SIGTYP 2023

The 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP

Proceedings of the Workshop

May 6, 2023

The SIGTYP organizers gratefully acknowledge the support from the following sponsors.

Supported By

Google

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-959429-56-2

Introduction

SIGTYP 2023 is the fifth edition of the workshop for typology-related research and its integration into multilingual Natural Language Processing (NLP). The workshop is co-located with the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023), which takes place in Dubrovnik, Croatia. This year our workshop features a shared task on cognate and derivative detection for low-resourced languages.

Encouraged by the 2019 – 2022 workshops, the aim of the fifth edition of SIGTYP workshop is to act as a platform and a forum for the exchange of information between typology-related research, multilingual NLP, and other research areas that can lead to the development of truly multilingual NLP methods. The workshop is specifically aimed at raising awareness of linguistic typology and its potential in supporting and widening the global reach of multilingual NLP, as well as at introducing computational approaches to linguistic typology. It fosters research and discussion on open problems, not only within the active community working on cross- and multilingual NLP but also inviting input from leading researchers in linguistic typology.

The workshop provides focused discussions on a range of topics, including the following:

1. Integration of typological features in language transfer and joint multilingual learning. In addition to established techniques such as "selective sharing", are there alternative ways to encode heterogeneous external knowledge in machine learning algorithms?
2. Development of unified taxonomy and resources.Building universal databases and models to facilitate understanding and processing of diverse languages.
3. Automatic inference of typological features. The pros and cons of existing techniques (e.g. heuristics derived from morphosyntactic annotation, propagation from features of other languages, supervised Bayesian and neural models) and discussion on emerging ones.
4. Typology and interpretability. The use of typological knowledge for interpretation of hidden representations of multilingual neural models, multilingual data generation and selection, and typological annotation of texts.
5. Improvement and completion of typological databases. Combining linguistic knowledge and automatic data-driven methods towards the joint goal of improving the knowledge on cross-linguistic variation and universals.
6. Linguistic diversity and universals. Challenges of cross-lingual annotation. Which linguistic phenomena or categories should be considered (near-)universal? How should they be annotated?
7. Bringing technology to document and revitalize endangered languages. Improving model performance and documentation of under-resourced and endangered languages using typological databases, multilingual models and data from high-resource languages.

The final program of SIGTYP contains 2 keynote talks, 3 shared task papers, 12 archival papers, and 5 extended abstracts. This workshop would not have been possible without the contribution of its program committee, to whom we would like to express our gratitude. We should also thank Ella Rabinovich and Natalia Levshina for kindly accepting our invitation as invited speakers. The workshop is sponsored by Google. Please find more details on the SIGTYP 2023 website: https://sigtyp.github.io/ws2023-sigtyp.html

Organizing Committee

Workshop Organizers

Lisa Beinborn, Vrije Universiteit Amsterdam
Koustava Goswami, Adobe Research
Saliha Muradoğlu, Australian National University
Alexey Sorokin, Moscow State University
Ritesh Kumar, Dr. Bhimrao Ambedkar University
Andreas Shcherbakov, The University of Melbourne
Edoardo M. Ponti, The University of Edinburgh
Ryan Cotterell, ETH Zürich
Ekaterina Vylomova, The University of Melbourne

Program Committee

Program Chairs

Emily Ahn, University of Washington
Miriam Butt, University of Konstanz
Daan van Esch, Google AI
Elisabetta Ježek, University of Pavia
Paola Merlo, University of Geneva
Joakim Nivre, Uppsala University
Robert Östling, Stockholm University
Ivan Vulić, The University of Cambridge
Richard Sproat, Google Japan
Željko Agić, Corti
Edoardo Ponti, The University of Edinburgh
Alexey Sorokin, Moscow State University
Andrey Shcherbakov, The University of Melbourne
Tanja Samardžić, University of Zurich
Aryaman Arora, Georgetown University
Samopriya Basu, The University of North Carolina at Chapel Hill
Badr M. Abdullah, Saarland University
Guglielmo Inglese, KU Leuven
Olga Zamaraeva, University of Washington
Borja Herce, University of Zurich
Michael Hahn, Stanford University
Giuseppe Celano, Leipzig University
Richard Futrell, University of California, Irvine
Gerhard Jäger, Universität Tübingen
Eitan Grossman, Hebrew University of Jerusalem
Johann-Mattis List, University of Passau
Miryam de Lhoneux, KU Leuven
Giulia Venturi, Istituto di Linguistica Computazionale Antonio Zampolli
Kristen Howell, University of Washington
Barend Beekhuizen, University of Toronto
Claire Bowern, Yale University
Thomas Proisl, University of Erlangen-Nuremberg
Michael Regan, University of Washington

Table of Contents

Program

Saturday, May 6, 2023

08:50 - 09:00 *Opening Remarks*

09:00 - 09:50 *Keynote by Ella Rabinovich*

09:50 - 10:20 *Cross-lingual Transfer*

Identifying the Correlation Between Language Distance and Cross-Lingual Transfer in a Multilingual Representation Space
Fred Philippy, Siwen Guo and Shohreh Haddadan

Gradual Language Model Adaptation Using Fine-Grained Typology
Marcell Richard Fekete and Johannes Bjerva

Cross-lingual Transfer Learning with Persian
Sepideh Mollanorozy, Marc Tanti and Malvina Nissim

10:20 - 10:35 *Cross-Lingual Transfer of Cognitive Complexity (Findings)*

10:35 - 11:15 *Break*

11:15 - 11:30 *Does Transliteration Help Multilingual Language Modeling Long (Findings)*

11:30 - 12:30 *Multilinguality*

Multilingual BERT has an Accent: Evaluating English Influences on Fluency in Multilingual Models
Isabel Papadimitriou, Kezia Lopez and Dan Jurafsky

The Denglisch Corpus of German-English Code-Switching
Doreen Osmelak and Shuly Wintner

Trimming Phonetic Alignments Improves the Inference of Sound Correspondence Patterns from Multilingual Wordlists
Frederic Blum and Johann-Mattis List

Saturday, May 6, 2023 (continued)

On the Nature of Discrete Speech Representations in Multilingual Self-supervised Models
Badr M. Abdullah, Mohammed Maqsood Shaik and Dietrich Klakow

You Can Have Your Data and Balance It Too: Towards Balanced and Efficient Multilingual Models
Tomasz Limisiewicz, Dan Malkin and Gabriel Stanovsky

12:30 - 12:45 *Evaluating the Diversity, Equity and Inclusion of NLP Technology: A Case Study for Indian Languages (Findings)*

12:45 - 13:00 *A Large-Scale Multilingual Study of Visual Constraints on Linguistic Selection of Descriptions (Findings)*

13:00 - 14:15 *Lunch (with Linguistic Trivia at 13:45–14:15)*

14:15 - 15:05 *Keynote by Natalia Levshina*

15:05 - 15:50 *Linguistic Complexity*

Information-Theoretic Characterization of Vowel Harmony: A Cross-Linguistic Study on Word Lists
Julius Steuer, Johann-Mattis List, Badr M. Abdullah and Dietrich Klakow

A Crosslinguistic Database for Combinatorial and Semantic Properties of Attitude Predicates
Deniz Özyıldız, Ciyang Qing, Floris Roelofsen, Maribel Romero and Wataru Uegaki

Revisiting Dependency Length and Intervener Complexity Minimisation on a Parallel Corpus in 35 Languages
Andrew Thomas Dyer

15:50 - 16:10 *Break*

16:10 - 16:40 *Shared task on Cognate and Derivative Detection For Low-Resourced Languages*

Findings of the SIGTYP 2023 Shared task on Cognate and Derivative Detection For Low-Resourced Languages
Priya Rani, Koustava Goswami, Adrian Doyle, Theodorus Fransen, Bernardo Stearns and John P. McCrae

Saturday, May 6, 2023 (continued)

ÚFAL Submission for SIGTYP Supervised Cognate Detection Task
Tomasz Limisiewicz

CoToHiLi at SIGTYP 2023: Ensemble Models for Cognate and Derivative Words Detection
Liviu P. Dinu, Ioan-Bogdan Iordache and Ana Sabina Uban

16:40 - 16:45 *Break*

16:45 - 18:05 *Syntax and Morphology*

Grambank's Typological Advances Support Computational Research on Diverse Languages
Hannah J. Haynie, Damián Blasi, Hedvig Skirgård, Simon J. Greenhill, Quentin D. Atkinson and Russell D. Gray

Language-Agnostic Measures Discriminate Inflection and Derivation
Coleman Haley, Edoardo M. Ponti and Sharon Goldwater

Does Topological Ordering of Morphological Segments Reduce Morphological Modeling Complexity? A Preliminary Study on 13 Languages
Andreas Shcherbakov and Ekaterina Vylomova

Multilingual End-to-end Dependency Parsing with Linguistic Typology knowledge
Chinmay Choudhary and Colm O'riordan

Using Modern Languages to Parse Ancient Ones: a Test on Old English
Luca Brigada Villa and Martina Giarda

Corpus-based Syntactic Typological Methods for Dependency Parsing Improvement
Diego Alves, Božo Bekavac, Daniel Zeman and Marko Tadić

18:05 - 18:10 *Best Paper Awards, Closing*