COLING

International Conference on Computational Linguistics

Proceedings of the Conference and Workshops

COLING

Volume 29 (2022), No. 17

Proceedings of 1st Workshop on NLP applications to field linguistics

The 29th International Conference on Computational Linguistics

October 16, 2022
Gyeongju, Republic of Korea

Copyright of each paper stays with the respective authors (or their employers).

ISSN 2951-2093

Message from the Program Chairs

Field Matters is a workshop focused on various applications of NLP methods to field linguistics and analysis of field data with the help of computational linguistics. On the one hand, field linguists document language data, but the fieldwork involves tons of manual annotation or analysis, which might be significantly sped up with computational instruments. On the other hand, NLP research brought methods for different tasks that show significant performance in high-resource languages, allowing to automate various routine tasks. The future development of NLP methods could gain from the language diversity of under-resourced languages.

Field Matters is aimed to combine linguistic fieldwork and NLP methods. Our workshop is hosted by the 29th International Conference on Computational Linguistics (COLING 2022). To provide the comprehensive diverse expertise in a multidisciplinary setting, we invited linguists and NLP researchers worldwide to our program committee. After the hard process of reviewing all submissions, the program committee chose nine papers for a poster or oral presentation at the workshop. Accepted papers illustrate the main idea of our workshop: how field linguistics may benefit from using contemporary methods of computational analysis and how the NLP community may evolve its methods with the help of under-resourced languages.

We are incredibly grateful to the Field Matters program committee, who worked on peer review to give meaningful comments for each submission and made this workshop possible. We want to thank the invited speakers, Antonios Anastasopoulos and Steven Bird, for contributing to the program. We would also like to mention all the authors who submitted their papers to our workshop, and we hope to continue to serve as a link between NLP specialists and field linguists.

You can find more details about the workshop on our website: https://field-matters.github.io/

Organising Committee

Program Chairs

Oleg Serikov, HSE University, AIRI, MIPT, RAS Linguistics
Ekaterina Voloshina, HSE University, AIRI
Anna Postnikova, HSE University
Elena Klyachko, Institute of Linguistics RAS, HSE University
Ekaterina Neminova, HSE University
Ekaterina Vylomova, University of Melbourne
Tatiana Shavrina, AIRI, SberDevices
Éric Le Ferrand, Charles Darwin University, Université Grenoble Alpes
Valentin Malykh, Huawei
Francis Tyers, Indiana University, HSE University
Timofey Arkhangelskiy, University of Hamburg
Vladislav Mikhailov, SberDevices, HSE University
Alena Fenogenova, SberDevices

Program Committee

Program Committee

Alexandre Arkhipov, University of Hamburg
Rolando Coto Solano, Dartmouth College
Emily Prud'hommeaux, Boston College
Robbie Jimmerson, Rochester Institute of Technology
Zoe Liu, Boston College
Harald Hammarström, Max Planck Institute for the Science of Human History
David R. Mortensen, Carnegie Mellon University
Saliha Muradoglu, The Australian National University (ANU)
Éric Le Ferrand, Charles Darwin University, Université Grenoble Alpes
He Zhou (Indiana University Bloomington)
Ezequiel Koile, HSE University, Max Planck Institute for the Science of Human History
Bonaventure Dossou, Jacobs University Bremen
Chris C. Emezue, Technical University Munich
William Lane, Charles Darwin University
John Mansfield, University of Melbourne
Vitaly Protasov, AIR Institute
Svetlana Toldova, HSE University
Daan van Esch, Google Research
Oleg Serikov, HSE University, AIRI, MIPT, RAS Linguistics
Ekaterina Voloshina, HSE University, AIRI
Elena Klyachko, Institute of Linguistics RAS, HSE University
Ekaterina Vylomova, University of Melbourne
Tatiana Shavrina, AIRI, SberDevices
Valentin Malykh, Huawei
Timofey Arkhangelskiy, University of Hamburg
Vladislav Mikhailov, SberDevices, HSE University
George Moroz, HSE University
Daniil Ignatiev, HSE University

Invited Speakers

Antonios Anastasopoulos, George Mason University
Steven Bird, Charles Darwin University

Table of Contents

A Finite State Aproach to Interactive Transcription
William Lane and Steven Bird 1

Corpus-Guided Contrast Sets for Morphosyntactic Feature Detection in Low-Resource English Varieties
Tessa Masis, Anissa Neal, Lisa Green and Brendan O'Connor 11

Machine Translation Between High-resource Languages in a Language Documentation Setting
Katharina Kann, Abteen Ebrahimi, Kristine Stenzel and Alexis Palmer 26

Automatic Detection of Borrowings in Low-Resource Languages of the Caucasus: Andic branch
Konstantin Zaitsev and Anzhelika Minchenko 34

The interaction between cognitive ease and informativeness shapes the lexicons of natural languages
Thomas Brochhagen and Gemma Boleda 42

The first neural machine translation system for the Erzya language
David Dale 45

Abui Wordnet: Using a Toolbox Dictionary to develop a wordnet for a low-resource language
Frantisek Kratochvil and Luís Morgado da Costa 54

How to encode arbitrarily complex morphology in word embeddings, no corpus needed
Lane Schwartz, Coleman Haley and Francis Tyers 64

Predictive Text for Agglutinative and Polysynthetic Languages
Sergey Kosyak and Francis Tyers 77