SIGTYP 2022

The 4th Workshop on Computational Typology and Multilingual NLP

Proceedings of the Workshop

July 14, 2022

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 978-1-955917-93-3

Introduction

SIGTYP 2022 is the fourth edition of the workshop for typology-related research and its integration into multilingual Natural Language Processing (NLP). The workshop is co-located with the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2022), which takes place in Seattle, Washington. This year our workshop features a shared task on prediction of cognate reflexes.

The final program of SIGTYP contains 3 keynote talks, 5 shared task papers, 6 archival papers, and 4 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 Kristen Howell, Isabel Papadimitriou, and Graham Neubig for kindly accepting our invitation as invited speakers. The workshop is generously sponsored by Google. Please find more details on the SIGTYP 2022 website: https://sigtyp.github.io/ws2022-sigtyp.html

Organizing Committee

Workshop Organizers

Ekaterina Vylomova, The University of Melbourne
Hila Gonen, University of Washington
Jonas Pfeiffer, New York University
Edoardo Ponti, The University of Edinburgh
Alexey Sorokin, Moscow State University
Andrey Shcherbakov, The University of Melbourne
Sabrina Mielke, Johns Hopkins University
Gabriella Lapesa, University of Stuttgart
Harald Hammarström, Uppsala University
Pranav A, Dayta AI
Ryan Cotterell, ETH Zürich
Ritesh Kumar, Dr. Bhimrao Ambedkar University

Program Committee

Program Chairs

Johannes Bjerva, Aalborg University
Emily Ahn, University of Washington
Miriam Butt, University of Konstanz
John Mansfield, The University of Melbourne
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
Agnieszka Falenska, University of Stuttgart
Edoardo Ponti, The University of Edinburgh
Alexey Sorokin, Moscow State University
Andrey Shcherbakov, The University of Melbourne
Tanja Samardžić, University of Zurich
Kemal Kurniawan, The University of Melbourne
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
Nianwen Xue, Brandeis University
Borja Herce, University of Zurich
Chinmay Choudhary, National University of Ireland, Galway
Bradley Hauer, The University of Alberta
Michael Hahn, Stanford University

Keynote Talk: Grammar Inference for Local Languages. Leveraging Typology for Automatic Grammar Generation

Kristen Howell
University of Washington

Abstract: In this talk I will describe the benefit of implemented grammars as well as the challenges involved in creating them. I present an inference system that can be used to automatically generate such grammars on the basis of interlinear glossed text (IGT) corpra. The inference system, called BASIL – Building Analyses from Syntactic Inference in Local Languages, leverages typologically informed heuristics to infer syntactic and morphological information from linguistic corpora to select analyses that model the language. We will engage with the question of whether and to what extent typological features are apparent in IGT data and how effectively grammars generated with these features can model human language.

Bio: Kristen Howell is a data scientist at LivePerson Inc. in Seattle, Washington. Her research interests range from grammar engineering and grammar inference to conversational NLP. Throughout this research, the common thread is multilingual NLP across typologically diverse languages. Kristen received her PhD from the University of Washington in 2020, where she engaged with typological literature to develop technology for automatically generating grammars for local languages. Recent work at LivePerson has focused on multilingual NLP, leveraging deep learning techniques for conversational AI.

Keynote Talk: Graham Neubig's Invited Talk

Graham Neubig
Carnegie Mellon University

Abstract: Will be announced later.

Bio: Graham is an associate professor at the Language Technologies Institute of Carnegie Mellon University. His research focuses on multilingual natural language processing, natural language interfaces to computers, and machine learning methods for NLP, with the final goal of every person in the world being able to communicate with each-other, and with computers in their own language. He also contributes to making NLP research more accessible through open publishing of research papers, advanced NLP course materials and video lectures, and open-source software, all of which are available on his web site.

Keynote Talk: Learning from our Differences. How Typologically Distinct Modalities of Data Help Demystify Language Models

Isabel Papadimitriou
Stanford University

Abstract: Looking beyond a single language, or to non-linguistic forms of data, can yield new insights into linguistic representation and use in language models. This talk will explore this theme in two threads: Firstly, what can we learn from passing non-linguistic data through language models? From natural modalities like music to controlled synthetic parentheses languages, we can use datasets with different underlying structures to explore knowledge in language model transfer learning. Knowing the structures in this data lets us understand if and how different features are acquired and generalized in language model training. Secondly, we will look at how typologically-aware analysis can help us understand joint multilingual representation in language models, with experiments that focus on agenthood and case in different languages in multilingual models. The typological diversity of agenthood gives us a handle into understanding how representations can be shared and also separated between languages. Examining language models at the points where diverse data differs – and systematically knowing the ways in which data differs – offers a useful window into how linguistic knowledge is represented in language models.

Bio: Isabel is a PhD student at Stanford in the Natural Language Processing group, advised by Dan Jurafsky. Her main research focuses on exploring the linguistic basis of computational language methods. She likes to focus on how language is both a discrete symbolic system and a system of continuous gradations, and exploring the limits of how large neural models can emcompass this combination. She is very interested in looking at the behavior of large language models in multilingual settings, and analyizing the ways in which languages and dialects co-occur and interfere in single models.

Table of Contents

Program

Thursday, July 14, 2022

08:30 - 08:40 *Opening Remarks*

08:40 - 09:30 *Grammar Inference for Local Languages: Leveraging Typology for Automatic Grammar Generation (Keynote by Kristen Howell)*

09:30 - 10:00 *Multilingual Representations (Long Talks)*

Multilingualism Encourages Recursion: a Transfer Study with mBERT
Andrea Gregor De Varda and Roberto Zamparelli

Cross-linguistic Comparison of Linguistic Feature Encoding in BERT Models for Typologically Different Languages
Yulia Otmakhova, Karin Verspoor and Jey Han Lau

10:00 - 10:10 *Break*

10:10 - 11:10 *Typology (Short Talks)*

Word-order Typology in Multilingual BERT: A Case Study in Subordinate-Clause Detection
Dmitry Nikolaev and Sebastian Pado

Investigating Information-Theoretic Properties of the Typology of Spatial Demonstratives
Sihan Chen, Richard Futrell and Kyle Mahowald

Tweaking UD Annotations to Investigate the Placement of Determiners, Quantifiers and Numerals in the Noun Phrase
Luigi Talamo

How Universal is Metonymy? Results from a Large-Scale Multilingual Analysis
Temuulen Khishigsuren, Gábor Bella, Thomas Brochhagen, Daariimaa Marav, Fausto Giunchiglia and Khuyagbaatar Batsuren

Typological Word Order Correlations with Logistic Brownian Motion
Kai Hartung, Gerhard Jäger, Sören Gröttrup and Munir Georges

11:10 - 12:00 *Graham Neubig's Keynote Talk*

Thursday, July 14, 2022 (continued)

12:00 - 13:30 *Lunch*

13:30 - 14:50 *Shared Task: Prediction of Cognate Reflexes*

The SIGTYP 2022 Shared Task on the Prediction of Cognate Reflexes
Johann-Mattis List, Ekaterina Vylomova, Robert Forkel, Nathan Hill and Ryan Cotterell

Bayesian Phylogenetic Cognate Prediction
Gerhard Jäger

Mockingbird at the SIGTYP 2022 Shared Task: Two Types of Models forthe Prediction of Cognate Reflexes
Christo Kirov, Richard Sproat and Alexander Gutkin

A Transformer Architecture for the Prediction of Cognate Reflexes
Giuseppe Celano

Approaching Reflex Predictions as a Classification Problem Using Extended Phonological Alignments
Tiago Tresoldi

14:50 - 15:20 *Linguistic Trivia*

15:20 - 15:30 *Break*

15:30 - 16:20 *Learning from our Differences: How Typologically Distinct Modalities of Data Help Demystify Language Models (Keynote by Isabel Papadimitriou)*

16:20 - 17:00 *Databases and Corpora*

A Database for Modal Semantic Typology
Qingxia Guo, Nathaniel Imel and Shane Steinert-Threlkeld

PaVeDa - Pavia Verbs Database: Challenges and Perspectives
Chiara Zanchi, Silvia Luraghi and Claudia Roberta Combei

Thursday, July 14, 2022 (continued)

ParaNames: A Massively Multilingual Entity Name Corpus
Jonne Sälevä and Constantine Lignos

17:00 - 17:10 *Best Paper Awards, Closing*