. He is working on areas of robotics, control, and human-robot interaction.

Shuhei Kurita is a Research Scientist at Center for Advanced Intelligence Project (AIP), RIKEN. He received his Ph.D. in informatics from Kyoto University in 2019. He is a visiting researcher in New York University for Assoc. Prof. Kyunghyun Cho from 2020. His paper "Neural Joint Model for Transition-based Chinese Syntactic Analysis" was selected as the outstanding paper of ACL2017 (Kurita et al., 2017). He is working on natural language understanding in the real world expressed in images, 3D scenes and photorealistic simulator. He has actively published papers in natural language processing, learning representations and computer vision venues.

Koichiro Yoshino is a Team Leader at the Institute of Physical and Chemical Research (RIKEN) and an Affiliate Professor at Nara Institute of Science and Technology (NAIST). He received his Ph.D. in informatics from Kyoto University in 2014. He worked at Kyoto University as a postdoc and at NAIST as an assistant professor. Since 2020, he has been working at Guardian Robot Project (GRP) of RIKEN as the team leader of Knowledge Acquisition and Dialogue research team. From 2019 to 2020, he was a visiting researcher at Heinrich-Heine-Universität Düsseldorf, Germany. He is working on spoken and natural language processing areas, especially robot dialogue systems. Dr. Koichiro Yoshino received several honors, including the best paper award of IWSDS2020 and the best paper award of the 1st NLP4ConvAI workshop. He is a member of IEEE Speech and Language Processing Technical Committee (SLTC), a member of Dialogue System Technology Challenge (DSTC) Steering Committee, an action editor of ACL Rolling Review (ARR), and a board member of SIGdial. He is a member of ACL, IEEE, SIGDIAL, IPSJ, JSAI, ANLP and RSJ.

Tutorial 2

Current Status of NLP in South East Asia with Insights from Multilingualism and Language Diversity

Alham Fikri Aji, **Jessica Zosa Forde**, **Alyssa Marie Loo**, **Lintang Sutawika**, **Skyler Wang**, **Genta Indra Winata**, **Zheng-Xin Yong**, **Ruochen Zhang**, **A. Seza Dogruoz**, **Yin Lin Tan**, **Jan Christian Blaise Cruz**

Wednesday, November 1, 2023, 9:00–12:30

Gianyar 1

South East Asia (SEA) is a region with immense cultural and linguistic diversity—a melting pot of cultures, religions, and languages, home to over 1000 languages. In addition, multilingualism (i.e., speaking more than one language or dialect) is widely practiced on a daily basis. Despite the variety of languages, there is relatively less research on natural language processing (NLP) of SEA languages and their users in the area compared to languages in other regions. The scarcity of available datasets for the region's languages presents a challenge for developing NLP technology for SEA languages. Other challenges include the complexity of language use in the region, such as code-switching, and the inaccessibility of language technology to certain groups of SEA researchers due to constraints on computing resources. This tutorial will present an overview of language issues in the SEA region, link multilingualism and computational sociolinguistics with historical and societal perspectives, and provide a summary of existing datasets for computational linguistics research, language models and NLP systems, and evaluation benchmarks. The tutorial will also characterize the existing research ecosystem in SEA, including community-based initiatives working on SEA languages and opportunities for developing NLP technologies for SEA languages.

Alham Fikri Aji is an assistant professor in MBZUAI. His research focuses on multi-lingual and cross-lingual NLP, especially for under resourced languages and communities. His work area also includes data construction as well as data-efficient systems, and compute-efficient models for better accessibility.

Jessica Zosa Forde is a PhD Candidate at Brown University. Jessica's research focuses on the evaluation of deep learning models, to improve their reliability in high stakes domains. Jessica presented a tutorial on reproduciblity in NLP at ACL in 2022.

Alyssa Marie Loo is an undergraduate in Linguistics and Computer Science at Brown University. Her research focuses on interpretability of large language models and their alignment with human linguistic behavior.

Lintang Sutawika is Researcher at EleutherAI. He is a proponent of open source software and machine learning artifacts. His work has comprised of extending language models to more languages, interpreting language models and maintaining software for language model evaluation.

Skyler Wang is a PhD Candidate at UC Berkeley and a Visiting Sociologist at Meta AI. Broadly, Skyler's research focuses on creating socially and ethically-grounded machine translation technologies for low-resource language communities. He is a Sociologist on Meta AI's "No Language Left Behind" team.

Genta Indra Winata is a Senior Research Scientist at Bloomberg. His research focuses on multilingual, cross-lingual, language models, dialogue system, and low-resource NLP. His work area includes few-shot learning and evaluation of large language models.

Zheng-Xin Yong is a PhD Candidate at Brown University. His research focuses on cross-lingual NLP, large language models, and AI safety.

Ruochen Zhang is a PhD Candidate at Brown University. Her research interests lie in multi-cross-lingual learning, evaluation and application of large language models.

A. Seza Doğruöz is a tenured Associate Professor at Ghent University. She conducts inter-disciplinary research on multilingualism, sociolinguistics and computational linguistics.

Yin Lin Tan is a PhD student at Stanford University. Her research focuses on sociolinguistic variation, phonetics, and multilingualism.

Jan Christian Blaise Cruz is an AI Research Engineer at Samsung R&D Institute Philippines. His research revolves around low-resource techniques for translation and language generation.

Tutorial 3

Practical Tools from Domain Adaptation for Designing Inclusive, Equitable, and Robust Generative AI

Anthony Sicilia, **Malihe Alikhani**

Wednesday, November 1, 2023, 9:00–12:30

Bangli 1

Generative language technologies have become integral to everyday communication, shaping social interactions and informing critical decision-making processes in areas such as recruitment, healthcare, and education. However, they often struggle to grasp the "long tail" of data distributions — concepts less frequently observed during training — which could have significant repercussions. These models may marginalize underrepresented groups by failing to comprehend preferred communication styles, such as code-switching, or perpetuating societal biases like gender bias. Sectors like healthcare, education, and law, requiring personalization and exhibiting nuanced linguistic features, are also particularly affected when pre-trained models misconstrue or overlook "long tail" data concepts. While methods like distillation of smaller language models, active learning, and other bias mitigation strategies can augment traditional training techniques, a careful statistical analysis is essential for their effective application. This tutorial offers a comprehensive examination of how to develop equitable, robust, and inclusive language technologies using statistical tools from Domain Adaptation (DA) that catalyze positive social change. We will delve into strategies for bias mitigation, explore how to measure bias, and examine open problems in creating culturally-grounded and inclusive language technologies. Accompanying materials including code notebooks, python packages, and coursework will be provided.

Anthony Sicilia is a 5th year Ph.D student, specializing in applications of learning theory and domain adaptation theory to NLP problems such as inclusivity, equity, and robustness. He has experience in practical deployment of NLP systems, leading an Alexa Prize TaskBot team (focused on inclusivity and collaboration) to 3rd place overall in this international contest. He has published 4 papers on robust NLP at *ACL venues, which are present in the reading list: Atwell et al. (2022); Sicilia and Alikhani (2022); Sicilia et al. (2022b); Sicilia and Alikhani (2023). He also received a best paper award at UAI 2022 for his work on novel PACBayesian DA theory for multiclass neural-networks (Sicilia et al., 2022a). His work spans application of DA theory to diverse areas.

Malihe Alikhani is an expert in natural language processing (NLP) and machine learning. Alikhani's research interests center on using representations of communicative structure to improve ethical and practical machine learning models. One of the main focuses of her recent research has been on studying formal methods of machine learning for designing equitable and robust NLP tasks. This includes using tools from learning theory for efficient dialogue management, text generation, classification and measuring and mitigating biases in generation and classification tasks (Atwell et al., 2022; Sicilia and Alikhani, 2022; Sicilia et al., 2022b; Atwell et al., 2021; Sicilia and Alikhani, 2023; Sicilia et al., 2022a). Her work in these areas have received three best paper awards at UAI 2022, ACM UMAP 2022 and INLG 2021.

Tutorial 4

Editing Large Language Models

Ningyu Zhang, **Yunzhi Yao**, **Shumin Deng**

Wednesday, November 1, 2023, 14:00–17:30

Tabanan

Even with their impressive abilities, Large Language Models (LLMs) such as ChatGPT are not immune to issues of factual or logically consistent. Concretely, the key concern is how to seamlessly update those LLMs to correct mistakes without resorting to an exhaustive retraining or continuous training procedure, both of which can demand significant computational resources and time. Thus, the capability to edit LLMs offers an efficient solution to alter a model's behavior, notably within a distinct area of interest, without negatively impacting its performance on other tasks. Through this tutorial, we strive to acquaint interested NLP researchers with recent and emerging techniques for editing LLMs. Specifically, we aim to present a systematic and current overview of cutting-edge methods, supplemented with practical tools, and unveil new research opportunities for our audience. All resources can be found at `https://github.com/zjunlp/ModelEditingPapers`.

Ningyu Zhang is an associate professor/doctoral supervisor at Zhejiang University, leading the group about KG and NLP technologies. He has supervised to construct a information extraction toolkit named DeepKE2 (1.9K+ stars on Github). His research interest include knowledge graph and natural language processing. He has published many papers in top international academic conferences and journals such as Natural Machine Intelligence, Nature Communications, NeurIPS, ICLR, AAAI, IJCAI,WWW, KDD, SIGIR, ACL, ENNLP, NAACL, and IEEE/ACM Transactions on Audio Speech and Language. He has served as Area Chair for ACL 2023, ARR Action Editor, Senior Program Committee member for IJCAI 2023, Program Committee member for EMNLP, NAACL, NeurIPS, ICLR, ICML, WWW, SIGIR, KDD, AAAI, and reviewer for TKDE, TKDD.

Yunzhi Yao is a Ph.D candidate at at School of Computer Science and Technology, Zhejiang University. Her research interests focus on Editing Large Language Models and Knowledge-enhanced Natural Language Processing. He has been research intern at Microsoft Research Asia supervised by Shaohan Huang, and research intern at Alibaba Group. He has published many papers in ACL, EMNLP, NAACL, SIGIR. For tutorial experience, he has given talks at AI-TIME to deliver his recent works. Moreover, he is the first author of the paper "Editing Large Language Models: Problems, Methods, and Opportunities" which is related to this tutorial.

Shumin Deng is a research fellow at Department of Computer Science, School of Computing (SoC), National University of Singapore. She have obtained her Ph.D. degree at School of Computer Science and Technology, Zhejiang University. Her research interests focus on Natural Language Processing, Knowledge Graph, Information Extraction, Neuro-Symbolic Reasoning and LLM Reasoning. She has been awarded 2022 Outstanding Graduate of Zhejiang Province, China; 2020 Outstanding Intern in Academic Cooperation of Alibaba Group. She is a member of ACL, and a member of the Youth Working Committee of the Chinese Information Processing Society of China. She has serves as a Research Session (Information Extraction) Chair for EMNLP 2022, and a Publication Chair for CoNLL 2023.

Tutorial 5

Learning WHO Saying WHAT to WHOM in Multi-Party Conversations

Jia-Chen Gu, **Zhuosheng Zhang**, **Zhen-Hua Ling**

Wednesday, November 1, 2023, 14:00–17:30

Gianyar 1

Multi-party conversations (MPC) are a more practical and challenging scenario involving more than two interlocutors. This research topic has drawn significant attention from both academia and industry, and it is nowadays counted as one of the most promising research areas in the field of dialogue systems. In general, MPC algorithms aim at addressing the issues of Who saying What to Whom, specifically, who speaks, say what, and address whom. The complicated interactions between interlocutors, between utterances, and between interlocutors and utterances develop many variant tasks of MPC worth investigation. In this tutorial, we present a comprehensive survey of recent advances in MPC. In particular, we summarize recent advances on the research of MPC modeling which is categorized by Who saying What to Whom. Finally, we highlight the challenges which are not yet well addressed in MPC and present future research directions.

Jia-Chen Gu is currently a Postdoctoral Researcher at University of Science and Technology of China. His research interests lie within machine learning for dialogue systems.

Zhuosheng Zhang is currently an Assistant Professor at Shanghai Jiao Tong University. His research interests include natural language processing, dialogue systems, and large language models. He has given a tutorial on Machine Reading Comprehension: The Role of Contextualized Language Models and Beyond at IJCAI 2021.

Zhen-Hua Ling is a Professor with the University of Science and Technology of China. He was a Visiting Scholar at the University of Washington and a Marie Curie Fellow at the University of Edinburgh. His research interests include speech processing and natural language processing. He was the recipient of the IEEE Signal Processing Society Young Author Best Paper Award in 2010. He was an Associate Editor of IEEE/ACM Transcations on Audio, Speech, and Language Processing from 2014 to 2018.

Tutorial 6

Developing State-Of-The-Art Massively Multilingual Machine Translation Systems for Related Languages

Jay Gala, **Pranjal A. Chitale**, **Raj Dabre**

Wednesday, November 1, 2023, 14:00–17:30

Bangli 1

The race for developing state-of-the-art (SOTA) machine translation (MT) systems often gives the impression that this can only be done by large organizations, most of which do not fully open-source their systems. In this tutorial, we dispel this myth by generalizing our experiences in developing SOTA MT systems for related languages. We cover topics ranging from (a) the history of MT systems for related languages, (b) curating high-quality datasets, manually created as well as mined, (c) creating domain-diverse benchmarks, (d) compact but high-quality open-source MT systems that surpass other systems despite being an order of magnitude smaller in terms of parameters and computational costs than massively multilingual generic systems and (e) robust automatic and human evaluation. We hope that our tutorial encourages other groups, regardless of scale, to engage in focussed efforts on related languages or language groups to develop open-source, high-quality MT systems.

Jay Gala received his B.E. from the University of Mumbai, India. He is an AI resident at AI4Bharat, where he primarily works on building open-source models, datasets, and benchmarks for Indian languages. His research interests broadly span in the area of multimodal and multilingual representation learning, specifically in the context of data-efficient learning, training dynamics, and generalization.

Pranjal A. Chitale received his B.E. from the University of Mumbai, India. He is currently an M.S. student at IIT Madras advised by Prof. Mitesh Khapra, working at the AI4Bharat lab. His interests lie in the fields of multilingual learning and data-efficient techniques and works on building open-source datasets, models, and benchmarks for Indic languages. His thesis work will be primarily focused on the development of multilingual NMT systems for Indian languages.

Raj Dabre received his M.Tech. from IIT Bombay, India and his Ph.D. from Kyoto University, Japan. He is a researcher at NICT, Japan and a visiting researcher at AI4Bharat. His research interests center on natural language processing, particularly neural machine translation for low resource languages, and on model compression and computing efficiency. He has MT and NLG related publications in ACL, EMNLP, AAAI, NAACL, COLING, INTERSPEECH and WMT. He is a current member of the organizing committee of the Workshop on Asian Translation. He has previously conducted tutorials on neural machine translation and multilingual machine translation at IJCNLP 2017 and COLING 2020, respectively.

5

Main Conference: November 2–4

Overview: Thursday, November 2

8:30 – 9:00 **Opening Address**
9:00 – 10:00 **Keynote Speech (Julia Hockenmaier)**
Karangasem 1+2
10:00 – 10:30 **Coffee Break**
10:30 – 12:00 **Session 1A: Dialog Systems and Generation**
Karangasem 1+2
10:30 – 12:00 **Session 1B: Question Answering**
Negara
12:00 – 14:00 **Lunch**
Karangasem 3
14:00 – 15:30 **Demo**
Negara
Session 1C: Regional Language Processing
Karangasem 1+2
15:30 – 16:00 **Coffee Break**
16:00 – 17:30 **Demo**
Negara
Findings Session
Karangasem 1+2

Overview: Friday, November 3

9:00 – 9:30 **Featured Plenary Talk (Katherine Zacarian)**
9:30 – 10:00 **Featured Plenary Talk (Ayu Purwarianti)**
Karangasem 1+2
10:00 – 10:30 **Coffee Break**
10:30 – 12:00 **Session 2A: Resources and Evaluation**
Karangasem 1+2
10:30 – 12:00 **Session 2B: Data Mining, Information Extraction and Retrieval**
Negara
12:00 – 14:00 **Lunch**
Karangasem 3
14:00 – 15:00 **Keynote Speech (Hinrich Schütze)**
Karangasem 1+2
15:00 – 15:30 **Coffee Break**
15:30 – 17:30 **Poster Session**
Foyer

18:00 – 22:00 **Dinner Banquet**
South Beach

Overview: Saturday, November 4

9:00 – 10:00 **Keynote Speech (Mohit Bansal)**
Karangasem 1+2
10:00 – 10:30 **Coffee Break**
10:30 – 12:00 **Session 3A: Multilingual and Multimodal Analysis**
Karangasem 1+2
10:30 – 12:00 **Session 3B: Machine Learning and Model Interpretability**
Negara
12:00 – 14:00 **Lunch**
Karangasem 3
14:00 – 15:30 **Session 3C: Semantics**
Karangasem 1+2
14:00 – 15:30 **Session 3D: NLP Applications**
Negara
15:30 – 16:00 **Coffee Break**
16:00 – 17:30 **Best paper session and closing remark**
Karangasem 1+2

Main Conference: Thursday, November 2

Session 1A – 10:30-12:00

Dialog Systems and Generation
Karangasem 1+2
Chair: Lili Mou

SILVER: Self Data Augmentation for Out-of-Scope Detection in Dialogues
Chunpeng Ma and Takuya Makino

Detecting out-of-scope (OOS) utterances is crucial in task-oriented dialogue systems, but obtaining enough annotated OOS dialogues to train a binary classifier directly is difficult in practice. Existing data augmentation methods generate OOS dialogues automatically, but their performance usually depends on an external corpus. This dependence not only induces uncertainty, but also reduces the quality of generated dialogues. Specifically, all of them are out-of-domain (OOD). Herein we propose SILVER, a self data augmentation method that does not use external data. It addresses issues of previous research and improves the accuracy of OOS detection (false positive rate: 90.5% to 47.4%). Furthermore, SILVER successfully generates high-quality in-domain (IND) OOS dialogues in terms of naturalness (percentage: 8% to 68%) and OOS correctness (percentage: 74% to 88%), as evaluated by human workers.

Sentiment Aided Graph Attentive Contextualization for Task Oriented Negotiation Dialogue Generation
Aritra Raut, Sriparna Saha, Anutosh Maitra, and Roshni Ramnani
Over the past several years, demand and popularity of using virtual assistants to finish jobs like service scheduling and

online shopping have increased. While keeping the user's request in mind, an effective task-oriented virtual agent must strive to improve the seller's profit. Therefore, in order to achieve the best possible trade-off between the parties, this form of virtual agents has to have strong negotiating abilities. Although current conversational agents are quite good at making fluent sentences, they are still unable to use strategic thinking. In order to more effectively contextualize the choice of the next set of negotiation methods while producing answers, we develop Nego-GAT, an end-to-end negotiation system that includes sentiment information and graph attention embedding into GPT-2. Our self-supervised model beats earlier cutting-edge negotiation models in terms of both the precision of strategy/dialogue act prediction and the caliber of the generated dialogue responses.

MQAG: Multiple-choice Question Answering and Generation for Assessing Information Consistency in Summarization
Potsawee Manakul, Adian Liusie, and Mark Gales
State-of-the-art summarization systems can generate highly fluent summaries. These summaries, however, may contain factual inconsistencies and/or information not present in the source. Hence, an important component of assessing the quality of summaries is to determine whether there is information consistency between the source and the summary. Existing approaches are typically based on lexical matching or representation-based methods. In this work, we introduce an alternative scheme based on standard information-theoretic measures in which the information present in the source and summary is directly compared. We propose a Multiple-choice Question Answering and Generation framework, MQAG, which approximates the information consistency by computing the expected statistical distance between summary and source answer distributions over automatically generated multiple-choice questions. This approach exploits multiple-choice answer probabilities, as predicted answer distributions can be compared. We conduct experiments on four summary evaluation datasets: QAG-CNNDM/XSum, XSum-Hallucination, Podcast Assessment, and SummEval. Experiments show that MQAG, using models trained on SQuAD or RACE, outperforms existing evaluation methods on the majority of tasks.

Toward Unified Controllable Text Generation via Regular Expression Instruction
Xin Zheng, Hongyu Lin, Xianpei Han, and Le Sun

Controllable text generation is a fundamental aspect of natural language generation, with numerous methods proposed for different constraint types. However, these approaches often require significant architectural or decoding modifications, making them challenging to apply to additional constraints or resolve different constraint combinations. To address this, our paper introduces Regular Expression Instruction (REI), which utilizes an instruction-based mechanism to fully exploit regular expressions' advantages to uniformly model diverse constraints. Specifically, our REI supports all popular fine-grained controllable generation constraints, i.e., lexical, positional, and length, as well as their complex combinations, via regular expression-style instructions. Our method only requires fine-tuning on medium-scale language models or few-shot, in-context learning on large language models, and requires no further adjustment when applied to various constraint combinations. Experiments demonstrate that our straightforward approach yields high success rates and adaptability to various constraints while maintaining competitiveness in automatic metrics and outperforming most previous baselines.

Reranking for Natural Language Generation from Logical Forms: A Study based on Large Language Models
Levon Haroutunian, Zhuang Li, Lucian Galescu, Philip Cohen, Raj Tumuluri, and Gholamreza Haffari
Large language models (LLMs) have demonstrated impressive capabilities in natural language generation. However, their output quality can be inconsistent, posing challenges for generating natural language from logical forms (LFs). This task requires the generated outputs to embody the exact semantics of LFs, without missing any LF semantics or creating any hallucinations. In this work, we tackle this issue by proposing a novel generate-and-rerank approach. Our approach involves initially generating a set of candidate outputs by prompting an LLM and subsequently reranking them using a task-specific reranker model. In addition, we curate a manually collected dataset to evaluate the alignment between different ranking metrics and human judgements. The chosen ranking metrics are utilized to enhance the training and evaluation of the reranker model. By conducting extensive experiments on three diverse datasets, we demonstrate that the candidates selected by our reranker outperform those selected by baseline methods in terms of semantic consistency and fluency, as measured by three comprehensive metrics. Our findings provide strong evidence for the effectiveness of our approach in improving the quality of generated outputs.

Incorporating Singletons and Mention-based Features in Coreference Resolution via Multi-task Learning for Better Generalization
Yilun Zhu, Siyao Peng, Sameer Pradhan, and Amir Zeldes
Previous attempts to incorporate a mention detection step into end-to-end neural coreference resolution for English have been hampered by a lack of singleton mention span data as well as other entity information. This paper presents a coreference model that learns singletons as well as features such as entity type and information status, via a multi-task learning-based approach. This approach achieves new state-of-the-art scores on the OntoGUM benchmark (+2.7 points) and increases robustness on multiple out-of-domain datasets (+2.3 points on average), likely due to greater generalizability for mention detection and utilization of more data from singletons, when compared to only coreferent mention pair matching.

Session 1B – 10:30-12:00

Question Answering
Negara
Chair: Shumin Deng

Attacking Open-domain Question Answering by Injecting Misinformation
Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang

With a rise in false, inaccurate, and misleading information in propaganda, news, and social media, real-world Question Answering (QA) systems face the challenges of synthesizing and reasoning over misinformation-polluted contexts to derive correct answers. This urgency gives rise to the need to make QA systems robust to misinformation, a topic previously unexplored. We study the risk of misinformation to QA models by investigating the sensitivity of open-domain QA models to corpus pollution with misinformation documents. We curate both human-written and model-generated false documents that we inject into the evidence corpus of QA models, and assess the impact on the performance of these systems. Experiments show that QA models are vulnerable to even small amounts of evidence contamination brought by misinformation, with large absolute performance drops on all models. Misinformation attack brings more threat when fake documents are produced at scale by neural models or the attacker targets on hacking specific questions of interest. To defend against such a threat, we discuss the necessity of building a misinformation-aware QA system that integrates question-answering and misinformation detection in a joint fashion.

Don't be Blind to Questions: Question-Oriented Math Word Problem Solving
Zhenwen Liang, Jipeng Zhang, and Xiangliang Zhang

Solving math word problems (MWP) is a challenging task for natural language processing systems, as it requires to not only identify and comprehend the problem description within the context, but also to deduce a solution in accordance with the posed question. Previous solvers have been found to prioritize the context over the question, resulting in low performance when solving multiple questions under the same context. In this paper, we present a question-oriented strategy to address this issue and improve the generalizability of MWP solvers. Our approach features an entity-aware encoder that enhances the connection between MWP context and question via entities in established dependency graphs, aiming at obtaining better problem representations. Then, a question-guided decoder is trained using a contrastive learning strategy to enhance the question representations. Empirical evaluations on four benchmarks demonstrate that our method outperforms previous solvers and exhibits a favorable balance between efficacy and efficiency in MWP solving. In addition, our solver is not reliant on any specific pre-trained model and demonstrates seamless compatibility with different pre-trained model backbones.

GrailQA++: A Challenging Zero-Shot Benchmark for Knowledge Base Question Answering
Ritam Dutt, Sopan Khosla, Vinayshekhar Bannihatti Kumar, and Rashmi Gangadharaiah

Most benchmarks designed for question answering over knowledge bases (KBQA) operate with the i.i.d. assumption where one encounters the same schema items during inference as those observed during training. Recently, the GrailQA dataset was established to evaluate zero-shot generalization capabilities of KBQA models as a departure from the i.i.d. assumption. Reasonable performance of current KBQA systems on the zero-shot GrailQA split hints that the field might be moving towards more generalizable systems. In this work, we observe a bias in the GrailQA dataset towards simpler one or two-hop questions, which results in an inaccurate assessment of the aforementioned prowess. We propose GrailQA++, a challenging zero-shot KBQA test set that contains more questions relying on complex reasoning. We leverage the concept of graph isomorphisms to control the complexity of the questions and to ensure that our proposed test set has a fair distribution of simple and complex questions. Existing KBQA models suffer a substantial drop in performance on our constructed new test set as compared to the GrailQA zero-shot split. Our analysis reveals how isomorphisms can be used to understand the complementary strengths of different KBQA models and provide a deeper insight into model mispredictions. Overall, our paper highlights the *non-generalizability* of existing models and the necessity for designing more challenging benchmarks. Our dataset is available at https://github.com/sopankhosla/GrailQA-PlusPlus

Automatic Translation of Span-Prediction Datasets
Ofri Masad, Kfir Bar, and Amir Cohen

Generating high-quality non-English language datasets is crucial for achieving high performance in various Natural Language Processing (NLP) tasks. In this paper, we propose a new approach for translating NLP datasets that relies on a two-phase pipeline and online translation services. Our approach focuses on solving the alignment problem that affects span prediction tasks and utilizes automatically labeled data for training an alignment model. We demonstrate that our model-based approach shows higher accuracy than any other alignment method and improves the average F1 score on several Question-Answering (QA) datasets, specifically on the XQuAD Translated-train dataset, achieving new state-of-the-art results.

Generating and Answering Simple and Complex Questions from Text and from Knowledge Graphs
Kelvin Han and Claire Gardent

While both text and Knowledge Graphs (KG) may be used to answer a question, most current Question Answering and Generation models only work on a single modality. In this paper, we introduce a multi-task model such that questions

can be generated and answered from both KG and text. The model has wide coverage and handles both simple (one KG fact) and complex (more than one KG fact) questions. Extensive internal, cross-modal and external consistency checks, and analysis of the quality of the generated questions, show that our approach outperforms previous work. Our data and modeling also leads to improvements in downstream tasks, including better performance with finetuning Open-Domain QA architectures and better correlation with human judgments than the Data-QuestEval metric which was previously proposed for evaluating the semantic adequacy of KG-to-Text generations.

SQUARE: Automatic Question Answering Evaluation using Multiple Positive and Negative References
Matteo Gabburo, Siddhant Garg, Rik Koncel-Kedziorski, and Alessandro Moschitti
Evaluation of QA systems is very challenging and expensive, with the most reliable approach being human annotations of correctness of answers for questions. Recent works (AVA, BEM) have shown that transformer LM encoder based similarity metrics transfer well for QA evaluation, but they are limited by the usage of a single correct reference answer. We propose a new evaluation metric: SQuArE (Sentence-level QUestion AnsweRing Evaluation), using multiple reference answers (combining multiple correct and incorrect references) for sentence-form QA. We evaluate SQuArE on both sentence-level extractive (Answer Selection) and generative (GenQA) QA systems, across multiple academic and industrial datasets, and show that it outperforms previous baselines and obtains the highest correlation with human annotations.

Session 1C – 14:00-15:30

Regional Language Processing
Karangasem 1+2
Chair: Xiang Dai

Assessment of Pre-Trained Models Across Languages and Grammars
Alberto Muñoz-Ortiz, David Vilares, and Carlos Gómez-Rodríguez

We present an approach for assessing how multilingual large language models (LLMs) learn syntax in terms of multi-formalism syntactic structures. We aim to recover constituent and dependency structures by casting parsing as sequence labeling. To do so, we select a few LLMs and study them on 13 diverse UD treebanks for dependency parsing and 10 treebanks for constituent parsing. Our results show that the framework is consistent across encodings, and notes the role played by factors such as the syntactic formalism, the LLM differences, and the pretraining and assessment data.

Generation of Korean Offensive Language by Leveraging Large Language Models via Prompt Design
Jisu Shin, Hoyun Song, Huije Lee, Fitsum Gaim, and Jong Park
The research for detecting offensive language on online platforms has much advanced. However, the majority of these studies have primarily focused on English. Given the unique characteristics of offensive language, where social and cultural contexts significantly influence content understanding, language-specific datasets are essential. Acquiring comprehensive datasets in Korean, a less-resourced language, has mostly relied on human annotations, suffering from inherent limitations in terms of labor intensity and potential annotator bias. Automatic generation of datasets using generative methods offers an alternative approach to address these limitations, yet faces challenges in capturing linguistic and cultural diversities while maintaining native-level fluency. To address these challenges, we introduce a prompt design methodology, Korean Offensive language Machine Generation (K-OMG), using large language models. By manipulating three prompt factors, we find an effective prompt design to generate culturally aligned offensive language with fluent expressions. Experimental results demonstrate the high quality and utility of our automatically generated dataset. Our detailed analysis shows that the proposed approach achieves exceptional fluency in generating texts while effectively incorporating social and cultural diversities.

Question Answer Generation in Bengali: Mitigating the scarcity of QA datasets in a low-resource language
Md Shihab Shahriar, Ahmad Al Fayad Chowdhury, Md. Amimul Ehsan, and Abu Raihan Kamal
The scarcity of comprehensive, high-quality Question-Answering (QA) datasets in low-resource languages has greatly limited the progress of research on QA for these languages. This has inspired research on Question-Answer Generation (QAG) which seeks to synthetically generate QA pairs and minimize the human effort required to compile labeled datasets. In this paper, we present the first QAG pipeline for the Bengali language, which consists of an answer span extraction model, a question generation model, and roundtrip consistency filtering to discard inconsistent QA pairs. To train our QAG pipeline, we translate SQuAD1.1 and SQuAD2.0 using the state-of-the-art NLLB machine translation model and accurately mark the answer spans using a novel embedding-based answer alignment algorithm to construct two Bengali QA datasets that we show are superior to the only two existing machine-translated datasets in terms of quality and quantity. We use our QAG pipeline to generate more than 170,000 QA pairs to build BanglaQA, a synthetic QA dataset from 16,000 Bengali news articles spanning 5 different news categories. We demonstrate the quality of BanglaQA by human evaluation on a variety of metrics. The best-performing model among several baselines on our dataset achieves an F1

score of 86.14 falling behind human performance of 95.72 F1. Our codebase and curated datasets are publicly available at https://github.com/shihabshahriar16/BengaliQAG.git.

MasakhaNEWS: News Topic Classification for African languages
David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, et al.

Despite representing roughly a fifth of the world population, African languages are underrepresented in NLP research, in part due to a lack of datasets. While there are individual language-specific datasets for several tasks, only a handful of tasks (e.g. named entity recognition and machine translation) have datasets covering geographical and typologically-diverse African languages. In this paper, we develop MasakhaNEWS—the largest dataset for news topic classification covering 16 languages widely spoken in Africa. We provide and evaluate a set of baseline models by training classical machine learning models and fine-tuning several language models. Furthermore, we explore several alternatives to full fine-tuning of language models that are better suited for zero-shot and few-shot learning, such as: cross-lingual parameter-efficient fine-tuning (MAD-X), pattern exploiting training (PET), prompting language models (ChatGPT), and prompt-free sentence transformer fine-tuning (SetFit and the co:here embedding API). Our evaluation in few-shot setting, shows that with as little as 10 examples per label, we achieve more than 90% (i.e. 86.0 F1 points) of the performance of full supervised training (92.6 F1 points) leveraging the PET approach. Our work shows that existing supervised approaches work well for all African languages and that language models with only a few supervised samples can reach competitive performance, both findings which demonstrate the applicability of existing NLP techniques for African languages.

Analysing Cross-Lingual Transfer in Low-Resourced African Named Entity Recognition
Michael Beukman and Manuel Fokam

Transfer learning has led to large gains in performance for nearly all NLP tasks while making downstream models easier and faster to train. This has also been extended to low-resourced languages, with some success. We investigate the properties of cross-lingual transfer learning between ten low-resourced languages, from the perspective of a named entity recognition task. We specifically investigate how much adaptive fine-tuning and the choice of transfer language affect zero-shot transfer performance. We find that models that perform well on a single language often do so at the expense of generalising to others, while models with the best generalisation to other languages suffer in individual language performance. Furthermore, the amount of data overlap between the source and target datasets is a better predictor of transfer performance than either the geographical or genetic distance between the languages.

Self-Augmentation Improves Zero-Shot Cross-Lingual Transfer
Fei Wang, Kuan-Hao Huang, Kai-Wei Chang, and Muhao Chen

Zero-shot cross-lingual transfer is a central task in multilingual NLP, allowing models trained in languages with more sufficient training resources to generalize to other low-resource languages. Earlier efforts on this task use parallel corpora, bilingual dictionaries, or other annotated alignment data to improve cross-lingual transferability, which are typically expensive to obtain. In this paper, we propose a simple yet effective method, SALT, to improve the zero-shot cross-lingual transfer of the multilingual PLM without the help of such external data. By incorporating code-switching and embedding mixup with self-augmentation, SALT effectively distills cross-lingual knowledge from the multilingual PLM and enhances its transferability on downstream tasks. Experimental results on XNLI and PAWS-X show that our method is able to improve zero-shot cross-lingual transferability without external data.

Findngs Session – 16:00-17:30

Karangasem 1+2
Chair: Pavlos Vougiouklis

An in-person spotlight session for Findings papers. Each presenter can have a 2-minute lightning talk.

Main Conference: Friday, November 3

Session 2A – 10:30-12:00

Resources and Evaluation
Karangasem 1+2
Chair: Lea Frermann

NusaWrites: Constructing High-Quality Corpora for Underrepresented and Extremely Low-Resource Languages
Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, Bryan Wilie, Galih Muridan, Genta Winata, David Moeljadi, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung
Democratizing access to natural language processing (NLP) technology is crucial, especially for underrepresented and extremely low-resource languages. Previous research has focused on developing labeled and unlabeled corpora for these languages through online scraping and document translation. While these methods have proven effective and cost-efficient, we have identified limitations in the resulting corpora, including a lack of lexical diversity and cultural relevance to local communities. To address this gap, we conduct a case study on Indonesian local languages. We compare the effectiveness of online scraping, human translation, and paragraph writing by native speakers in constructing datasets. Our findings demonstrate that datasets generated through paragraph writing by native speakers exhibit superior quality in terms of lexical diversity and cultural content. In addition, we present the NusaWrites benchmark, encompassing 12 underrepresented and extremely low-resource languages spoken by millions of individuals in Indonesia. Our empirical experiment results using existing multilingual large language models emphasize the need to extend these models to more underrepresented languages.

Valla: Standardizing and Benchmarking Authorship Attribution and Verification Through Empirical Evaluation and Comparative Analysis
Jacob Tyo, Bhuwan Dhingra, and Zachary C. Lipton
Despite decades of research on authorship attribution (AA) and authorship verification (AV), inconsistent dataset splits/filtering and mismatched evaluation methods make it difficult to assess the state of the art. In this paper, we present a survey of the fields, resolve points of confusion, introduce Valla that standardizes and benchmarks AA/AV datasets and metrics, provide a large-scale empirical evaluation, and provide apples-to-apples comparisons between existing methods. We evaluate eight promising methods on fifteen datasets (including distribution shifted challenge sets) and introduce a new dataset based on texts archived by Project Gutenberg. Surprisingly, we find that a traditional ngram-based model performs best on 5 (of 7) AA tasks, achieving an average macro-accuracy of 76.50% (compared to 66.71% for a BERT-based model). However, on the two AA datasets with the greatest number of words per author, as well as on the AV datasets, BERT-based models perform best. While AV methods are easily applied to AA, they are seldom included as baselines in AA papers. We show that through the application of hard-negative mining, AV methods are competitive alternatives to AA methods. Valla and all experiment code can be found at https://github.com/JacobTyo/Valla

MedRedQA for Medical Consumer Question Answering: Dataset, Tasks, and Neural Baselines
Vincent Nguyen, Sarvnaz Karimi, Maciej Rybinski, and Zhenchang Xing

Medical question answering for consumers aims to assist consumers in finding trustworthy and relevant information for their concerns. Although some datasets exist for consumer question answering, they use synthetic questions or present difficult-to-understand answers. We introduce MedRedQA, a large non-factoid English consumer Question Answering (QA) dataset containing 51,000 pairs of consumer questions and their corresponding expert answers. MedRedQA facilitates research that aims to provide consumer-friendly responses to real-world consumer questions. We propose and benchmark three tasks for consumer medical question answering for our dataset, including (1) candidate answer ranking, (2) open-ended answer generation, and (3) answer generation with scientific evidence. Our benchmarking experiments reveal that, for the ranking task, it is feasible to retrieve expert answers within five responses in an oracle retrieval. Though, in an answer generation task, it remains challenging to align the generation toward expert answers. However, our experiments show that including scientific evidence in the prompt may reduce hallucinations in an answer generation setup.

RECESS: Resource for Extracting Cause, Effect, and Signal Spans
Fiona Anting Tan, Hansi Hettiarachchi, Ali Hürriyetoğlu, Nelleke Oostdijk, Tommaso Caselli, Tadashi Nomoto, Onur Uca, Farhana Ferdousi Liza, and See-Kiong Ng
Causality expresses the relation between two arguments, one of which represents the cause and the other the effect (or consequence). Causal relations are fundamental to human decision making and reasoning, and extracting them from natural language texts is crucial for building effective natural language understanding models. However, the scarcity of annotated corpora for causal relations poses a challenge in the development of such tools. Thus, we created Resource for Extracting Cause, Effect, and Signal Spans (RECESS), a comprehensive corpus annotated for causality at different levels, including Cause, Effect, and Signal spans. The corpus contains 3,767 sentences, of which, 1,982 are causal sentences that contain a total of 2,754 causal relations. We report baseline experiments on two natural language tasks (Causal Sentence Classification, and Cause-Effect-Signal Span Detection), and establish initial benchmarks for future work. We conduct an in-depth analysis of the corpus and the properties of causal relations in text. RECESS is a valuable resource for developing and evaluating causal relation extraction models, benefiting researchers working on topics from information retrieval to natural language understanding and inference.

The Persuasive Memescape: Understanding Effectiveness and Societal Implications of Internet Memes
Gitanjali Kumari, Pranali Shinde, and Asif Ekbal
Persuasive meme identification is a crucial task in automatically categorizing memes based on their persuasive nature. Memes, being highly influential in online communication, have the ability to shape individuals' attitudes, behaviors, and beliefs, both positively and negatively. They can be utilized to promote positive actions, challenge social norms, and raise awareness, but they can also perpetuate harmful ideologies, spread misinformation, stereotype, and manipulate emotions.

In this paper, we are addressing this challenge by empirically investigating three novel tasks: (i) Task 1: Persuasive meme detection (ii) Task 2: Identification of the effectiveness of persuasive memes, and (iii) Task 3: Identification of persuasion techniques used in persuasive memes. To this end, we make the very first attempt to release a high-quality, large-scale dataset, Persuasive_meme since there is no publicly available such dataset for Hindi-English code-mixed (Hinglish) domain. We further developed several baseline unimodal and multimodal models for these tasks. Empirical evaluations, including both qualitative and quantitative analysis, on the Persuasive_meme dataset highlight the significance of multimodality in addressing these tasks effectively. Additionally, also we discuss the limitations of the current models and emphasize the need for further research to overcome these challenges.

The Impact of Debiasing on the Performance of Language Models in Downstream Tasks is Underestimated

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki

Pre-trained language models trained on large-scale data have learned serious levels of social biases. Consequently, various methods have been proposed to debias pre-trained models. Debiasing methods need to mitigate only discriminatory bias information from the pre-trained models, while retaining information that is useful for the downstream tasks. In previous research, whether useful information is retained has been confirmed by the performance of downstream tasks in debiased pre-trained models. On the other hand, it is not clear whether these benchmarks consist of data pertaining to social biases and are appropriate for investigating the impact of debiasing. For example in gender-related social biases, data containing female words, male words, and stereotypical words are considered to be the most affected by debiasing. If there is not much data containing these words in a benchmark dataset for a target task, there is the possibility of erroneously evaluating the effects of debiasing. In this study, we compare the impact of debiasing on performance across multiple downstream tasks using a wide-range of benchmark datasets that containing female, male, and stereotypical words. Experiments show that the effects of debiasing are consistently *underestimated* across all tasks. Moreover, the effects of debiasing could be reliably evaluated by separately considering instances containing female, male, and stereotypical words than all of the instances in a benchmark dataset.

Session 2B – 10:30-12:00

Data Mining, Information Extraction and Retrieval

Negara

Chair: Preslav Nakov

Query Rewriting for Effective Misinformation Discovery

Ashkan Kazemi, Artem Abzaliev, Naihao Deng, Rui Hou, Scott Hale, Veronica Perez-Rosas, and Rada Mihalcea

We propose a novel system to help fact-checkers formulate search queries for known misinformation claims and effectively search across multiple social media platforms. We introduce an adaptable rewriting strategy, where editing actions for queries containing claims (e.g., swap a word with its synonym; change verb tense into present simple) are automatically learned through offline reinforcement learning. Our model uses a decision transformer to learn a sequence of editing actions that maximizes query retrieval metrics such as mean average precision. We conduct a series of experiments showing that our query rewriting system achieves a relative increase in the effectiveness of the queries of up to 42%, while producing editing action sequences that are human interpretable.

Target-Aware Contextual Political Bias Detection in News

Iffat Maab, Edison Marrese-Taylor, and Yutaka Matsuo

Media bias detection requires comprehensive integration of information derived from multiple news sources. Sentence-level political bias detection in news is no exception, and has proven to be a challenging task that requires an understanding of bias in consideration of the context. Inspired by the fact that humans exhibit varying degrees of writing styles, resulting in a diverse range of statements with different local and global contexts, previous work in media bias detection has proposed augmentation techniques to exploit this fact. Despite their success, we observe that these techniques introduce noise by over-generalizing bias context boundaries, which hinders performance. To alleviate this issue, we propose techniques to more carefully search for context using a bias-sensitive, target-aware approach for data augmentation. Comprehensive experiments on the well-known BASIL dataset show that when combined with pre-trained models such as BERT, our augmentation techniques lead to state-of-the-art results. Our approach outperforms significantly, obtaining an F1-score of 58.15 over state-of-the-art bias detection task.

Zero-shot Triplet Extraction by Template Infilling

Bosung Kim, Hayate Iso, Nikita Bhutani, Estevam Hruschka, Ndapa Nakashole, and Tom Mitchell

The task of triplet extraction aims to extract pairs of entities and their corresponding relations from unstructured text. Most existing methods train an extraction model on training data involving specific target relations, and are incapable of extracting new relations that were not observed at training time. Generalizing the model to unseen relations typically requires fine-

tuning on synthetic training data which is often noisy and unreliable. We show that by reducing triplet extraction to a template infilling task over a pre-trained language model (LM), we can equip the extraction model with zero-shot learning capabilities and eliminate the need for additional training data. We propose a novel framework, ZETT (Zero-shot Triplet extraction by Template infilling), that aligns the task objective to the pre-training objective of generative transformers to generalize to unseen relations. Experiments on FewRel and Wiki-ZSL datasets demonstrate that ZETT shows consistent and stable performance, outperforming previous state-of-the-art methods, even when using automatically generated templates.

Rethinking the Role of Entity Type in Relation Classification
Xiang Dai, Sarvnaz Karimi, and Stephen Wan

Relation Classification (RC)—the task of identifying the relation between a pair of target entities—is a fundamental sub-task of information extraction. RC models built on top of entity information are prevalent, with different variants using entity information, especially entity type information, differently. However, RC models are often benchmarked on datasets that human annotators provide near-perfect entity information, and, state-of-the-art results are reported using gold entity type information. We believe there is a need to understand how the effectiveness of RC models is affected by the correctness of entity type information because in practice this information is provided by imperfect entity recognition models. Our results on six datasets across four domains show that although using gold entity type improves the effectiveness of RC models, incorrect entity types may cause large effectiveness drops on some (but not all) datasets. We propose using Pointwise Mutual Information (PMI) to identify datasets on which RC models may be negatively impacted by incorrect entity type information.

FollowupQG: Towards information-seeking follow-up question generation
Yan Meng, Liangming Pan, Yixin Cao, and Min-Yen Kan

Humans ask follow-up questions driven by curiosity, which reflects a creative human cognitive process. We introduce the task of real-world information-seeking follow-up question generation (FQG), which aims to generate follow-up questions seeking a more in-depth understanding of an initial question and answer. We construct FollowupQG, a dataset of over 3K real-world (initial question, answer, follow-up question) tuples collected from a Reddit forum providing layman-friendly explanations for open-ended questions. In contrast to existing datasets, questions in FollowupQG use more diverse pragmatic strategies to seek information, and they also show higher-order cognitive skills (such as applying and relating). We evaluate current question generation models on their efficacy for generating follow-up questions, exploring how to generate specific types of follow-up questions based on step-by-step demonstrations. Our results validate FollowupQG as a challenging benchmark, as model-generated questions are adequate but far from human-raised questions in terms of informativeness and complexity.

Do the Benefits of Joint Models for Relation Extraction Extend to Document-level Tasks?
Pratik Saini, Tapas Nayak, and Indrajit Bhattacharya

Two distinct approaches have been proposed for relational triple extraction - pipeline and joint. Joint models, which capture interactions across triples, are the more recent development, and have been shown to outperform pipeline models for sentence-level extraction tasks. Document-level extraction is a more challenging setting where interactions across triples can be long-range, and individual triples can also span across sentences. Joint models have not been applied for document-level tasks so far. In this paper, we benchmark state-of-the-art pipeline and joint extraction models on sentence-level as well as document-level datasets. Our experiments show that while joint models outperform pipeline models significantly for sentence-level extraction, their performance drops sharply below that of pipeline models for the document-level dataset.

Poster Session – 15:30-17:30

Foyer

Conversation Style Transfer using Few-Shot Learning
Shamik Roy, Raphael Shu, Nikolaos Pappas, Elman Mansimov, Yi Zhang, Saab Mansour, and Dan Roth

Conventional text style transfer approaches focus on sentence-level style transfer without considering contextual information, and the style is described with attributes (e.g., formality). When applying style transfer in conversations such as task-oriented dialogues, existing approaches suffer from these limitations as context can play an important role and the style attributes are often difficult to define in conversations. In this paper, we introduce conversation style transfer as a few-shot learning problem, where the model learns to perform style transfer by observing only a few example dialogues in the target style. We propose a novel in-context learning approach to solve the task with style-free dialogues as a pivot. Human evaluation shows that by incorporating multi-turn context, the model is able to match the target style while having better appropriateness and semantic correctness compared to utterance/sentence-level style transfer. Additionally, we show that conversation style transfer can also benefit downstream tasks. For example, in multi-domain intent classification tasks, the F1 scores improve after transferring the style of training data to match the style of the test data.

PICK: Polished & Informed Candidate Scoring for Knowledge-Grounded Dialogue Systems
Bryan Wilie, Yan Xu, Willy Chung, Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung

Grounding dialogue response generation on external knowledge is proposed to produce informative and engaging responses. However, current knowledge-grounded dialogue (KGD) systems often fail to align the resulting responses with human-preferred qualities due to several issues like hallucination, unfaithfulness, and the lack of coherence. To address these challenges and driven by these observations, we propose Polished & Informed Candidate Scoring (PICK), a generation re-scoring framework that empowers models to generate faithful and relevant responses without requiring additional labeled data or model tuning. Through comprehensive automatic and human evaluations, we demonstrate the effectiveness of PICK in generating responses that are more faithful while keeping them relevant to the dialogue history. Further, we investigate the performance disparity between PICK and the theoretical upper bound re-ranking approach to provide insights for future research.

Controllable Discovery of Intents: Incremental Deep Clustering Using Semi-Supervised Contrastive Learning
Mrinal Rawat, Hithesh Sankararaman, and Victor Barres
Deriving value from a conversational AI system depends on the capacity of a user to translate the prior knowledge into a configuration. In most cases, discovering the set of relevant turn-level speaker intents is often one of the key steps. Purely unsupervised algorithms provide a natural way to tackle discovery problems but make it difficult to incorporate constraints and only offer very limited control over the outcomes. Previous work has shown that semi-supervised (deep) clustering techniques can allow the system to incorporate prior knowledge and constraints in the intent discovery process. But they did not address how to allow for control through human feedback. In our Controllable Discovery of Intents (CDI) framework domain and prior knowledge are incorporated using a sequence of unsupervised contrastive learning on unlabeled data followed by fine-tuning on partially labeled data, and finally iterative refinement of clustering and representations through repeated clustering and pseudo-label fine-tuning. In addition, we draw from continual learning literature and use learning-without-forgetting to prevent catastrophic forgetting across those training stages. Finally, we show how this deep-clustering process can become part of an incremental discovery strategy with human-in-the-loop. We report results on both CLINC and BANKING datasets. CDI outperforms previous works by a significant margin: 10.26% and 11.72% respectively.

Phylogeny-Inspired Soft Prompts For Data-to-Text Generation in Low-Resource Languages
William Soto Martinez, Yannick Parmentier, and Claire Gardent

Most work on verbalising Knowledge-Graphs (KG) has focused on high-resource languages such as English, Russian, Czech or Arabic. In this paper, we focus on KG-to-Text generation where the output text is in Breton, Irish or Welsh. To overcome the small size of the parallel training data, we combine the strengths of a multilingual language model with denoising fine-tuning on monolingual data and Soft Prompt fine-tuning on a small quantity of KG/text data. We furthermore structure the soft prompt into multiple sub-prompts designed to capture the similarities and differences between English, Knowledge graphs and the three target languages. Our experiments show that our approach outperforms strong baselines and that all sub-prompts contribute to performance.

Retrieval Augmented Generation with Rich Answer Encoding
Wenyu Huang, Mirella Lapata, Pavlos Vougiouklis, Nikos Papasarantopoulos, and Jeff Pan

Knowledge-intensive generation tasks like generative question answering require models to retrieve appropriate passages from external knowledge sources to support answer generation. The generation quality relies heavily on the retrieved passages, which serve as contextual information. State-of-the-art Retrieval Augmented Generation models with marginalized output dominate this area but focus too much on label-relevant passages, rather than question-relevant passages and answers. This work addresses this issue by incorporating rich answer encoding through Dense Knowledge Similarity (DKS) and Retriever as Answer Classifier (RAC). We demonstrate the advantages of our proposed approach in open domain question answering (MSMARCO) and conversation (Wizard of Wikipedia) datasets, reporting both generation and retrieval metrics. In the MSMARCO development set, our best model achieves 12.1% relative improvement1 on Recall@1 and 4.5% relative improvement on BLEU-4 compared to the baseline model. In the KILT-WoW leaderboard, our best model achieves 8.9% relative improvement on R-Precision and 13.3% relative improvement on KILT-RL compared to the baseline model. Our codes and models are available at https://github.com/hwy9855/rag-ae.

Reimagining Complaint Analysis: Adopting Seq2Path for a Generative Text-to-Text Framework
Apoorva Singh, Raghav Jain, and Sriparna Saha

The escalating volume and frequency of social media complaints necessitate robust automated complaint analysis techniques. Much of the existing body of research in this area has been devoted to two primary aspects: identifying complaint-specific content amidst other non-complaint communications, and predicting the severity of a complaint, which involves classifying complaints into different severity levels based on the anticipated resolution from the complainant's perspective. These automated analysis tools equip companies with the means to effectively manage complaints and generate suitable responses. In our study, we present a unified generative approach for complaint detection, transforming the multitask learning problem into a text-to-text generation task. As part of our training strategy, we adopt the Seq2Path training paradigm that conceptualizes the outcome as a tree structure as opposed to a traditional sequence. This innovative approach tackles the drawbacks of conventional sequences, such as the lack of order among the outputs, yielding a more coherent and structured output. Our model's effectiveness is assessed against the benchmark *Complaints* dataset, highlighting its superior performance across diverse evaluation metrics when compared with state-of-the-art models and other baselines.

A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity
Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung
This paper proposes a framework for quantitatively evaluating interactive LLMs such as ChatGPT using publicly available data sets, using 23 data sets covering 8 different common NLP application tasks. We extensively evaluate the multitask, multilingual and multi-modal aspects of ChatGPT based on these data sets and a newly designed multimodal dataset. We find that ChatGPT outperforms LLMs with zero-shot learning on most tasks and even outperforms fine-tuned models on some tasks. We find that it is better at understanding non-Latin script languages than generating them. It is able to generate multimodal content from textual prompts, via an intermediate code generation step. Moreover, we find that ChatGPT is 63.41% accurate on average in 10 different reasoning categories under logical reasoning, non-textual reasoning, and commonsense reasoning, hence making it an unreliable reasoner. ChatGPT suffers from hallucination problems like other LLMs.Finally, the interactive feature of ChatGPT enables human collaboration with the underlying LLM to improve its performance, i.e, 8% ROUGE-1 on summarization and 2% ChrF++ on machine translation, in a multi-turn "prompt engineering" fashion. We also release a codebase for evaluation set extraction.

KoBigBird-large: Transformation of Transformer for Korean Language Understanding
Kisu Yang

This work presents KoBigBird-large, a large size of Korean BigBird that achieves state-of-the-art performance and allows long sequence processing for Korean language understanding. Without further pretraining, we only transform the architecture and extend the positional encoding with our proposed Tapered Absolute Positional Encoding Representations (TAPER). In experiments, KoBigBird-large shows state-of-the-art overall performance on Korean language understanding benchmarks and the best performance on document classification and question answering tasks for longer sequences against the competitive baseline models. We publicly release our model here.

Human-Like Distractor Response in Vision-Language Model
Xiaonan Xu and Haoshuo Chen

Previous studies exploring the human-like capabilities of machine-learning models have primarily focused on pure language models. Limited attention has been given to investigating whether models exhibit human-like behavior when performing tasks that require the integration of visual and language information. In this study, we investigate the impact of tags of semantic, phonological, and bilingual features on the visual question answering task performance of an unsupervised model. Our findings reveal its similarities with the influence of distractors in the picture-naming task (known as the picture-word-interference paradigm) observed in human experiments: 1) Semantically-related tags have a more negative effect on task performance compared to unrelated tags, indicating a more robust competition between visual and tag information which are semantically closer to each other when generating an answer. 2) Even presenting a partial section (wordpiece) of the originally detected tag significantly improves task performance, with the portion that plays a lesser role in determining the overall meaning of the original tag leading to a more pronounced improvement. 3) Tags in two languages that refer to the same meaning exhibit a symmetrical-like effect on performance in balanced bilingual models. Datasets and code of this project are released at https://github.com/NLPbelllabs/PWI.

Linguistic Productivity: the Case of Determiners in English
Raquel G. Alhama, Ruthe Foushee, Daniel Byrne, Allyson Ettinger, Susan Goldin-Meadow, and Afra Alishahi
Having heard "a pimwit", English-speakers assume that "the pimwit" is also possible. This type of productivity is attributed to syntactic categories such as NOUN and DETERMINER, but the key question is how do humans become endowed with these categories in the first place. We propose a novel approach that combines corpus analysis with computational modeling to analyze the productivity of DETERMINER+NOUN constructions in child-produced utterances. Our experiments on two corpora of child-adult interactions using two different methods of quantifying linguistic productivity show that children do not display productivity at early stages. Using a model trained on child-directed utterances, we simulate children's developmental trajectory with great precision, suggesting that the emergence of productivity in human language can be explained without the need to postulate a priori access to syntactic categories.

Improving Neural Machine Translation with Offline Evaluations
Minkyung Park and Byung-Jun Lee

Reinforcement learning (RL) offers a principled framework for optimizing a given reward function and has been applied to a neural machine translation (NMT) problem to maximize arbitrary task metrics. However, previously adopted RL algorithms for NMT (e.g., policy gradient) are generally slow as they require online data collection, and limits the algorithm's applicability to specific reward functions that can be evaluated online. In this paper, we present an offline RL algorithm called CER (conservative expectile regression). Despite the demanding nature of offline RL tasks, which are even more difficult with large models, this algorithm is capable to learn stably by explicitly exploiting the properties of NMT's RL formulation, such as the deterministic transition function. We analyze and discuss the design choices of CER, and demonstrate in the experiments that the proposed method outperforms its competitors for offline reward optimization in NMT.

Model-based Subsampling for Knowledge Graph Completion
Xincan Feng, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe

Subsampling is effective in Knowledge Graph Embedding (KGE) for reducing overfitting caused by the sparsity in Knowledge Graph (KG) datasets. However, current subsampling approaches consider only frequencies of queries that consist of entities and their relations. Thus, the existing subsampling potentially underestimates the appearance probabilities of infrequent queries even if the frequencies of their entities or relations are high. To address this problem, we propose Model-based Subsampling (MBS) and Mixed Subsampling (MIX) to estimate their appearance probabilities through predictions of KGE models. Evaluation results on datasets FB15k-237, WN18RR, and YAGO3-10 showed that our proposed subsampling methods actually improved the KG completion performances for popular KGE models, RotatE, TransE, HAKE, ComplEx, and DistMult.

Informative Evidence-guided Prompt-based Fine-tuning for English-Korean Critical Error Detection

DaHyun Jung, Sugyeong Eo, Chanjun Park, Hyeonseok Moon, Jaehyung Seo, and Heuiseok Lim

Critical error detection (CED) aims to identify the presence of catastrophic meaning distortion in machine translation. Fatal errors require significant attention because of their potential to cause personal or societal harm. The CED for Korean, an agglutinative language, is particularly highlighted, as minor variations in morphemes often bring substantial shifts in semantic interpretation. However, research on Korean is still underexplored and has room for improvement. In this study, we conduct the first investigation of CED for English-Korean to the best of our knowledge. We adopt prompt-based fine-tuning and propose various informative evidence to incorporate into the input prompt. Subsequently, we perform comprehensive verification and analysis to identify the most helpful guidance for detecting critical errors. The experimental results show that prompt-based fine-tuning with informative evidence outperforms standard fine-tuning by a large margin, demonstrating its remarkable effectiveness in English-Korean CED.

Interactive-Chain-Prompting: Ambiguity Resolution for Crosslingual Conditional Generation with Interaction

Jonathan Pilault, Xavier Garcia, Arthur Bražinskas, and Orhan Firat

Crosslingual conditional generation (e.g., machine translation) has long enjoyed the benefits of scaling. Nonetheless, there are still issues that scale alone may not overcome. For instance, in the absence of additional context, a source query in one language may yield several translation options in another language. Only one translation could be acceptable however, depending on the translator's preferences. Choosing the incorrect option may significantly affect translation usefulness and quality. We propose a novel method *interactive-chain prompting* (InterCPt) — a series of question, answering and generation intermediate steps between a *Translator* model and a *User* model — that reduces translations into a list of subproblems addressing ambiguities and then resolving such subproblems before producing the final translated text. To check ambiguity resolution capabilities and evaluate translation quality, we create a dataset exhibiting different linguistic phenomena which lead to ambiguities for four languages. To encourage further exploration in this direction, we release all datasets publicly. Using eight interactions as exemplars, InterCPt consistently surpasses other prompt-based methods with direct access to background information to resolve ambiguities.

Prover: Generating Intermediate Steps for NLI with Commonsense Knowledge Retrieval and Next-Step Prediction

Deepanway Ghosal, Somak Aditya, and Monojit Choudhury

The Natural Language Inference (NLI) task often requires reasoning over multiple steps to reach the conclusion. While the necessity of generating such intermediate steps (instead of a summary explanation) has gained popular support, it is unclear how to generate such steps without complete end-to-end supervision and how such generated steps can be further utilized. In this work, we train and enhance a sequence-to-sequence next-step prediction model with external commonsense knowledge and search to generate intermediate steps with limited next-step supervision. We show the correctness of such generated steps through automated and human verification, on MNLI and MED datasets (and discuss the limitations through qualitative examples). We show that such generated steps can help improve end-to-end NLI task performance using simple data augmentation strategies. Using a CheckList dataset for NLI, we also explore the effect of augmentation on specific reasoning types.

Analyzing and Predicting Persistence of News Tweets

Maggie Liu, Jing Wang, and Daniel Preotiuc-Pietro

News is read and consumed differently based on its topic and timeliness to the reader. Some stories attract readers immediately after they are published, while others capture readership consistently over multiple days after their publication, regardless of their overall popularity. This paper studies this less explored facet of news story consumption, which we name persistence, operationalized as the time for a story to reach a certain percent of its total interest. In particular, we study persistence though a novel, publicly available data set of news tweets from 353 news outlets. We perform an extensive linguistic analysis of news persistence in social media to uncover the underlying topical and stylistic cues that impact short- or long-term interest in a story. We train several models for predicting news persistence that achieve predictive performance of up to 0.353 Spearman correlation when extrapolating to tweets from days unseen in training and retain significant predictive performance even on tweets from accounts unseen in training. The ability to predict news persistence can be useful in several practical applications that drive news and social media consumption including alerting, search ranking or recommendations.

FiRo: Finite-context Indexing of Restricted Output Space for NLP Models Facing Noisy Input
Minh Nguyen and Nancy Chen

NLP models excel on tasks with clean inputs, but are less accurate with noisy inputs. In particular, character-level noise such as human-written typos and adversarially-engineered realistic-looking misspellings often appears in text and can easily trip up NLP models. Prior solutions to address character-level noise often alter the content of the inputs (low fidelity), thus inadvertently lowering model accuracy on clean inputs. We proposed FiRo, an approach to boost NLP model performance on noisy inputs without sacrificing performance on clean inputs. FiRo sanitizes the input text while preserving its fidelity by inferring the noise-free form for each token in the input. FiRo uses finite-context aggregation to obtain contextual embeddings which is then used to find the noise-free form within a restricted output space. The output space is restricted to a small cluster of probable candidates in order to predict the noise-free tokens more accurately. Although the clusters are small, FiRo's effective vocabulary (union of all clusters) can be scaled up to better preserve the input content. Experimental results show NLP models that use FiRo outperforming baselines on six classification tasks and one sequence labeling task at various degrees of noise.

DisCGen: A Framework for Discourse-Informed Counterspeech Generation
Sabit Hassan and Malihe Alikhani

Counterspeech can be an effective method for battling hateful content on social media. Automated counterspeech generation can aid in this process. Generated counterspeech, however, can be viable only when grounded in the context of topic, audience and sensitivity as these factors influence both the efficacy and appropriateness. In this work, we propose a novel framework based on theories of discourse to model inferential links of the context. Within this framework, we propose: i) a taxonomy of counterspeech derived from discourse frameworks, and ii) discourse-informed prompting strategies for generating contextually-grounded counterspeech. To construct and validate this framework, we present a process for collecting an in-the-wild dataset of counterspeech from Reddit. Using this process, we manually annotate a dataset of 3.9k Reddit comment pairs for the presence of hatespeech and counterspeech. The positive pairs are annotated for 10 classes in our proposed taxonomy. We annotate these pairs with style-transferred counterparts to remove offensiveness and first-person references. We show that by using our dataset and framework, large language models can generate contextually-grounded counterspeech informed by theories of discourse. According to our human evaluation, our approaches can act as a safeguard against critical failures of discourse-agnostic models.

We Need to Talk About Classification Evaluation Metrics in NLP
Peter Vickers, Loic Barrault, Emilio Monti, and Nikolaos Aletras

In Natural Language Processing (NLP) classification tasks such as topic categorisation and sentiment analysis, model generalizability is generally measured with standard metrics such as Accuracy, F-Measure, or AUC-ROC. The diversity of metrics, and the arbitrariness of their application suggest that there is no agreement within NLP on a single best metric to use. This lack suggests there has not been sufficient examination of the underlying heuristics which each metric encodes. To address this we compare several standard classification metrics with more 'exotic' metrics and demonstrate that a random-guess normalised Informedness metric is a parsimonious baseline for task performance. To show how important the choice of metric is, we perform extensive experiments on a wide range of NLP tasks including a synthetic scenario, natural language understanding, question answering and machine translation. Across these tasks we use a superset of metrics to rank models and find that Informedness best captures the ideal model characteristics. Finally, we release a Python implementation of Informedness following the SciKitLearn classifier format.

A Review of Datasets for Aspect-based Sentiment Analysis
Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio

Aspect-based sentiment analysis (ABSA) is a natural language processing problem that analyzes user-generated reviews to determine a) the target entity being reviewed, b) the high-level aspect to which it belongs, and c) the sentiment expressed toward the targets and the aspects. Numerous yet scattered corpora for ABSA make it difficult for researchers to identify corpora best suited for a specific ABSA subtask quickly. This study presents a database of corpora that can be used to train and evaluate autonomous ABSA systems. Additionally, we provide an overview of the major corpora for ABSA and its subtasks and highlight several features that researchers should consider when selecting a corpus. Finally, we discuss the advantages and disadvantages of existing dataset collection approaches and make recommendations for future corpora creation. This survey examines 98 publicly available ABSA datasets covering over 25 domains, including 77 English and 21 other languages datasets (https://github.com/RiTUAL-UH/ABSA-Datasets-Info).

Uncertainty Estimation for Debiased Models: Does Fairness Hurt Reliability?
Gleb Kuzmin, Artem Vazhentsev, Artem Shelmanov, Xudong Han, Simon Suster, Maxim Panov, Alexander Panchenko, and Timothy Baldwin

When deploying a machine learning model, one should aim not only to optimize performance metrics such as accuracy but also care about model fairness and reliability. Fairness means that the model is prevented from learning spurious correlations between a target variable and socio-economic attributes, and is generally achieved by applying debiasing techniques. Model reliability stems from the ability to determine whether we can trust model predictions for the given data. This can be achieved using uncertainty estimation (UE) methods. Debiasing and UE techniques potentially interfere with each other, raising the question of whether we can achieve both reliability and fairness at the same time. This work aims to answer this question empirically based on an extensive series of experiments combining state-of-the-art UE and debiasing methods, and examining the impact on model performance, fairness, and reliability.

Benchmarking Procedural Language Understanding for Low-Resource Languages: A Case Study on Turkish
Arda Uzunoglu and Gözde Şahin
Understanding procedural natural language (e.g., step-by-step instructions) is a crucial step to execution and planning. However, while there are ample corpora and downstream tasks available in English, the field lacks such resources for most languages. To address this gap, we conduct a case study on Turkish procedural texts. We first expand the number of tutorials in Turkish wikiHow from 2,000 to 52,000 using automated translation tools, where the translation quality and loyalty to the original meaning are validated by a team of experts on a random set. Then, we generate several downstream tasks on the corpus, such as linking actions, goal inference, and summarization. To tackle these tasks, we implement strong baseline models via fine-tuning large language-specific models such as TR-BART and BERTurk, as well as multilingual models such as mBART, mT5, and XLM. We find that language-specific models consistently outperform their multilingual models by a significant margin across most procedural language understanding (PLU) tasks. We release our corpus, downstream tasks and the baseline models with https://github.com/GGLAB-KU/turkish-plu.

Semi-supervised News Discourse Profiling with Contrastive Learning
Ming Li and Ruihong Huang

News Discourse Profiling seeks to scrutinize the event-related role of each sentence in a news article and has been proven useful across various downstream applications. Specifically, within the context of a given news discourse, each sentence is assigned to a pre-defined category contingent upon its depiction of the news event structure. However, existing approaches suffer from an inadequacy of available human-annotated data, due to the laborious and time-intensive nature of generating discourse-level annotations. In this paper, we present a novel approach, denoted as Intra-document Contrastive Learning with Distillation (ICLD), for addressing the news discourse profiling task, capitalizing on its unique structural characteristics. Notably, we are the first to apply a semi-supervised methodology within this task paradigm, and evaluation demonstrates the effectiveness of the presented approach.

J-Guard: Journalism Guided Adversarially Robust Detection of AI-generated News
Tharindu Kumarage, Amrita Bhattacharjee, Djordje Padejski, Kristy Roschke, Dan Gillmor, Scott Ruston, Huan Liu, and Joshua Garland
The rapid proliferation of AI-generated text online is profoundly reshaping the information landscape. Among various types of AI-generated text, AI-generated news presents a significant threat as it can be a prominent source of misinformation online. While several recent efforts have focused on detecting AI-generated text in general, these methods require enhanced reliability, given concerns about their vulnerability to simple adversarial attacks. Furthermore, due to the eccentricities of news writing, applying these detection methods for AI-generated news can produce false positives, potentially damaging the reputation of news organizations. To address these challenges, we leverage the expertise of an interdisciplinary team to develop a framework, J-Guard, capable of steering existing supervised AI text detectors for detecting AI-generated news while boosting adversarial robustness. By incorporating stylistic cues inspired by the unique journalistic attributes, J-Guard effectively distinguishes between real-world journalism and AI-generated news articles. Our experiments on news articles generated by a vast array of AI models, including ChatGPT (GPT3.5), demonstrate the effectiveness of J-Guard in enhancing detection capabilities while maintaining an average performance decrease of as low as 7% when faced with adversarial attacks.

Learning to Predict Concept Ordering for Common Sense Generation
Tianhui Zhang, Danushka Bollegala, and Bei Peng

Prior work has shown that the ordering in which concepts are shown to a commonsense generator plays an important role, affecting the quality of the generated sentence. However, it remains a challenge to determine the optimal ordering of a given set of concepts such that a natural sentence covering all the concepts could be generated from a pretrained generator. To understand the relationship between the ordering of the input concepts and the quality of the generated sentences, we conduct a systematic study considering multiple language models (LMs) and concept ordering strategies. We find that BART-large model consistently outperforms all other LMs considered in this study when fine-tuned using the ordering of concepts as they appear in CommonGen training data as measured using multiple evaluation metrics. Moreover, the larger GPT3-based large language models (LLMs) variants do not necessarily outperform much smaller LMs on this task, even when fine-tuned on task-specific training data. Interestingly, human annotators significantly reorder input concept sets when manually writing sentences covering those concepts, and this ordering provides the best sentence generations independently of the LM used for the generation, outperforming a probabilistic concept ordering baseline.

Enhancing Volatility Forecasting in Financial Markets: A General Numeral Attachment Dataset for Understanding Earnings Calls
Ming-Xuan Shi, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen
Volatility, a crucial statistical measure in the financial market, serves as an indicator of financial instrument risk. Accurate volatility capture aids in predicting stock movements and is valuable in derivative trading, such as options trading. While recent research focuses on volatility forecasting using earnings call transcriptions, most approaches rely on end-to-end models that directly process textual or vocal data. However, limited efforts have been made to simulate the reading and comprehension processes of financial professionals, thereby enhancing the capabilities of language models. To address this gap, we propose a general numeral attachment dataset designed to train language models to understand earnings calls with the expertise of professionals. Additionally, we introduce a pre-training process that improves the semantic understanding of earnings calls. Experimental results demonstrate that our pre-trained language model enhances the accuracy of 3-day

volatility forecasting.

All Labels Together: Low-shot Intent Detection with an Efficient Label Semantic Encoding Paradigm
Jiangshu Du, Congying Xia, Wenpeng Yin, Tingting Liang, and Philip Yu
In intent detection tasks, leveraging meaningful semantic information from intent labels can be particularly beneficial for few-shot scenarios. However, existing few-shot intent detection methods either ignore the intent labels, (e.g. treating intents as indices) or do not fully utilize this information (e.g. only using part of the intent labels). In this work, we present an end-to-end One-to-All system that enables the comparison of an input utterance with all label candidates. The system can then fully utilize label semantics in this way. Experiments on three few-shot intent detection tasks demonstrate that One-to-All is especially effective when the training resource is extremely scarce, achieving state-of-the-art performance in 1-, 3- and 5-shot settings. Moreover, we present a novel pretraining strategy for our model that utilizes indirect supervision from paraphrasing, enabling zero-shot cross-domain generalization on intent detection tasks. Our code is at https://github.com/jiangshdd/AllLablesTogether.

Enhancing Open-Domain Table Question Answering via Syntax- and Structure-aware Dense Retrieval
Nengzheng Jin, Dongfang Li, Junying Chen, Joanna Siebert, and Qingcai Chen
Open-domain table question answering aims to provide answers to a question by retrieving and extracting information from a large collection of tables. Existing studies of open-domain table QA either directly adopt text retrieval methods or consider the table structure only in the encoding layer for table retrieval, which may cause syntactical and structural information loss during table scoring. To address this issue, we propose a syntax- and structure-aware retrieval method for the open-domain table QA task. It provides syntactical representations for the question and uses the structural header and value representations for the tables to avoid the loss of fine-grained syntactical and structural information. Then, a syntactical-to-structural aggregator is used to obtain the matching score between the question and a candidate table by mimicking the human retrieval process. Experimental results show that our method achieves the state-of-the-art on the NQ-tables dataset and overwhelms strong baselines on a newly curated open-domain Text-to-SQL dataset.

On the Challenges of Fully Incremental Neural Dependency Parsing
Ana Ezquerro, Carlos Gómez-Rodríguez, and David Vilares

Since the popularization of BiLSTMs and Transformer-based bidirectional encoders, state-of-the-art syntactic parsers have lacked incrementality, requiring access to the whole sentence and deviating from human language processing. This paper explores whether fully incremental dependency parsing with modern architectures can be competitive. We build parsers combining strictly left-to-right neural encoders with fully incremental sequence-labeling and transition-based decoders. The results show that fully incremental parsing with modern architectures considerably lags behind bidirectional parsing, noting the challenges of psycholinguistically plausible parsing.

Who Are All The Stochastic Parrots Imitating? They Should Tell Us!
Sagi Shaier, Lawrence Hunter, and Katharina Kann

Both standalone language models (LMs) as well as LMs within downstream-task systems have been shown to generate statements which are factually untrue. This problem is especially severe for low-resource languages, where training data is scarce and of worse quality than for high-resource languages. In this opinion piece, we argue that LMs in their current state will never be fully trustworthy in critical settings and suggest a possible novel strategy to handle this issue: by building LMs such that can cite their sources – i.e., point a user to the parts of their training data that back up their outputs. We first discuss which current NLP tasks would or would not benefit from such models. We then highlight the expected benefits such models would bring, e.g., quick verifiability of statements. We end by outlining the individual tasks that would need to be solved on the way to developing LMs with the ability to cite. We hope to start a discussion about the field's current approach to building LMs, especially for low-resource languages, and the role of the training data in explaining model generations.

Issues Surrounding the Use of ChatGPT in Similar Languages: The Case of Malay and Indonesian
Hiroki Nomoto

We report a problem that one faces when using ChatGPT in similar languages, taking Malay and Indonesian as examples: ChatGPT often responds to prompts in Malay (the language with fewer speakers) in Indonesian (the language with more speakers). We examined ChatGPT's identification (LangID) ability to find out whether this language choice problem arises from LangID errors. The results show that LangID errors alone cannot explain the problem's severity. By comparing the patterns of responses to Malay prompts and those to Javanese prompts, we conclude that the problem happens mainly because ChatGPT does not treat Malay and Indonesian equally as distinct languages. Rather, it behaves as if Malay were a non-standard variety of Indonesian. We also discuss social issues the language choice problem causes and possible solutions to them.

Can You Translate for Me? Code-Switched Machine Translation with Large Language Models
Jyotsana Khatri, Vivek Srivastava, and Lovekesh Vig

Large language models (LLMs) have shown remarkable performance on a variety of multilingual NLP tasks. Code-switching is one of the most convenient styles of communication in multilingual communities. It is known to present

several challenges to the existing language models and task-specific models. In this paper, we evaluate the capability of multilingual LLMs for the code-switched machine translation (CSMT) task in traditional and novel settings and present our insights. We observe that ChatGPT outperforms other LLMs and shows competitive performance to the supervised fine-tuned models. Though promising, ChatGPT shows major limitations, such as high gender bias, stereotypes, and factual inconsistencies. It further demands a multi-dimensional large-scale evaluation of the multilingual LLMs for code-switched languages.

Efficient Zero-Shot Cross-lingual Inference via Retrieval
Genta Winata, Lingjue Xie, Karthik Radhakrishnan, Yifan Gao, and Daniel Preotiuc-Pietro

Resources for building NLP applications, such as data and models, are usually only created and curated for a limited set of high resource languages. Thus, the ability to transfer knowledge to a new language is a key way in which to enable access to NLP technology for a wider population. This paper presents a framework to perform zero-shot inference in a target language by using cross-lingual retrieval from another language where limited annotated data for a comparable domain is available. Results on two large-scale multilingual datasets show that, in this setup, this framework improves over fine-tuning multilingual models or translating annotated data, and achieves results relatively close to fine-tuning the model on the target language directly. These results show that models can be transferred efficiently across languages for a given task and domain, even for languages not covered by multilingual model training approaches.

The Language Model, Resources, and Computational Pipelines for the Under-Resourced Iranian Turkic
Marzia Nouri, Mahsa Amani, Reihaneh Zohrabi, and Ehsaneddin Asgari
Iranian Azerbaijani is a dialect of the Azerbaijani language spoken by more than 16% of the population in Iran (>14 million). Unfortunately, a lack of computational resources is one of the factors that puts this language and its rich culture at risk of extinction. This work aims to create fundamental natural language processing (NLP) resources and pipelines for the processing and analysis of Iranian Azerbaijani introducing standard datasets and starter models for various NLP tasks such as language modeling, text classification, part-of-speech (POS) tagging, and machine translation. The proposed resources have been curated and preprocessed to facilitate the development of NLP models for Iranian Azerbaijani and provide a strong baseline for further research and development. This study is an example of bridging the gap in NLP for low-resource languages and promoting the advancement of language technologies in underrepresented languages. To the best of our knowledge, for the first time, this paper presents major infrastructures for the processing and analysis of Iranian Azerbaijani, with the ultimate goal of improving communication and information access for millions of individuals. Furthermore, our translation model's online demo is accessible at https://azeri.parsi.ai/.

Borderless Azerbaijani Processing: Linguistic Resources and a Transformer-based Approach for Azerbaijani Transliteration
Reihaneh Zohrabi, Mostafa Masumi, Omid Ghahroodi, Parham AbedAzad, Hamid Beigy, Mohammad Hossein Rohban, and Ehsaneddin Asgari
Recent advancements in neural language models have revolutionized natural language understanding. However, many languages still face the risk of being left behind without the benefits of such advancements, potentially leading to their extinction. One such language is Azerbaijani in Iran, which suffers from limited digital resources and a lack of alignment between spoken and written forms. In contrast, Azerbaijani in the Republic of Azerbaijan has seen more resources and is not considered as low-resource as its Iranian counterpart. In this context, our research focuses on the computational progress made in Iranian Azerbaijani language. We propose a transliteration model that leverages an Azerbaijani parallel dataset, effectively bridging the gap between the Latin and Persian scripts. By enabling seamless communication between these two scripts, our model facilitates cultural exchange and serves as a valuable tool for transfer learning. The effectiveness of our approach surpasses traditional rule-based methods, as evidenced by the significant improvements in performance metrics. We observe a minimum 15% increase in BLEU scores and a reduction of at least 1/3 in edit distance.

Exploring Methods for Cross-lingual Text Style Transfer: The Case of Text Detoxification
Daryna Dementieva, Daniil Moskovskiy, David Dale, and Alexander Panchenko

Text detoxification is the task of transferring the style of text from toxic to neutral. While there are approaches yielding promising results in monolingual setup, e.g., (Dale et al.,2021; Hallinan et al., 2022), cross-lingual transfer for this task remains a challenging open problem (Moskovskiy et al., 2022). In this work, we present a large-scale study of strategies for cross-lingual text detoxification – given a parallel detoxification corpus for one language; the goal is to transfer detoxification ability to another language for which we do not have such a corpus. Moreover, we are the first to explore a new task where text translation and detoxification are performed simultaneously, providing several strong baselines for this task. Finally, we introduce new automatic detoxification evaluation metrics with higher correlations with human judgments than previous benchmarks. We assess the most promising approaches also with manual markup, determining the answer for the best strategy to transfer the knowledge of text detoxification between languages.

PACT: Pretraining with Adversarial Contrastive Learning for Text Classification
Md Tawkat Islam Khondaker, Muhammad Abdul-Mageed, Laks Lakshmanan, and V.S.

We present PACT (Pretraining with Adversarial Contrastive Learning for Text Classification), a novel self-supervised framework for text classification. Instead of contrasting against in-batch negatives, a popular approach in the literature, PACT mines negatives closer to the anchor representation. PACT operates by endowing the standard pretraining mechanisms of

BERT with adversarial contrastive learning objectives, allowing for effective joint optimization of token- and sentence-level pretraining of the BERT model. Our experiments on 13 diverse datasets including token-level, single-sentence, and sentence-pair text classification tasks show that PACT achieves consistent improvements over SOTA baselines. We further show that PACT regularizes both token-level and sentence-level embedding spaces into more uniform representations, thereby alleviating the undesirable anisotropic phenomenon of language models.

VACASPATI: A Diverse Corpus of Bangla Literature
Pramit Bhattacharyya, Joydeep Mondal, Subhadip Maji, and Arnab Bhattacharya

Bangla (or Bengali) is the fifth most spoken language globally; yet, the state-of-the-art NLP in Bangla is lagging for even simple tasks such as lemmatization, POS tagging, etc. This is partly due to lack of a varied quality corpus. To alleviate this need, we build VACASPATI, a diverse corpus of Bangla literature. The literary works are collected from various websites; only those works that are publicly available without copyright violations or restrictions are collected. We believe that published literature captures the features of a language much better than newspapers, blogs or social media posts which tend to follow only a certain literary pattern and, therefore, miss out on language variety and vocabulary. Our corpus VACASPATI is varied from multiple aspects, including type of composition, topic, author, time, space, etc. It contains more than 11 million sentences and 115 million words. We have also built a word embedding model, VAC-FT, using FastText from VACASPATI as well as trained an Electra model, VAC-BERT, using the corpus. VAC-BERT has far fewer parameters and requires only a fraction of resources compared to other state-of-the-art transformer models and yet performs either better or similar on various downstream tasks. Similarly, VAC-FT outperforms other FastText-based models on multiple downstream tasks. We also demonstrate the efficacy of VACASPATI as a corpus by showing that similar models built from other corpora are not as effective. The models are available at https://bangla.iitk.ac.in/projects/vacaspati.html.

Are Machine Reading Comprehension Systems Robust to Context Paraphrasing?
Yulong Wu, Viktor Schlegel, and Riza Batista-Navarro

Investigating the behaviour of Machine Reading Comprehension (MRC) models under various types of test-time perturbations can shed light on the enhancement of their robustness and generalisation capability, despite the superhuman performance they have achieved on existing benchmark datasets. In this paper, we study the robustness of contemporary MRC systems to context paraphrasing, i.e., whether these models are still able to correctly answer the questions once the reading passages have been paraphrased. To this end, we systematically design a pipeline to semi-automatically generate perturbed MRC instances which ultimately lead to the creation of a paraphrased test set. We conduct experiments on this dataset with six state-of-the-art neural MRC models and we find that even the minimum performance drop of all these models exceeds 41%, whereas human performance remains high. Retraining models with augmented perturbed examples results in improved robustness, though the performance remains lower than on the original dataset. These results demonstrate that the existing high-performing MRC systems are still far away from real language understanding.

It's not only What You Say, It's also Who It's Said to: Counterfactual Analysis of Interactive Behavior in the Courtroom
Biaoyan Fang, Trevor Cohn, Timothy Baldwin, and Lea Frermann
To what extent do personal attributes affect the way we are spoken to? Answering this question requires the precise reproduction of a conversational context except for one personal attribute of interest, amounting to a classical, yet infeasible, causal inference problem. We present a method based on counterfactual analysis by manipulating speaker attributes in observational data. We present a case study of Advocate responses to Justices in debates in the Supreme Court of the United States. Specifically, we measure changes in politeness and coordination of Advocates when responding to (a) real Justices and (b) counterfactually-manipulated Justices, with responses generated with GPT2. We first validate our method, showing that GPT2-generated outputs capture coordination and politeness. Our results confirm a known impact of the attribute gender, and suggest a weaker effect of seniority on coordination.

Main Conference: Saturday, November 4

Session 3A – 10:30-12:00

Multilingual and Multimodal Analysis
Karangasem 1+2
Chair: Thamar Solorio

Only 5% Attention Is All You Need: Efficient Long-range Document-level Neural Machine Translation
Zihan Liu, Zewei Sun, Shanbo Cheng, Shujian Huang, and Mingxuan Wang

Document-level Neural Machine Translation (DocNMT) has been proven crucial for handling discourse phenomena by introducing document-level context information. One of the most important directions is to input the whole document directly to the standard Transformer model. In this case, efficiency becomes a critical concern due to the quadratic complexity of the attention module. Existing studies either focus on the encoder part, which cannot be deployed on sequence-to-sequence generation tasks, e.g., Machine Translation (MT), or suffer from a significant performance drop. In this work, we keep the translation performance while gaining 20% speed up by introducing extra selection layer based on lightweight attention that selects a small portion of tokens to be attended. It takes advantage of the original attention to ensure performance and dimension reduction to accelerate inference. Experimental results show that our method could achieve up to 95% sparsity (only 5% tokens attended) approximately, and save 93% computation cost on the attention module compared with the original Transformer, while maintaining the performance.

Examining Consistency of Visual Commonsense Reasoning based on Person Grounding

Huiju Kim, Youjin Kang, and SangKeun Lee

Given an image depicting multiple individuals, humans are capable of inferring each individual's emotions, intentions, and social norms based on commonsense understanding. However, a machine's ability of commonsense reasoning about distinct individuals in images remains underexplored. In this study, we examine the consistency of visual commonsense reasoning based on person grounding. We introduce a novel test dataset called Visual Commonsense Reasoning-Contrast Sets (VCR-CS) to evaluate whether models can reason about individual people in an image by changing the person tags in the questions and answers. We benchmark various vision-language models on VCR-CS and observe that they fail in consistent commonsense reasoning about different people in one image, showing a performance decrease of up to 31.5%. To mitigate such failures, we propose a multi-task learning framework called Personcentric groundIng eNhanced Tuning (PINT). Our framework enhances a model's ability to perform person-grounded commonsense reasoning by leveraging two novel person-centric pretraining tasks: Image Person-based Text Matching and Person-Masked Language Modeling. The experimental results revealed the effectiveness of PINT by showing the lowest performance degradation on VCR-CS and the improvements in consistency and sensitivity metrics. Our dataset and code are publicly available.

Exploring the Impact of Training Data Distribution and Subword Tokenization on Gender Bias in Machine Translation

Bar Iluz, Tomasz Limisiewicz, Gabriel Stanovsky, and David Mareček

We study the effect of tokenization on gender bias in machine translation, an aspect that has been largely overlooked in previous works. Specifically, we focus on the interactions between the frequency of gendered profession names in training data, their representation in the subword tokenizer's vocabulary, and gender bias. We observe that female and non-stereotypical gender inflections of profession names (e.g., Spanish "doctora" for "female doctor") tend to be split into multiple subword tokens. Our results indicate that the imbalance of gender forms in the model's training corpus is a major factor contributing to gender bias and has a greater impact than subword splitting. We show that analyzing subword splits provides good estimates of gender-form imbalance in the training data and can be used even when the corpus is not publicly available. We also demonstrate that fine-tuning just the token embedding layer can decrease the gap in gender prediction accuracy between female and male forms without impairing the translation quality.

Implicit Affordance Acquisition via Causal Action–Effect Modeling in the Video Domain

Hsiu-Yu Yang and Carina Silberer

Affordance knowledge is a fundamental aspect of commonsense knowledge. Recent findings indicate that world knowledge emerges through large-scale self-supervised pretraining, motivating our exploration of acquiring affordance knowledge from the visual domain. To this end, we augment an existing instructional video resource to create the new Causal Action–Effect (CAE) dataset and design two novel pretraining tasks—Masked Action Modeling (MAM) and Masked Effect Modeling (MEM)—promoting the acquisition of two affordance properties in models: behavior and entity equivalence, respectively. We empirically demonstrate the effectiveness of our proposed methods in learning affordance properties. Furthermore, we show that a model pretrained on both tasks outperforms a strong image-based visual–linguistic foundation model (FLAVA) as well as pure linguistic models on a zero-shot physical reasoning probing task.

A Multimodal Analysis of Influencer Content on Twitter

Danae Sánchez Villegas, Catalina Goanta, and Nikolaos Aletras

Influencer marketing involves a wide range of strategies in which brands collaborate with popular content creators (i.e., influencers) to leverage their reach, trust, and impact on their audience to promote and endorse products or services. Because followers of influencers are more likely to buy a product after receiving an authentic product endorsement rather than an explicit direct product promotion, the line between personal opinions and commercial content promotion is frequently blurred. This makes automatic detection of regulatory compliance breaches related to influencer advertising (e.g., misleading advertising or hidden sponsorships) particularly difficult. In this work, we (1) introduce a new Twitter (now X) dataset consisting of 15,998 influencer posts mapped into commercial and non-commercial categories for assisting in the automatic detection of commercial influencer content; (2) experiment with an extensive set of predictive models that combine text and visual information showing that our proposed cross-attention approach outperforms state-of-the-art multimodal models; and (3) conduct a thorough analysis of strengths and limitations of our models. We show that multimodal modeling is useful for identifying commercial posts, reducing the amount of false positives, and capturing relevant context that aids in the

discovery of undisclosed commercial posts.

Theia: Weakly Supervised Multimodal Event Extraction from Incomplete Data
Farhad Moghimifar, Fatemeh Shiri, Van Nguyen, Yuan-Fang Li, and Gholamreza Haffari

Event extraction from multimodal documents is an important yet under-explored problem. One challenge faced by this task is the scarcity of paired image-text datasets, making it difficult to fully exploit the strong representation power of multimodal language models. In this paper, we present Theia, an end-to-end multimodal event extraction framework that can be trained on incomplete data. Specifically, we couple a generation-based event extraction model with a customised image synthesizer that can generate images from text. Our model leverages capabilities of pre-trained vision-language models and can be trained on incomplete (i.e. text-only) data. Experimental results on existing multimodal datasets demonstrate the effectiveness of our approach for both synthesising missing data and extracting events over state-of-the-art approaches.

Session 3B – 10:30-12:00

Machine Learning and Model Interpretability

Negara

Chair: Zhuosheng Zhang

Emerging Challenges in Personalized Medicine: Assessing Demographic Effects on Biomedical Question Answering Systems
Sagi Shaier, Kevin Bennett, Lawrence Hunter, and Katharina Kann

State-of-the-art question answering (QA) models exhibit a variety of social biases (e.g., with respect to sex or race), generally explained by similar issues in their training data. However, what has been overlooked so far is that in the critical domain of biomedicine, any unjustified change in model output due to patient demographics is problematic: it results in the unfair treatment of patients. Selecting only questions on biomedical topics whose answers do not depend on ethnicity, sex, or sexual orientation, we ask the following research questions: (RQ1) Do the answers of QA models change when being provided with irrelevant demographic information? (RQ2) Does the answer of RQ1 differ between knowledge graph (KG)-grounded and text-based QA systems? We find that irrelevant demographic information change up to 15% of the answers of a KG-grounded system and up to 23% of the answers of a text-based system, including changes that affect accuracy. We conclude that unjustified answer changes caused by patient demographics are a frequent phenomenon, which raises fairness concerns and should be paid more attention to.

Faithful Chain-of-Thought Reasoning
Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch

While Chain-of-Thought (CoT) prompting boosts Language Models' (LM) performance on a gamut of complex reasoning tasks, the generated reasoning chain does not necessarily reflect how the model arrives at the answer (aka. faithfulness). We propose Faithful CoT, a reasoning framework involving two stages: Translation (Natural Language query $\rightarrow$ symbolic reasoning chain) and Problem Solving (reasoning chain $\rightarrow$ answer), using an LM and a deterministic solver respectively. This guarantees that the reasoning chain provides a faithful explanation of the final answer. Aside from interpretability, Faithful CoT also improves empirical performance: it outperforms standard CoT on 9 of 10 benchmarks from 4 diverse domains, with a relative accuracy gain of 6.3% on Math Word Problems (MWP), 3.4% on Planning, 5.5% on Multi-hop Question Answering (QA), and 21.4% on Relational Inference. Furthermore, with GPT-4 and Codex, it sets the new state-of-the-art few-shot performance on 7 datasets (with 95.0+ accuracy on 6 of them), showing a strong synergy between faithfulness and accuracy.

24-bit Languages
Yiran Wang, Taro Watanabe, Masao Utiyama, and Yuji Matsumoto

We propose a contrastive hashing method to compress and interpret the contextual representation of pre-trained language models into binary codes. Unlike previous work that generates token-level tags, our method narrows the representation bottleneck to codes with only 24 bits, retaining task-relevant information in a more interpretable and fine-grained format without sacrificing performance (in most cases). We provide experiments and discussions on various structured prediction tasks, such as part-of-speech tagging, named entity recognition, and constituency parsing, to demonstrate the effectiveness and interpretability of our method.

ConDA: Contrastive Domain Adaptation for AI-generated Text Detection
Amrita Bhattacharjee, Tharindu Kumarage, Raha Moraffah, and Huan Liu

Large language models (LLMs) are increasingly being used for generating text in a variety of use cases, including journalistic news articles. Given the potential malicious nature in which these LLMs can be used to generate disinformation at scale, it is important to build effective detectors for such AI-generated text. Given the surge in development of new LLMs, acquiring labeled training data for supervised detectors is a bottleneck. However, there might be plenty of unlabeled text data available, without information on which generator it came from. In this work we tackle this data problem, in detecting AI-generated news text, and frame the problem as an unsupervised domain adaptation task. Here the domains are the differ-

ent text generators, i.e. LLMs, and we assume we have access to only the labeled source data and unlabeled target data. We develop a Contrastive Domain Adaptation framework, called ConDA, that blends standard domain adaptation techniques with the representation power of contrastive learning to learn domain invariant representations that are effective for the final unsupervised detection task. Our experiments demonstrate the effectiveness of our framework, resulting in average performance gains of 31.7% from the best performing baselines, and within 0.8% margin of a fully supervised detector.

On a Benefit of Masked Language Model Pretraining: Robustness to Simplicity Bias
Ting-Rui Chiang

Despite the success of pretrained masked language models (MLM), why MLM pretraining is useful is still a question not fully answered. In this work we theoretically and empirically that MLM pretraining makes models robust to lexicon-level spurious features, partly answering the question. Our explanation is that MLM pretraining may alleviate problems brought by simplicity bias (Shah et al., 2020), which refers to the phenomenon that a deep model tends to rely excessively on simple features. In NLP tasks, those simple features could be token-level features whose spurious association with the label can be learned easily. We show that MLM pretraining makes learning from the context easier. Thus, pretrained models are less likely to rely excessively on a single token. We also explore the theoretical explanations of MLM's efficacy in causal settings. Compared with Wei et al. (2021), we achieve similar results with milder assumption. Finally, we close the gap between our theories and real-world practices by conducting experiments on real-world tasks.

Perplexity-Driven Case Encoding Needs Augmentation for CAPITALIZATION Robustness
Rohit Jain, Huda Khayrallah, Roman Grundkiewicz, and Marcin Junczys-Dowmunt

Subword segmentation methods are the predominant solution to vocab sparsity in NMT. However, they cannot currently handle capitalization well. We re-encode case to allow the perplexity-driven SPM unigram language model algorithm to learn how to segment capitalization. Since naturally occurring data accurately describes the prevalence of capitalization but underestimates the importance humans ascribe to capitalization robustness, we propose data augmentation to fill this gap. We demonstrate that our proposed method improves translation quality on ALL CAPS, lower cased, and Title Case, while maintaining quality on standard test sets. In contrast to prior work, our proposed method has minimal impact on decoding speed. We release our code: github.com/marian-nmt/sentencepiece.

Session 3C – 14:00-15:30

Semantics

Karangasem 1+2

Chair: Tim Baldwin

Smoothing Entailment Graphs with Language Models
Nick McKenna, Tianyi Li, Mark Johnson, and Mark Steedman

The diversity and Zipfian frequency distribution of natural language predicates in corpora leads to sparsity in Entailment Graphs (EGs) built by Open Relation Extraction (ORE). EGs are computationally efficient and explainable models of natural language inference, but as symbolic models, they fail if a novel premise or hypothesis vertex is missing at test-time. We present theory and methodology for overcoming such sparsity in symbolic models. First, we introduce a theory of optimal smoothing of EGs by constructing transitive chains. We then demonstrate an efficient, open-domain, and unsupervised smoothing method using an off-the-shelf Language Model to find approximations of missing premise predicates. This improves recall by 25.1 and 16.3 percentage points on two difficult directional entailment datasets, while raising average precision and maintaining model explainability. Further, in a QA task we show that EG smoothing is most useful for answering questions with lesser supporting text, where missing premise predicates are more costly. Finally, controlled experiments with WordNet confirm our theory and show that hypothesis smoothing is difficult, but possible in principle.

LexicoMatic: Automatic Creation of Multilingual Lexical-Semantic Dictionaries
Federico Martelli, Luigi Procopio, Edoardo Barba, and Roberto Navigli

Lexical-semantic resources such as wordnets and multilingual dictionaries often suffer from significant coverage issues, especially in languages other than English. While improving their coverage manually is a prohibitively expensive undertaking, current approaches to the automatic creation of such resources fail to investigate the latest advances achieved in relevant fields, such as cross-lingual annotation projection. In this work, we address these shortcomings and propose LexicoMatic, a novel resource-independent approach to the automatic construction and expansion of multilingual semantic dictionaries, in which we formulate the task as an annotation projection problem. In addition, we tackle the lack of a comprehensive multilingual evaluation framework and put forward a new entirely manually-curated benchmark featuring 9 languages. We evaluate LexicoMatic with an extensive array of experiments and demonstrate the effectiveness of our approach, achieving a new state of the art across all languages under consideration. We release our novel evaluation benchmark at: https://github.com/SapienzaNLP/lexicomatic.

One Sense per Translation
Bradley Hauer and Grzegorz Kondrak

Word sense disambiguation (WSD) is the task of determining the sense of a word in context. Translations have been used

in WSD as a source of knowledge, and even as a means of delimiting word senses. In this paper, we define three theoretical properties of the relationship between senses and translations, and argue that they constitute necessary conditions for using translations as sense inventories. The key property of One Sense per Translation (OSPT) provides a foundation for a translation-based WSD method. The results of an intrinsic evaluation experiment indicate that our method achieves a precision of approximately 93% compared to manual corpus annotations. Our extrinsic evaluation experiments demonstrate WSD improvements of up to 4.6% F1-score on difficult WSD datasets.

Self-Consistent Narrative Prompts on Abductive Natural Language Inference
Chunkit Chan, Xin Liu, Tsz Ho CHAN, Jiayang Cheng, Yangqiu Song, Ginny Wong, and Simon See

Abduction has long been seen as crucial for narrative comprehension and reasoning about everyday situations. The abductive natural language inference (αNLI) task has been proposed, and this narrative text-based task aims to infer the most plausible hypothesis from the candidates given two observations. However, the inter-sentential coherence and the model consistency have not been well exploited in the previous works on this task. In this work, we propose a prompt tuning model α-PACE, which takes self-consistency and inter-sentential coherence into consideration. Besides, we propose a general self-consistent framework that considers various narrative sequences (e.g., linear narrative and reverse chronology) for guiding the pre-trained language model in understanding the narrative context of input. We conduct extensive experiments and thorough ablation studies to illustrate the necessity and effectiveness of α-PACE. The performance of our method shows significant improvement against extensive competitive baselines.

FastRAT: Fast and Efficient Cross-lingual Text-to-SQL Semantic Parsing
Pavlos Vougiouklis, Nikos Papasarantopoulos, Danna Zheng, David Tuckey, Chenxin Diao, Zhili Shen, and Jeff Pan
Recent advances of large pre-trained language models have motivated significant breakthroughs in various Text-to-SQL tasks. However, a number of challenges inhibit the deployment of SQL parsers in commercial applications. In this paper, we focus on two such challenges: decoding speed and multilingual input, and introduce FastRAT, a model that includes (i) a decoder-free framework to quickly generate SQL queries from natural language questions based on SQL Semantic Predictions, (ii) a cross-lingual multi-task pre-training scheme, and (iii) a method, based on distant supervision, to extend a semantic parser to new languages. We apply FastRAT on CSpider and Spider, two challenging zero-shot semantic parsing benchmarks. Our system achieves an average of 10x decoding speedup over a set of competitive baselines based on auto- or semi-auto-regressive decoding. In the cross-lingual CSpider dataset, our approach achieves an exact query match accuracy score of 61.3, outperforming the relevant competition. In the monolingual task, it maintains competitive performance by exhibiting less than 5% accuracy drop compared to disproportionately slower solutions.

Learning a Better Initialization for Soft Prompts via Meta-Learning
Yukun Huang, Kun Qian, and Zhou Yu

Prompt tuning (PT) is an effective approach to adapting pre-trained language models to downstream tasks. However, prompt tuning doesn't perform well under few-shot settings due to the poor initialization. So pre-trained prompt tuning (PPT) is proposed to adapt prompt tuning to few-shot settings by initializing prompts with source data. We propose Meta-learned Prompt Tuning to further improve PPT's few-shot learning performance by considering latent structure within the source data. Specifically, we introduce the framework by first clustering source data into different meta-training tasks in an unsupervised manner. Then we leverage these tasks to meta-train prompts with a meta-learning algorithm. Such a process enables prompts to learn a better initialization by discovering commonalities among these meta-training tasks. We evaluate our method on seven downstream sentiment tasks. The results demonstrate that our MetaPT achieves better performance and stability than the state-of-the-art method.

Session 3D – 14:00-15:30

NLP Applications

Negara

Chair: Yixin Cao

Investigating Zero- and Few-shot Generalization in Fact Verification
Liangming Pan, Yunxiang Zhang, and Min-Yen Kan

We explore zero- and few-shot generalization for fact verification (FV), which aims to generalize the FV model trained on well-resourced domains (e.g., Wikipedia) to low-resourced domains that lack human annotations. To this end, we first construct a benchmark dataset collection that contains 11 FV datasets representing 6 domains. We conduct an empirical analysis of generalization across these FV datasets, finding that current models generalize poorly. Our analysis reveals that several factors affect generalization, including dataset size, length of evidence, and the type of claims. Finally, we show that two directions of work improve generalization: 1) incorporating domain knowledge via pretraining on specialized domains, and 2) automatically generating training data via claim generation.

SYNC: A Structurally Guided Hard Negative Curricula for Generalizable Neural Code Search
Atharva Naik, Soumitra Das, Jyothi Vedurada, and Somak Aditya

In neural code search, a Transformers-based pre-trained language model (such as CodeBERT) is used to embed both the query (NL) and the code snippet (PL) into a joint representation space; which is used to retrieve the relevant PLs satisfying the query. These models often make mistakes such as retrieving snippets with incorrect data types, and incorrect method names or signatures. The generalization ability beyond training data is also limited (as the code retrieval datasets vary in the ways NL-PL pairs are collected). In this work, we propose a novel contrastive learning technique (SYNC) that enables efficient finetuning of code LMs with soft and hard negatives, where the hard negatives are constructed using a set of structure-aware AST-based perturbations; targeted towards possible syntactic and semantic variations. Our method achieves significant improvements in retrieval performance for three code LMs (CodeBERT, GraphCodeBERT, UniXCoder) over four Python code retrieval datasets. We also open-source our [code](https://github.com/atharva-naik/SYNC) for reproducibility (https://github.com/atharva-naik/SYNC).

ProMap: Effective Bilingual Lexicon Induction via Language Model Prompting

Abdellah El Mekki, Muhammad Abdul-Mageed, ElMoatez Billah Nagoudi, Ismail Berrada, and Ahmed Khoumsi

Bilingual Lexicon Induction (BLI), where words are translated between two languages, is an important NLP task. While noticeable progress on BLI in rich resource languages using static word embeddings has been achieved. The word translation performance can be further improved by incorporating information from contextualized word embeddings. In this paper, we introduce ProMap, a novel approach for BLI that leverages the power of prompting pretrained multilingual and multidialectal language models to address these challenges. To overcome the employment of subword tokens in these models, ProMap relies on an effective *padded prompting* of language models with a seed dictionary that achieves good performance when used independently. We also demonstrate the effectiveness of ProMap in re-ranking results from other BLI methods such as with aligned static word embeddings. When evaluated on both rich-resource and low-resource languages, ProMap consistently achieves state-of-the-art results. Furthermore, ProMap enables strong performance in few-shot scenarios (even with less than 10 training examples), making it a valuable tool for low-resource language translation. Overall, we believe our method offers both exciting and promising direction for BLI in general and low-resource languages in particular. ProMap code and data are available at https://github.com/4mekki4/promap.

Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method

Xuan ZHANG and Wei Gao

While large pre-trained language models (LLMs) have shown their impressive capabilities in various NLP tasks, they are still under-explored in the misinformation domain. In this paper, we examine LLMs with in-context learning (ICL) for news claim verification, and find that only with 4-shot demonstration examples, the performance of several prompting methods can be comparable with previous supervised models. To further boost performance, we introduce a Hierarchical Step-by-Step (HiSS) prompting method which directs LLMs to separate a claim into several subclaims and then verify each of them via multiple questions-answering steps progressively. Experiment results on two public misinformation datasets show that HiSS prompting outperforms state-of-the-art fully-supervised approach and strong few-shot ICL-enabled baselines.

MCML: A Novel Memory-based Contrastive Meta-Learning Method for Few Shot Slot Tagging

Hongru Wang, Zezhong WANG, Wai Chung Kwan, and Kam-Fai Wong

Meta-learning is widely used for few-shot slot tagging in tasks of few-shot learning. The performance of existing methods is, however, seriously affected by sample forgetting issues, where the model forgets the historically learned meta-training tasks while solely relying on support sets when adapting to new tasks. To overcome this predicament, we propose the Memory-based Contrastive Meta-Learning (aka, MCML) method, including learn-from-the-memory and adaption-from-the-memory modules, which bridge the distribution gap between training episodes and between training and testing respectively. Specifically, the former uses an explicit memory bank to keep track of the label representations of previously trained episodes, with a contrastive constraint between the label representations in the current episode with the historical ones stored in the memory. In addition, the adaption-from-memory mechanism is introduced to learn more accurate and robust representations based on the shift between the same labels embedded in the testing episodes and memory. Experimental results show that the MCML outperforms several state-of-the-art methods on both SNIPS and NER datasets and demonstrates strong scalability with consistent improvement when the number of shots gets more.

Minimum Bayes' Risk Decoding for System Combination of Grammatical Error Correction Systems

Vyas Raina and Mark Gales

For sequence-to-sequence tasks it is challenging to combine individual system outputs. Further, there is also often a mismatch between the decoding criterion and the one used for assessment. Minimum Bayes' Risk (MBR) decoding can be used to combine system outputs in a manner that encourages better alignment with the final assessment criterion. This paper examines MBR decoding for Grammatical Error Correction (GEC) systems, where performance is usually evaluated in terms of edits and an associated F-score. Hence, we propose a novel MBR loss function directly linked to this form of criterion. Furthermore, an approach to expand the possible set of candidate sentences is described. This builds on a current max-voting combination scheme, as well as individual edit-level selection. Experiments on three popular GEC datasets and with state-of-the-art GEC systems demonstrate the efficacy of the proposed MBR approach. Additionally, the paper highlights how varying reward metrics within the MBR decoding framework can provide control over precision, recall, and

the F-score in combined GEC systems.

6

Workshops: Wednesday, November 1

Overview: Workshops

In parallel with *the tutorials* on the same day (p.19).

Wednesday, November 1, 9:00–12:30

Badung 1	2nd Workshop on Information Extraction from Scientific Publications (WIESP)	p.59
Singaraja 1	The Fourth Workshop on Evaluation & Comparison of NLP Systems (Eval4NLP)	p.60
Badung 2	The IJCNLP-AACL 2023 Student Research Workshop (SRW)	p.61

Wednesday, November 1, 14:00-17:30

Badung 1	The 6th Workshop on Financial Technology and Natural Language Processing (FinNLP)	p.62
Singaraja 1	The 11th International Workshop on Natural Language Processing for Social Media (SocialNLP)	p.63
Badung 2	The First Workshop on South East Asian Language Processing (SEALP)	p.64

Wednesday, November 1, *Online*

Online	3rd Workshop on NLP for Medical Conversations (NLPMC)	p.65
Online	Second Workshop on Natural Language Interfaces (NLInt)	p.66
Online	The ART of Safety: Workshop on Adversarial testing and Red-Teaming for generative AI (ArtOfSafety)	p.67

Workshop 1

2nd Workshop on Information Extraction from Scientific Publications (WIESP)

Tirthankar Ghosal (Oak Ridge National Laboratory),
Felix Grezes (Harvard & Smithsonian), **Thomas Allen (Harvard & Smithsonian)**,
Kelly Lockhart Harvard & Smithsonian), **Alberto Accomazzi**,
(Harvard & Smithsonian), **Sergi Blanco-Cuaresma**, **(Harvard & Smithsonian)**

`https://ui.adsabs.harvard.edu/WIESP/2023/`
Badung 1

WIESP provides a platform to researchers to foster discussion and research on information extraction, mining, generation, and knowledge discovery from scientific publications using Natural Language Processing and Machine Learning techniques.

Workshop 2

The Fourth Workshop on Evaluation & Comparison of NLP Systems (Eval4NLP)

Daniel Deutsch (Google Research),
Rotem Dror (University of Pennsylvania and Haifa University),
Steffen Eger (Bielefeld University), **Yang Gao (Google Research)**,
Christoph Leiter (Bielefeld University), **Juri Opitz (Heidelberg University)**,
Andreas Ruckle (Amazon)

`https://eval4nlp.github.io/2023/index.html`
Singaraja 1

Fair evaluations and comparisons are of fundamental importance to the NLP community to properly track progress, especially within the current deep learning revolution, with new state-of-the-art results reported in ever shorter intervals. This concerns the creation of benchmark datasets that cover typical use cases and blind spots of existing systems, the designing of metrics for evaluating the performance of NLP systems on different dimensions, and the reporting of evaluation results in an unbiased manner. Although certain aspects of NLP evaluation and comparison have been addressed in previous workshops (e.g., Metrics Tasks at WMT, NeuralGen, NLG-Evaluation, and New Frontiers in Summarization), we believe that new insights and methodology, particularly in the last 2-3 years, have led to much renewed interest in the workshop topic. The first workshop in the series, Eval4NLP'20 (collocated with EMNLP'20), was the first workshop to take a broad and unifying perspective on the subject matter. The second (Eval4NLP'21 collocated with EMNLP'21) and third (Eval4NLP'22 collocated with AACL'22) workshop extended this perspective. The fourth workshop will continue the tradition and become a reputed platform for presenting and discussing latest advances in NLP evaluation methods and resources. The special topic of this year's version is evaluation of and evaluation through LLM. Therefore, Eval4NLP will feature a shared task on LLM evaluation and encourages the submission of LLM-evaluation focused papers

Workshop 3

The IJCNLP-AACL 2023 Student Research Workshop (SRW)

Dongfang Li (Harbin Institute of Technology),
Rahmad Mahendra (RMIT University), **Zilu Peter Tang (Boston University)**

`https://aacl2023-srw.github.io/`
Badung 2

The IJCNLP-AACL 2023 Student Research Workshop (SRW) will be held in conjunction with the IJCNLP-AACL 2023 conference. The SRW gives student researchers in Computational Linguistics and Natural Language Processing the opportunity to present their work and receive constructive feedback and mentorship by experienced members of the ACL community.

Workshop 4

The 6th Workshop on Financial Technology and Natural Language Processing (FinNLP)

Chung-Chi Chen (National Institute of Advanced Industrial Science and Technology),
Hen-Hsen Huang (Academia Sinica),
Hiroya Takamura (National Institute of Advanced Industrial Science and Technology),
Hsin-Hsi Chen (National Taiwan University), **Hiroki Sakaji (The University of Tokyo)**,
Kiyoshi Izumi (The University of Tokyo)

`https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp2023/home`
Badung 1

The aim of this workshop is to provide a forum where international participants share knowledge on applying NLP to the FinTech domain. Recently, analyzing documents related to finance and economics has attracted much attention in the AI community. In the financial field, FinTech is a new industry that focuses on improving financial activity with technology. Thus, in order to bridge the gap between the NLP research and the financial applications, we organize FinNLP workshop series. One of the expected accomplishments of FinNLP is to introduce insights from the financial domain to the NLP community. With the sharing of the researchers in FinNLP, the challenging problems of blending FinTech and NLP will be identified, and the future research direction will be shaped. That can broaden the scope of this interdisciplinary research area.

Workshop 5

The 11th International Workshop on Natural Language Processing for Social Media (SocialNLP)

Lun-Wei Ku (Academia Sinica), Cheng-Te Li (National Cheng Kung University)

https://sites.google.com/view/socialnlp2023/
Singaraja 1

With the rapid growing of social networks and Web 2.0 services (e.g. Facebook and Twitter), being able to process data come from such platforms has gained much attention in recent years. SocialNLP is a new inter-disciplinary area of natural language processing (NLP) and social computing. We consider three plausible directions of SocialNLP: (1) addressing issues in social computing using NLP techniques; (2) solving NLP problems using information from social networks or social media; and (3) handling new problems related to both social computing and natural language processing.

Several challenges are foreseeable in SocialNLP. First, the message lengths on social media services are usually short (e.g. 140 characters per tweet in Twitter) and thus it is difficult to apply traditional NLP approaches directly. Second, social media contains heterogeneous social information (e.g. tags, friends, followers, endorsements, profiles, and retweets) that should be considered together with the contents for better quality of analysis. Finally, microblogs and the social media contents always involve the interactions among multiple persons with slangs and jargons, and usually require special techniques to distill reasonable information and discover useful knowledge.

To encourage research in the area of SocialNLP, we have organized a SocialNLP SIG-group in AFNLP since 2012 and made it first a yearly now twice a year based workshop series. The latest SocialNLP workshop website is available at 10th SocialNLP 2022 . Through this workshop, we provide a platform for research outcome presentations and head-to-head discussions in the research area of SocialNLP, with the hope to combine the insight and experience of prominent researchers from both NLP and social computing fields to jointly contribute to this area.

Workshop 6

The First Workshop on South East Asian Language Processing (SEALP)

Derry Wijaya (Monash Indonesia), **Alham Fikri Aji (MBZUAI)**,
Clara Vania (Amazon), **Genta Indra Winata (Bloomberg)**, **Ayu Purwarianti (ITB)**

`https://sealp-workshop.github.io/`
Badung 2

South East Asia is one of the most linguistically diverse regions in the world, with over 1200 languages spoken by 680 million people. However, the diversity of South East Asian languages has long been at risk due to the emphasis on national languages as lingua franca in South East Asian countries at the end of colonization; and the increasing prominence of English due to the necessities of globalization.

This workshop will bring together practitioners from academia, government, and industry interested in the research and development of language technologies for SEA languages. The workshop also aims to build an inclusive community of everyone passionate about SEA languages, increase community awareness of works that have been developed to date on these languages, and foster collaborations that will strengthen and spur NLP research and development in SEA languages.

Workshop 7

3rd Workshop on NLP for Medical Conversations (NLPMC)

Sopan Khosla (AWS AI Labs), **Chaitanya Shivade (AWS AI Labs)**,
Vinayshekhar Bannihatti (AWS AI Labs), **Rashmi Gangadharaiah (AWS AI Labs)**,
Thomas Schaaf (3M), **Sandeep Konam (Hitloop)**,
Kevin Lybarger (George Mason University), **Aleksandar Savkov (JABS MedTech)**

`https://nlpmc-2023.github.io/`
Online

The goal of this workshop is to discuss state-of-the-art approaches in conversational AI, as well as share insights and challenges when applied in healthcare. This is critical in order to bridge gaps between research and real-world product deployments, this will further shed light on future directions. This will be a one-day workshop including keynotes, spotlight talks, posters, and panel sessions.

Workshop 8

Second Workshop on Natural Language Interfaces (NLInt)

Vinayshekhar Bannihatti Kumar (AWS AI Labs), **Sopan Khosla (AWS AI Labs)**,
Rashmi Gangadharaiah (AWS AI Labs), **Scott Wen-tau Yih (META AI - FAIR)**,
Ahmed Hassan Awadallah (Microsoft Research),
Tania Bedrax-Weiss (Google Research), **Dan Roth (AWS AI Labs)**,
Katrin Kirchoff (AWS AI Labs)

`https://nlint-workshop.github.io/`
Online

The aim of this workshop is to bring together researchers in both industry and academia who are interested in the advancement of NLInt. Historically, a lot of work on NLInts has happened within the industry, but this workshop aims to bring together both academic and industry researchers to combine their expertise and knowledge. Researchers who work in Natural Language Processing (NLP) and Human Computer Interaction (HCI) are welcome to discuss the existing issues of today's NLInt and suggest improvements that will help in the advancement of the field. We welcome papers along both the modeling domains as well as the interfaces domain. For modeling, researchers could focus on highlighting the issues with existing NLP models/algorithms and/or suggest improvements both in terms of absolute performance as well as latency. For HCI, researchers can focus on the issues with the existing NLInt and discuss interesting new ideas to improve the user-experience of NL interfaces.

Workshop 9

The ART of Safety: Workshop on Adversarial testing and Red-Teaming for generative AI (ArtOfSafety)

Alicia Parrish (Google), **Hannah Rose Kirk (University of Oxford)**, **Jessica Quaye (Harvard University)**, **Charvi Rastogi (Carnegie Mellon University)**, **Max Bartolo (University College London & Cohere)**, **Oana Inel (University of Zurich)**, **Vijay Janapa Reddi (Harvard University)**, **Lora Aroyo (Google)**

`https://sites.google.com/view/art-of-safety/home`
Online

"The ART of Safety" is a workshop on the promise and pitfalls of adversarial testing and red-teaming for safety issues in generative AI. Towards this end, the workshop has two main goals: (i) Collect and compare red-teaming results for Text-2-Image (T2I) models. We welcome both empirical and position papers discussing experiences and results from using the Adversarial Nibbler challenge; (ii) Provide an overview of current approaches, methods and techniques for red-teaming and adversarial testing. We welcome both empirical and position papers discussing various workshop topics including data quality, safety evaluations, safety ground truth, etc.

Index

7 Local Guide

This guide was written by the local chairs of IJCNLP-AACL 2023. For the most up-to-date version, please visit `http://www.afnlp.org/conferences/ijcnlp2023/wp/conference-venue/`

Conference Venue

IJCNLP-AACL 2023 will take place in the Grand Hyatt Bali, located in Nusa Dua, Bali.

Address: Kawasan Wisata Nusa Dua BTDC, Jl. Nusa Dua, Benoa, South Kuta, Badung Regency, Bali 80363 (Phone: +62-361-771234)

Location map (*25 mins by car from Bali airport*)

About Nusa Dua, Bali

Nusa Dua, located just 15 minutes from Bali International Airport, is a popular resort area located on the southern coast of Bali, Indonesia. It's known for its upscale accommodations, beautiful beaches, and serene atmosphere. Nusa Dua boasts some of Bali's most pristine and picturesque beaches. The coastline is adorned with soft, white sands and clear turquoise waters.

Bali, the Island of the Gods, also boasts an array of activities to do for visitors of all ages, including Balinese Kecak Dance show at Uluwatu Temple, best outdoor shopping destination at Bali Collection, stylish dining with live performances and Water Rafting at Ubud's Ayung River.

Area Attractions

Information partly from: `https://www.hyatt.com/en-US/hotel/indonesia/grand-hyatt-bali/balgh/area-attractions`

Dine & Drink

Lunch and coffee breaks are included in the registration fees and will be served in the venue. In addition, there are many dining options nearby, offering both local and international cuisines.

Various dining options at Bali Collection (website)(map) ***4 mins walking***
A dining complex with a series of restaurants, Lounge Bars, and Cafes.

Bebek Bengil Nusa Dua (website)(map) ***9 mins walking***
Bebek Bengil is one of the legendary culinary destinations in Bali. Located in Nusa Dua, near Grand Hyatt Bali, Bebek Bengil Nusa Dua boasts beautiful beach views, and its signature crispy duck dish.

Kekeb Restaurant (website)(map) ***10 mins walking***
Restaurant with a beachside view. Serve Indonesian food, authentic Balinese food, and live Seafood

Ginger House Restaurant (website)(map) ***8 mins walking***
Italian restaurant. Vegetarian Friendly, Vegan Options, Gluten Free Options.

The Royal Kitchen Bali (website)(map) ***9 mins walking***
As one of the very few authentic Indian restaurants in Indonesia, The Royal Kitchen serves its signature North Indian dishes and new age cocktails.

Nusa By/Suka – Restaurant & Bar (website)(map) ***9 mins walking***
Comfort food, modern and inspired wood fired fish, steaks and roasts, expansive international wine list and hand crafted signature cocktails.

Kagura Authentic Japanese Cuisine (website)(map) ***12 mins walking***
Tokyo style fine dining restaurant.

Tropical Restaurant Mengiat (website)(map) ***12 mins walking***
Pizza, Seafood, Asian and Indonesian food

Warung Nasi Ayam Ibu Oki (map) ***18 mins walking***
Balinese authentic chicken dishes (*ayam betutu*) served in a relaxed local eatery.

Warung Babi Guling Sari Dewi Bp. Dobil (map) ***22 mins walking***
Balinese authentic suckling pig (*babi guling*) served with crispy skin, rice, and fiery chillies.

Natys Sigilita (website)(map) ***18 mins walking***
Located at a strategic place and next to Coco Supermarket Shopping Center. Serve foods and drinks with international flavours at domestic prices.

Mr Bob Bar and Grill Nusadua (Second Outlet) (map) ***25 mins walking***
Watering hole & cafe known for ribs, grilled chicken & fish in a comfortable setting.

Udupi (website)(map) ***30 mins walking***
Vegetarian Indian restaurant serving foods from both North and South India regions as well as Nusantara and Indo-Chinese dishes.

Kebabs and Kurries (website)(map) ***32 mins walking***
The Kebabs & Kurries story began in 2020 when Sattvik by Nature extended its operation to Bali. With authentic Indian cuisine prepared by Indian chefs, serving a good feast of fresh Veg and Non Veg Indian food.

Balicious (map) ***36 mins walking***
Late-night haunt serving traditional Indonesian fare and seafood in bright digs with outdoor tables.

N Café Jimbaran (website)(map) ***18 mins driving***
N Café Jimbaran provides a wide range of delectable Indonesian and Western cuisine in the bustling centre of the Jimbaran area. Guests can opt for a relaxed dining time with friends and loved ones indoors or just simply chill out under Bali's sun at the restaurant's terrace.

"La Brasserie" by Melting Wok (website)(map) ***17 mins driving***

Situated at the heart of the Jimbaran bay area, La Brasserie by Melting Wok offers a variety of international delights featuring Asian fusion, French, and selections of vegetarian and gluten-free dishes.

Cuca Restaurant (website)(map) ***18 mins driving***

Cuca serves-up Tapas, Cocktails and Desserts. Expect inventive comfort food meant for sharing, bursting with tropical flavours, balancing textures, colours and tastes.

Queen's Tandoor – Seminyak (website)(map) ***27 mins driving***

Established since 2004, Queen's Tandoor has been grabbing the attention of Indian cuisine enthusiasts ever since. With four Bali outlets in Seminyak, Kuta, Nusa Dua and Ubud, it's a must-visit for authentic Indian culinary delights.

Classic Arts and Culture

Kecak and Barong Dance The Nusa Dua (website)(map) ***16 mins walking***

It is a concept of Balinese Barong and Kecak Dance arts which are packaged in a single performance at Taksu Art Stage on Peninsula Island at Nusa Dua, Bali. The show begins with a Barong performance and then continues with a Kecak Dance performance with stories taken from Ramayana epic. The show is from 6-7pm every Friday.

GWK Cultural Park (website)(map) ***30 mins driving***

Covering about 60 hectares of land on the elevated area of Ungasan, GWK Cultural Park is a one-stop destination that offers a variety of F&B outlets, cultural performances, and the most breathtaking of all, the towering Garuda Wisnu Kencana statue – standing tall at 120.9 metres high.

Uluwatu Kecak & Fire Dance (website)(map) ***50 mins driving***

Defined as one of Bali's most popular cultural attractions, Kecak dance follows the story of the epic tale of Ramayana featuring the haunting chants of kecak. Here at Uluwatu cliff, visitors can witness the ever-famous Kecak dance with a stunning view of the island's southwestern coast and its beautiful sunset. Kecak dance takes place at the Uluwatu Temple from 5:00 PM to 6:00 PM every day.

Agung Rai Museum of Art Bali (website)(map) ***1.5 hours driving***

The Agung Rai Museum of Art – more familiarly known as ARMA – is home to a vast collection of preserved artworks, from paintings and sculptures, to music, dance and other cultural art forms. First opened in 1996 by art connoisseur Agung Rai, the permanent exhibition at ARMA is not to be missed.

Historic Destinations

Tanah Lot (website)(map) ***1 hour driving***

Tanah Lot is a rock formation off the Indonesian island of Bali. It is home to the ancient Hindu pilgrimage temple Pura Tanah Lot. Tanah Lot means "Land [in the] Sea" in the Balinese language. The temple sits on a large offshore rock which has been shaped continuously over the years by the ocean tide.

Sacred Monkey Forest Sanctuary (website)(map) ***1.5 hours driving***

Watch and interact with Balinese long-tailed monkeys in their sanctuary and natural habitat of Ubud Monkey Forest. Home to more than a thousand monkeys, the famous monkey forest also opens its doors to its three temples that are worth visiting.

Pura Tirta Empul (website)(map) ***1.5 hours driving***

Located near the town of Tampaksiring, the Balinese Hindu water temple Pura Tirta Empul is a popular destination, for locals and travellers alike. Many come to Tirta Empul to partake in the traditional purification ritual known as "melukat", but many others also come to enjoy the sceneries.

Pura Besakih (website)(map) ***2 hours driving***

Pura Besakih is the largest and one of the most important holy temples in Bali. Its location on the slopes of Bali's majestic Mount Agung makes it a sought-after travel destination, in addition to its enchanting tiered construction.

Tirta Gangga (website)(map) ***2 hours driving***

This former royal water palace is a popular destination in the eastern part of Bali with its surrounding lagoons and gardens.

Ulun Danu Beratan Temple (website)(map) ***2 hours driving***

Scenic Hindu temple on a lake with gardens, boating & wild animals.

Bali Unique Outdoor Scenes

Nusa Dua Beach (map) ***5 mins walking***

White sandy beach just next to the venue with water sports on offer.

Waterblow (website)(map) ***16 mins walking***

Water Blow in Nusa Dua lets you witness the awesome power of nature as large

waves from the Indian Ocean constantly crash against the jagged limestone edges on the peninsula's south-eastern cliff. Here you'll find 240 degrees of dramatic seascape, with the irregular splashes and sprays simply adding to the fun of it. The so-called 'water blows' result from the narrowing crag below the cliff face that channels a massive surge of water up to 30m high from its base following strong currents

Kuta beach (website)(map) ***30 mins driving***

Bustling sandy stretch lined with restaurants & resorts, a well-known spot for surfing & sunsets.

Sanur beach (map) ***40 mins driving***

Sanur is a seaside town in the southeast of the island of Bali, in Indonesia. Its long stretch of beach offers shallow waters. Colourful jukung fishing boats rest on the sand, backed by a paved cycling path. The Pura Blanjong temple is built from coral and has inscriptions dating from the 10th century. The leafy main street Jalan Danau Tamblingan is lined with art galleries and restaurants.

Blue Point Beach (map) ***50 mins driving***

Originally called Suluban Beach, Blue Point Beach is one of Bali's most beautiful beaches with turquoise blue water and white sand. Slightly hidden underneath Uluwatu's dramatic cliffs, Blue Point is popular among surfers and beach hunters.

Kanto Lampo Waterfall Gianyar (map) ***1.5 hours driving***

Kanto Lampo Waterfall offers a mesmerising sight to behold, and the destination is relatively easy to reach. Standing 15 metres tall, the tiered waterfall and the surrounding rocks make picturesque shapes, a true idyllic photo-op setting.

Tegallalang Rice Terrace Ubud (website)(map) ***2 hours driving***

Tegallalang is known and loved for the soothing views of rice terraces. Travellers can see how the traditional Balinese cooperative irrigation system called "subak" works, or simply just breathe in the fresh air and the serene and scenic views.

Bali Pulina Coffee Plantation (website)(map) ***2 hours driving***

Coffee enthusiasts should not miss a visit to Bali Pulina Coffee Plantation. Here you can learn about and take part in luwak coffee production. The plantation is located in the lush Tegallalang district, so in addition to the good coffee, you will also get to bask in the refreshing ambience.

Trunyan (website)(map) ***3 hours driving***

Trunyan or Terunyan is a Balinese village located on the eastern shore of Lake Batur.

The village is one of the most notable homes of the Bali Aga people. Trunyan is notable for its peculiar treatment of dead bodies, in which they are placed openly on the ground, simply covered with cloth and bamboo canopies, and left to decompose. The influence of a nearby tree is said to remove the putrid smell of the corpses.

Day-trip Getaways

Bali Zoo (website)(map) ***1 hour driving***

Offering fun for the whole family, Bali Zoo showcases an extensive variety of animals, in addition to the animal-friendly and educational interactive exhibits. Book ahead to make the most of Bali Zoo's programs and activities.

Water Rafting at Ubud's Ayung River (website)(map) ***1.5 hours driving***

Water rafting at Ubud's Ayung River offers a different kind of thrill for adventure seekers to pump some adrenaline with an approximately 10 kilometre route that will take around 1.5 to 2 hours to explore. While crossing through the river, travellers will be greeted by the fresh mountain air and Bali's lush green rainforest.

Bali Strawberry Farm and Restaurant (map) ***2 hours driving***

If you're looking for something more relaxing in the Bedugul area, head to Bali Strawberry Farm and Restaurant. You can pick your own strawberries to bring home, unwind at the restaurant and enjoy the menu on offer.

Gitgit Waterfall (website)(map) ***2.5 hours driving***

Gitgit is one of the most well-known waterfalls in Bali, thanks to its height and breathtaking natural surroundings, including verdant foliage and natural swimming pools that are accessible by a rocky walking trail.

Mount Batur (website)(map) ***2.5 hours driving***

Mount Batur (Gunung Batur) is an active volcano located at the centre of two concentric calderas north west of Mount Agung on the island of Bali, Indonesia. Mount Batur, which last erupted in 2000, is a popular hiking spot. It's recommended to make the hour-long climb in the early hours so you can reach the peak in time for sunrise.

Lake Batur (website)(map) ***2.5 hours driving***

Lake Batur is a volcanic crater lake in Kintamani, Bali, Bangli Regency of Bali, located about 30 km northeast of Ubud in Bali. The lake is inside of the caldera of an active volcano, Mount Batur, located along the Ring of Fire of volcanic activity.

Also, a selection of ways to tour Bali from TripAdvisor including touring Ubud, ATV Quad Biking, touring Nusa Penida and snorkelling, snorkelling at blue lagoon beach, Mount Batur Sunrise hiking and natural hot spring visit, and Bali waterfalls tour.

Other beautiful islands near Bali

If you have time to spend before or after the conference, you can also take a trip to beautiful islands near Bali that include:

Lombok (website)
Lombok is an island east of Bali. It's known for beaches and surfing spots, particularly at Kuta and Banko Banko. The motor-vehicle-free Gili Islands (Gili Trawangan, Gili Air and Gili Meno), off Lombok's west coast, offer more beaches, reefs for diving and snorkelling, and a sea turtle hatchery.

Komodo Island (website)
Komodo island is particularly notable as the habitat of the Komodo dragon, the largest lizard on Earth, which is named after the island. In addition, the island is a popular destination for diving.

Shopping

Bali Collection (website)(map) ***4 mins walking***
Open-air mall with clothing, jewellery & souvenir stores, supermarkets & diverse dining options.

Uluwatu Handmade Balinese Lace (website)(map) ***16 mins walking***
Fashion enthusiasts can't miss an array of stunning and premium linen dresses, clothes and accessories at Uluwatu Handmade Balinese Lace. Each piece is beautifully handsewn by the local artisans, offering a unique and authentic design in every collection. In the Nusa Dua area, the store is located inside Bali Collection and is open every day from 11AM to 7PM local time.

T Galleria By DFS Bali (website)(map) ***20 mins driving***
T Galleria brings duty-free shopping to another level by offering concierge service, complimentary Wi-Fi, luggage storage, mobile charging stations, and more facilities for your convenience. You can even shop here and pick up the items at Ngurah Rai International Airport.

Discovery Shopping Mall (map) ***25 mins driving***

This beachfront shopping mall in Kuta sprawls across 3.8 hectares and is one of the most popular shopping destinations in Bali. Discovery Shopping Mall hosts more than 150 tenants, including two renowned department stores, Centro and Sogo. It also serves as a nice spot to catch the sunset.

Beachwalk Shopping Center (website)(map) ***30 mins driving***

Beachwalk Shopping Center is a beautifully designed shopping mall featuring high-end brands. More than just a shopping destination, Beachwalk also carries popular F&B household names, entertainment offers, and more, making it an elaborate lifestyle hub.

Other accommodations

The convenient option is to stay at the conference venue, Grand Hyatt Bali. You can also choose other accommodation options. Some nearby hotels and homestays (*30 mins walking distance*) with at least 4.5 rating on Google are listed below for your convenience.

Hotels

Amarterra Villas Resort Bali Nusa Dua (website)(map) ***7 mins walking***

Merusaka Nusa Dua (website)(map) ***10 mins walking***

Ayodya Resort Bali (website(map) ***13 mins walking***

Courtyard by Marriott Bali Nusa Dua Resort (website)(map) ***13 mins walking***

The Laguna, a Luxury Collection Resort & Spa (website)(map) ***16 mins walking***

Melia Bali (map) ***17 mins walking***

Bali National Golf Club (website)(map) ***17 mins walking***

The Westin Resort Nusa Dua, Bali (website)(map) ***17 mins walking***

Menzel Villa Nusa Dua (website)(map) ***18 mins walking***

Marriott's Bali Nusa Dua Gardens (website)(map) ***19 mins walking***

Kayumanis Nusa Dua Private Villa (website)(map) ***20 mins walking***

Novotel Bali Nusa Dua – Hotel & Residences (website)(map) ***20 mins walking***

The Grand Bali Nusa Dua (website)(map) ***21 mins walking***

Bali Nusa Dua Hotel (website)(map) ***22 mins walking***

Nusa Dua Beach Hotel & Spa, Bali (website)(map) ***21 mins walking***

The St. Regis Bali Resort (website)(map) ***28 mins walking***

Nusa Dua Moon Hotel (website)(map) ***30 mins walking***

Allia Residence (website)(map) ***30 mins walking***

Homestays

The Maj Nusa Dua (map) ***13 mins walking***

Kubu Dimel Homestay (website)(map) ***20 mins walking***

Ayu Homestay Nusa (map) ***28 mins walking***

Bali Sunshine Inn by Udupi (map) ***30 mins walking***

Jepun Bali Homestay (website)(map) ***30 mins walking***

vivo

人工智能全球研究院

vivo AI Lab

团队介绍 About Us

团队愿景 Vision

用AI为用户带来无处不在的惊喜和激动人心的体验。

Surprise and excite users with AI on all sides.

团队工作 Teamwork

vivo 人工智能全球研究院成立于2017年，致力于打造行业一流的AI技术，为用户提供极致的产品体验。专注于计算机视觉、语音技术、自然语言处理和机器学习等领域的基础研究和应用探索。

目前已经在深圳、北京、杭州、南京建立了人工智能研发基地，拥有超过1000名人工智能工程师，其中海外名校背景的人才占比超过两成。2018年开始到现在，vivo人工智能团队累计发表70多篇全球顶会的论文，申请超近千项AI专利。

vivo AI Global Lab, founded in 2017, has been committed to building industry-leading AI technologies and providing users with ultimate product experience. Areas of work include Computer Vision, Speech Technology, Natural Language Processing and Machine Learning.

We have AI R&D bases in Shenzhen, Beijing, Hangzhou and Nanjing, with more than 1,000 artificial intelligence engineers, of which more than 20% are talents with backgrounds from prestigious overseas schools. Since 2018, the vivo artificial intelligence team has published more than 70 papers at top global conferences and applied for nearly a thousand AI patents.

招聘信息 Join Us

岗位名称：NLP算法实习生/大模型算法实习生

简历投递邮箱：lifangyuan@vivo.com（李女士）

Job Title: NLP/Large Language Model Intern

Application Email: lifangyuan@vivo.com (Ms. Li)

岗位名称：语音算法实习生

简历投递邮箱：bb.chen@vivo.com（陈先生）

Job Title: Speech Algorithm Intern

Application Email: bb.chen@vivo.com (Mr. Chen)

联系我们 Contact Us

vivo人工智能技术公众号

vivo AI Technology

人工智能小喇叭

vivo AI-WeChat ID

AFNLP

ZHIPU · AI

open models | open code | open evaluation

ChatGLM-6/12/32/66/130B

Open Bilingual Foundation & Dialogue Models
chatglm.ai

GLM-130B

An Open Bilingual Pre-Trained Model

CodeGeeX

Open Multilingual Code Generation Models

CogView **CogVideo** **CogVLM**

Open Vision Language Models

We Are More Open

We have openly released various LLMs, including GLM-130B, ChatGLM-6B, ChatGLM2-6B, VisualGLM-6B, CodeGeeX2-6B, CodeGeeX-13B, CogView, CogView2, CogVideo, and CogVLM, which have been downloaded for more than 10 million times worldwide.

All models, code, and evaluation are at github.com/THUDM & https://huggingface.co/THUDM.

Sponsors

IJCNLP-AACL 2023 is extremely grateful to all sponsors. We simply couldn't run the conference without the help of these generous organizations. We thank them sincerely for their ongoing support of the NLP/CL community.

PLATINUM

GOLD

SILVER

BRONZE