SIGTYP 2024

The 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP

Proceedings of the Workshop

March 22, 2024

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 979-8-89176-071-4

Introduction

SIGTYP 2024 is the sixth edition of the workshop for typology-related research and its integration into multilingual Natural Language Processing (NLP). The workshop is co-located with the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2024), which takes place in St Julian's, Malta. This year our workshop features a shared task on Word Embedding Evaluation for Ancient and Historical Languages.

Encouraged by the 2019 – 2024 workshops, the aim of the sixth 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 encoding 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 universal? How should they be annotated?
7. Language-specific studies to support or contradict universals. Framing a study on 1-3 languages that would shed more light on common linguistic structures and properties.

The final program of SIGTYP contains 2 keynote talks, 5 shared task papers, 11 archival papers, and 2 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 Chris Bentz and Ximena Gutierrez-Vasques for kindly accepting our invitation as invited speakers. The workshop is sponsored by Google. Please find more details on the SIGTYP 2024 website: https://sigtyp.github.io/ws2024-sigtyp.html

Organizing Committee

Workshop Organizers

Michael Hahn, Saarland University
Alexey Sorokin, Yandex and Lomonosov Moscow State University
Ritesh Kumar, Dr. Bhimrao Ambedkar University
Andreas Shcherbakov, University of Melbourne
Yulia Otmakhova, The University of Melbourne
Jinrui Yang, The University of Melbourne
Oleg Serikov, King Abdullah University of Science and Technology
Priya Rani, University of Galway
Edoardo M. Ponti, University of Edinburgh
Saliha Muradoğlu, Australian National University
Rena Gao, University of Melbourne
Ryan Cotterell, Swiss Federal Institute of Technology
Ekaterina Vylomova, The University of Melbourne
Oksana Dereza, University of Galway

Program Committee

Program Chairs

Michael Hahn, Saarland University
Alexey Sorokin, Yandex and Lomonosov Moscow State University
Ritesh Kumar, Dr. Bhimrao Ambedkar University
Andreas Shcherbakov, University of Melbourne
Yulia Otmakhova, The University of Melbourne
Jinrui Yang, The University of Melbourne
Oleg Serikov, King Abdullah University of Science and Technology
Priya Rani, University of Galway
Edoardo M. Ponti, University of Edinburgh
Saliha Muradoğlu, Australian National University
Rena Gao, University of Melbourne
Ryan Cotterell, Swiss Federal Institute of Technology
Ekaterina Vylomova, The University of Melbourne

Reviewers

Badr M. Abdullah, Aryaman Arora

Barend Beekhuizen, Claire Bowern, Miriam Butt

Giuseppe G. A. Celano

Richard Futrell

Rena Wei Gao

Borja Herce, Kristen Howell

Elisabetta Jezek, Gerhard Jäger

Ritesh Kumar, Kemal Kurniawan

Johann-Mattis List

Saliha Muradoglu

Joakim Nivre

Yulia Otmakhova, Robert Östling

Edoardo Ponti

Priya Rani

Oleg Serikov, Andreas Shcherbakov, Alexey Sorokin, Richard Sproat

Daan Van Esch, Giulia Venturi, Ivan Vulić, Ekaterina Vylomova

Jinrui Yang

Olga Zamaraeva

Keynote Talk: Zipfian laws across diverse languages

Christian Bentz
University of Tübingen
2024-03-22 09:00:00 – Room: **Radisson, Marie Loise 1**

Abstract: There are few - if any - universals which hold across all known languages. Promising candidates are quantitative laws such as Zipf's law of word frequencies and Zipf's law of abbreviation. This talk will review some of the current research into these laws from a cross-linguistic perspective. This includes a discussion of the methodological challenges when working with diverse languages, modalities, and writing systems, as well as the controversial question how "meaningful" the laws are given random baselines. Finally, an avenue for further research is explored: the challenge of defining a statistical fingerprint for human languages.

Bio: Christian Bentz is currently an Assistant Professor at the Department of General Linguistics, University of Tübingen. He received his PhD in Computation, Cognition, and Language from the University of Cambridge. His research interests include information theory, quantitative linguistics, language typology, and language evolution.

Keynote Talk: Text-based typology for modeling linguistic diversity in NLP

Ximena Gutierrez-Vasques
UNAM, Mexico City
2024-03-22 13:45:00 – Room: **Radisson, Marie Loise 1**

Abstract: During this presentation, I will elaborate on the importance of capturing the immense diversity inherent in natural languages. This extends beyond advancing language technologies; it also serves to answer interdisciplinary research questions and enrich the exploration of linguistic typology through computational lenses. By harnessing textual data and unsupervised NLP techniques, we can induce typological knowledge, thereby facilitating the expansion of existing typological databases and facilitating more comprehensive language comparisons for various NLP applications.
I will illustrate these concepts through a case study that demonstrates how simple techniques such as subword tokenization and the analysis of multilingual text corpora enable the study of the morphological typology of languages and the complexity of their morphological systems. We will also examine the implications and constraints associated with these methodologies.

Bio: Ximena Gutierrez-Vasques is a computational linguist with an interdisciplinary focus to deepen the study of human language. Her lines of research cover multilingual NLP, computational morphology, and NLP under-resourced languages of the Americas. She was a postdoctoral researcher at the University of Zürich where she specialized in approaches for modeling linguistic complexity and typology using text corpora and inspired by information theory. She recently joined an interdisciplinary research center in Mexico (CEIICH, UNAM), where she works in the interface between humanities and the field of AI.

Table of Contents

Program

Friday, March 22, 2024

08:50 - 09:00 *Opening Remarks*

09:00 - 10:00 *Keynote by Christian Bentz*

10:00 - 10:30 *Low-Resource NLP*

GUIDE: Creating Semantic Domain Dictionaries for Low-Resource Languages
Jonathan Janetzki, Gerard De Melo, Joshua Nemecek and Daniel Lee Whitenack

Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens
Nay San, Georgios Paraskevopoulos, Aryaman Arora, Xiluo He, Prabhjot Kaur, Oliver Adams and Dan Jurafsky

10:30 - 11:15 *Break*

11:15 - 12:20 *Typology and Language Comparison*

A New Dataset for Tonal and Segmental Dialectometry from the Yue- and Pinghua-Speaking Area
Ho Wang Matthew Sung, Jelena Prokic and Yiya Chen

Language Atlas of Japanese and Ryukyuan (LAJaR): A Linguistic Typology Database for Endangered Japonic Languages
Kanji Kato, So Miyagawa and Natsuko Nakagawa

A Call for Consistency in Reporting Typological Diversity
Wessel Poelman, Esther Ploeger, Miryam De Lhoneux and Johannes Bjerva

Are Sounds Sound for Phylogenetic Reconstruction?
Luise Häuser, Gerhard Jäger, Johann-Mattis List, Taraka Rama and Alexandros Stamatakis

The Typology of Ellipsis: A Corpus for Linguistic Analysis and Machine Learning Applications
Damir Cavar, Ludovic Mompelat and Muhammad S. Abdo

12:30 - 13:45 *Lunch*

Friday, March 22, 2024 (continued)

13:45 - 14:45 *Keynote by Ximena Gutierrez-Vasques*

14:45 - 15:30 *Multilinguality*

GTNC: A Many-To-One Dataset of Google Translations from NewsCrawl
Damiaan J W Reijnaers and Charlotte Pouw

ModeLing: A Novel Dataset for Testing Linguistic Reasoning in Language Models
Nathan Andrew Chi, Teodor Malchev, Riley Kong, Ryan Andrew Chi, Lucas Huang, Ethan A Chi, R. Thomas McCoy and Dragomir Radev

Compounds in Universal Dependencies: A Survey in Five European Languages
Emil Svoboda and Magda Ševčíková

Sociolinguistically Informed Interpretability: A Case Study on Hinglish Emotion Classification
Kushal Tatariya, Heather Lent, Johannes Bjerva and Miryam De Lhoneux

15:30 - 16:15 *Break*

16:15 - 17:00 *Shared task on Word Embedding Evaluation for Ancient and Historical Languages*

Findings of the SIGTYP 2024 Shared Task on Word Embedding Evaluation for Ancient and Historical Languages
Oksana Dereza, Adrian Doyle, Priya Rani, Atul Ojha, Pádraic Moran and John McCrae

TartuNLP @ SIGTYP 2024 Shared Task: Adapting XLM-RoBERTa for Ancient and Historical Languages
Aleksei Dorkin and Kairit Sirts

Heidelberg-Boston @ SIGTYP 2024 Shared Task: Enhancing Low-Resource Language Analysis With Character-Aware Hierarchical Transformers
Frederick Riemenschneider and Kevin Krahn

UDParse @ SIGTYP 2024 Shared Task : Modern Language Models for Historical Languages
Johannes Heinecke

Friday, March 22, 2024 (continued)

Allen Institute for AI @ SIGTYP 2024 Shared Task on Word Embedding Evaluation for Ancient and Historical Languages
Lester James V. Miranda

17:00 - 17:30 *Typology and Human Language Processing*

Syntactic dependency length shaped by strategic memory allocation
Weijie Xu and Richard Futrell

A Computational Model for the Assessment of Mutual Intelligibility Among Closely Related Languages
Jessica Nieder and Johann-Mattis List

Predicting Mandarin and Cantonese Adult Speakers' Eye-Movement Patterns in Natural Reading
LI Junlin, Yu-Yin Hsu, Emmanuele Chersoni and Bo Peng

17:35 - 17:50 *Computational Morphology and Lexicography Modeling of Modern Standard Arabic Nominals (EACL Findings)*

17:50 - 18:00 *Best Paper Awards, Closing*