The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing

Proceedings of the Student Research Workshop

August 5-6, 2021
Bangkok, Thailand (online)
Introduction

Welcome to the ACL-IJCNLP 2021 Student Research Workshop!

The ACL-IJCNLP 2021 Student Research Workshop (SRW) is a forum for student researchers in computational linguistics and natural language processing. The workshop provides a unique opportunity for student participants to present their work and receive valuable feedback from the international research community as well as from faculty mentors.

Following the tradition of the previous student research workshops, we have two tracks: research papers and thesis proposals. The research paper track is a venue for Ph.D. students, Masters students, and advanced undergraduates to describe completed work or work-in-progress along with preliminary results. The thesis proposal track is offered for advanced Masters and Ph.D. students who have decided on a thesis topic and are interested in feedback on their proposal and ideas about future directions for their work.

This year, the student research workshop has again received wide attention. We received 114 submissions including 109 research papers and 5 thesis proposals. The submissions included 68 long papers and 46 short papers. Following withdrawals and desk rejects, 45 were accepted for an acceptance rate of 39%. Excluding non-archival papers, 36 papers appear in these proceedings. All the accepted papers will be presented virtually in three sessions during the course of August 3rd.

Mentoring is at the heart of the SRW. In keeping with previous years, we had a pre-submission mentoring program before the submission deadline. A total of 36 papers participated in the pre-submission mentoring program. This program offered students the opportunity to receive comments from an experienced researcher to improve the writing style and presentation of their submissions.

We are deeply grateful to the Swiss National Science Foundation (SNSF) for providing funds that covered student registrations. We thank our program committee members for their careful reviews of each paper and all of our mentors for donating their time to provide feedback to our student authors. Thank you to our faculty advisors, Jing Jiang, Rico Sennrich, Derek F. Wong and Nianwen Xue, for their essential advice and guidance, and to the ACL-IJCNLP 2021 organizing committee for their support. Finally, thank you to our student participants!
Organizers:

Jad Kabbara, McGill University and the Montreal Institute for Learning Algorithms (MILA)
Haitao Lin, Institute of Automation, Chinese Academy of Sciences
Amandalynne Paullada, University of Washington
Jannis Vamvas, University of Zurich

Faculty Advisors:

Jing Jiang, Singapore Management University
Rico Sennrich, University of Edinburgh
Derek F. Wong, University of Maca
Nianwen Xue, Brandeis University

Pre-submission Mentors:

Duygu Ataman, University of Zürich
Valerio Basile, University of Turin
Eduardo Blanco, University of North Texas
David Chiang, University of Notre Dame
Marta R. Costa-Jussà, Universitat Politècnica de Catalunya
Lucia Donatelli, Saarland University
Greg Durrett, UT Austin
Sarah Ebling, University of Zurich
Yansong Feng, Peking University
Orhan Firat, Google AI
Lea Frermann, Melbourne University
Shujian Huang, National Key Laboratory for Novel Software Technology, Nanjing University
Kentaro Inui, Tohoku University / Riken
Robin Jia, Facebook AI Research
Katharina Kann, University of Colorado Boulder
Mamoru Komachi, Tokyo Metropolitan University
Parisa Kordjamshidi, Michigan State University
Jindřich Libovický, Ludwig Maximilian University of Munich
Pengfei Liu, Carnegie Mellon University
Vincent Ng, University of Texas at Dallas
Sai Krishna Rallabandi, Carnegie Mellon University
Masoud Rouhizadeh, Johns Hopkins University
Dipti Sharma, IIIT, Hyderabad
Manish Shrivastava, International Institute of Information Technology Hyderabad
Sunayana Sitaram, Microsoft Research India
Gabriel Stanovsky, The Hebrew University of Jerusalem
Amanda Stent, Bloomberg
Hanna Suominen, The Australian National University, Data61/CSIRO, and University of Turku
Mihai Surdeanu, University of Arizona
Masashi Toyoda, The University of Tokyo
Chen-Tse Tsai, Bloomberg LP
Bonnie Webber, University of Edinburgh
Yujiu Yang, tsinghua.edu.cn
Arkaitz Zubiaga, Queen Mary University of London
Program Committee:

Assina Abdussaitova, Suleyman Demirel University
Ibrahim Abu Farha, University of Edinburgh
Oshin Agarwal, University of Pennsylvania
Piush Aggarwal, University of Duisburg-Essen, Language Technology Lab
Roee Aharoni, Google
Miguel A. Alonso, Universidade da Coruña
Malik Altakrori, McGill University /Mila
Rami Aly, University of Cambridge
Bharat Ram Ambati, Apple Inc.
Aida Amini, University of Washington
Maria Antoniak, Cornell University
Tal August, University of Washington
Vidhisha Balachandran, Carnegie Mellon University
Anusha Balakrishnan, Microsoft Semantic Machines
Jorge Balazs, Amazon
Roberto Basili, University of Roma, Tor Vergata
Rachel Bawden, Inria
Chris Biemann, Universität Hamburg
Tatiana Bladier, Heinrich Heine University Düsseldorf
Nikolay Bogoychev, University of Edinburgh
Avishek Joey Bose, Mila/McGill
Ruken Cakici, METU
Ronald Cardenas, University of Edinburgh
Arlene Casey, University of Edinburgh
Aishik Chakraborty, McGill University
Jonathan P. Chang, Cornell University
Jifan Chen, UT Austin
Sihao Chen, University of Pennsylvania
Elizabeth Clark, University of Washington
Xiang Dai, University of Copenhagen
Siddharth Dalmia, Carnegie Mellon University
Samvit Dammalapati, Indian Institute of Technology Delhi
Alok Debnath, Factmata
Louise Deléger, INRAE - Université Paris-Saclay
Pieter Delobelle, KU Leuven, Department of Computer Science
Dorottya Demszky, Stanford University
Etienne Denis, McGill
Chris Develder, Ghent University
Anne Dirkson, Leiden University
Radina Dobrev, University of Edinburgh
Zi-Yi Dou, UCLA
Hicham El Boukouri, LIMSI, CNRS, Université Paris-Saclay
Carlos Escolano, Universitat Politècnica de Catalunya
Luis Espinosa Anke, Cardiff University
Tina Fang, University of Waterloo
Murhaf Fares, University of Oslo
Amir Feder, Technion - Israel Institute of Technology
Jared Fernandez, Carnegie Mellon University
Gosse Minnema, University of Groningen
Amita Misra, IBM
Omid Moradiannasab, Saarland University
Nora Muheim, University of Bern
Masaaki Nagata, NTT Corporation
Aakanksha Naik, Carnegie Mellon University
Denis Newman-Griffis, University of Pittsburgh
Dat Quoc Nguyen, VinAI Research
Vincent Nguyen, Australian National University & CSIRO Data61
Shinji Nishimoto, CiNet
Yasumasa Onoe, The University of Texas at Austin
Silviu Oprea, University of Edinburgh
Naoki Otani, Carnegie Mellon University
Ashwin Paranjape, Stanford University
Archita Pathak, University at Buffalo (SUNY)
Viviana Patti, University of Turin, Dipartimento di Informatica
Siyao Peng, Georgetown University
Ian Porada, Mila, McGill University
Jakob Prange, Georgetown University
Adithya Pratapa, Carnegie Mellon University
Yusu Qian, New York University
Long Qiu, Onehome (Beijing) Network Technology Co. Ltd.
Ivaylo Radev, IICT-BAS
Sai Krishna Rallabandi, Carnegie Mellon University
Vikas Raunak, Microsoft
Lina M. Rojas Barahona, Orange Labs
Guy Rotman, Faculty of Industrial Engineering and Management, Technion, IIT
Maria Ryskina, Carnegie Mellon University
Farig Sadeque, Educational Testing Service
Jin Sakuma, University of Tokyo
Elizabeth Salesky, Johns Hopkins University
Younes Samih, University of Düsseldorf
Ramon Sanabria, The University Of Edinburgh
Michael Sejr Schlichtkrull, University of Amsterdam
Sebastian Schuster, New York University
Olga Seminck, CNRS
Indira Sen, GESIS
Vasu Sharma, Carnegie Mellon University
Sina Sheikholeslami, KTH Royal Institute of Technology
A.B. Siddique, University of California, Riverside
Kevin Small, Amazon
Marco Antonio Sobrevilla Cabezudo, University of São Paulo
Katira Soleymanzadeh, Ege University
Swapna Somasundaran, Educational Testing Service
Sandeep Soni, Georgia Institute of Technology
Richard Sproat, Google, Japan
Makesh Narsimhan Sreedhar, Mila, Universite de Montreal
Tejas Srinivasan, Microsoft
Vamshi Krishna Srirangam, International Institute of Information Technology, Hyderabad
Marija Stanojevic, Center for Data Analytics and Biomedical Informatics, Temple University
Shane Steinert-Threlkeld, University of Washington

viii
Investigation on Data Adaptation Techniques for Neural Named Entity Recognition
Evgeniia Tokarchuk, David Thulke, Weiyue Wang, Christian Dugast and Hermann Ney ........... 1

Stage-wise Fine-tuning for Graph-to-Text Generation
Qingyun Wang, Semih Yavuz, Xi Victoria Lin, Heng Ji and Nazneen Rajani .................. 16

Transformer-Based Direct Hidden Markov Model for Machine Translation
Weiyue Wang, Zijian Yang, Yingbo Gao and Hermann Ney ................................ 23

AutoRC: Improving BERT Based Relation Classification Models via Architecture Search
Wei Zhu ................................................................................................................. 33

How Low is Too Low? A Computational Perspective on Extremely Low-Resource Languages
Rachit Bansal, Himanshu Choudhary, Ravneet Punia, Niko Schenk, Émilie Pagé-Perron and Jacob Dahl ............................................................ 44

On the Relationship between Zipf’s Law of Abbreviation and Interfering Noise in Emergent Languages
Ryo Ueda and Koki Washio ................................................................. 60

Long Document Summarization in a Low Resource Setting using Pretrained Language Models
Ahsaas Bajaj, Pavitra Dangati, Kalpesh Krishna, Pradhiksha Ashok Kumar, Rheeya Uppaal, Bradford Windsor, Eliot Brenner, Dominic Dotterrer, Rajarshi Das and Andrew McCallum .............. 71

Attending Self-Attention: A Case Study of Visually Grounded Supervision in Vision-and-Language Transformers
Jules Samaran, Noa Garcia, Mayu Otani, Chenhui Chu and Yuta Nakashima ............... 81

Video-guided Machine Translation with Spatial Hierarchical Attention Network
Weiqi Gu, Haiyue Song, Chenhui Chu and Sadao Kurohashi ................................ 87

Stylistic approaches to predicting Reddit popularity in diglossia
Huikai Chua ......................................................................................................... 93

"I've Seen Things You People Wouldn't Believe": Hallucinating Entities in GuessWhat?!
Alberto Testoni and Raffaella Bernardi .......................................................... 101

How do different factors Impact the Inter-language Similarity? A Case Study on Indian languages
Sourav Kumar, Salil Aggarwal, Dipti Misra Sharma and Radhika Mamidi ............... 112

COVID-19 and Misinformation: A Large-Scale Lexical Analysis on Twitter
Dimosthenis Antypas, Jose Camacho-Collados, Alun Preece and David Rogers ........ 119

Situation-Based Multiparticipant Chat Summarization: a Concept, an Exploration-Annotation Tool and an Example Collection
Anna Smirnova, Evgeniy Slobodkin and George Chernishev ............................... 127

Modeling Text using the Continuous Space Topic Model with Pre-Trained Word Embeddings
Seiichi Inoue, Taichi Aida, Mamoru Komachi and Manabu Asai .......................... 138

Semantics of the Unwritten: The Effect of End of Paragraph and Sequence Tokens on Text Generation with GPT2
He Bai, Peng Shi, Jimmy Lin, Luchen Tan, Kun Xiong, Wen Gao, Jie Liu and Ming Li .... 148
Data Augmentation with Unsupervised Machine Translation Improves the Structural Similarity of Cross-lingual Word Embeddings
Sosuke Nishikawa, Ryokan Ri and Yoshimasa Tsuruoka ............................................... 163

Joint Detection and Coreference Resolution of Entities and Events with Document-level Context Aggregation
Samuel Kriman and Heng Ji .......................................................................................... 174

"Hold on honey, men at work": A semi-supervised approach to detecting sexism in sitcoms
Smriti Singh, Tanvi Anand, Arijit Ghosh Chowdhury and Zeerak Waseem .................... 180

Observing the Learning Curve of NMT Systems With Regard to Linguistic Phenomena
Patrick Stadler, Vivien Macketanz and Eleftherios Avramidis ................................... 186

Improving the Robustness of QA Models to Challenge Sets with Variational Question-Answer Pair Generation
Kazutoshi Shinoda, Saku Sugawara and Akiko Aizawa ................................................ 197

Tools Impact on the Quality of Annotations for Chat Untangling
Jhonny Cerezo, Felipe Bravo-Marquez and Alexandre Henri Bergel ........................... 215

How Many Layers and Why? An Analysis of the Model Depth in Transformers
Antoine Simoulin and Benoit Crabbé ........................................................................... 221

Edit Distance Based Curriculum Learning for Paraphrase Generation
Sora Kadotani, Tomoyuki Kajiwara, Yuki Arase and Makoto Onizuka ....................... 229

Changing the Basis of Contextual Representations with Explicit Semantics
Tamás Ficsor and Gábor Berend .................................................................................. 235

Personal Bias in Prediction of Emotions Elicited by Textual Opinions
Piotr Milkowski, Marcin Gruza, Kamil Kanclerz, Przemyslaw Kazienko, Damian Grimling and Jan Kocon ................................................................................................................. 248

MVP-BERT: Multi-Vocab Pre-training for Chinese BERT
Wei Zhu ......................................................................................................................... 260

CMTA: COVID-19 Misinformation Multilingual Analysis on Twitter
Raj Pranesh, Mehrdad Farokhenajd, Ambesh Shekhar and Genoveva Vargas-Solar .... 270

Predicting pragmatic discourse features in the language of adults with autism spectrum disorder
Christine Yang, Duanchen Liu, Qingyun Yang, Zoey Liu and Emily Prud’hommeaux .......... 284

SumPubMed: Summarization Dataset of PubMed Scientific Articles
Vivek Gupta, Prerna Bharti, Pegah Nokhiz and Harish Karnick ..................................... 292

A Case Study of Analysis of Construals in Language on Social Media Surrounding a Crisis Event
Lolo Aboufoul, Khyati Mahajan, Tiffany Gallicano, Sara Levens and Samira Shaikh ....... 304

Cross-lingual Evidence Improves Monolingual Fake News Detection
Daryna Dementieva and Alexander Panchenko ......................................................... 310

Neural Machine Translation with Synchronous Latent Phrase Structure
Shintaro Harada and Taro Watanabe ............................................................................ 321
<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero Pronouns Identification based on Span prediction</td>
<td>Sei Iwata, Taro Watanabe and Masaaki Nagata</td>
<td>331</td>
</tr>
<tr>
<td>On the differences between BERT and MT encoder spaces and how to address them in translation tasks</td>
<td>Raúl Vázquez, Hande Celikkanat, Mathias Creutz and Jörg Tiedemann</td>
<td>337</td>
</tr>
<tr>
<td>Synchronous Syntactic Attention for Transformer Neural Machine Translation</td>
<td>Hiroyuki Deguchi, Akihiro Tamura and Takashi Ninomiya</td>
<td>348</td>
</tr>
</tbody>
</table>