ACL 2019

TyP-NLP: The Workshop on Typology for Polyglot NLP

Proceedings of the First Workshop

August 1, 2019
Florence, Italy


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 978-1-950737-29-1

Introduction

The TyP-NLP workshop will be the first dedicated venue for typology-related research and its integration into multilingual Natural Language Processing (NLP). The workshop will be hosted by the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019) in Florence, Italy.

The ultimate goal of TyP-NLP is the development of robust language technology applicable across the world's languages. Long overdue, the workshop is specifically aimed at raising awareness of linguistic typology and its potential in supporting and widening the global reach of multilingual NLP. It will foster research and discussion on open problems relevant to the multilingual NLP community, but it will also invite input from leading researchers in linguistics and cognitive sciences.

The final program of TyP-NLP contains 5 keynote talks and 18 accepted posters, selected among a large number of non-archival submissions. This workshop would not have been possible without the hard work of its program committee. We would like to express our gratitude to them for writing meticulous reviews in a very constrained span of time. We should also thank our invited speakers, Isabelle Augenstein, Emily Bender, Balthasar Bickel, Jason Eisner, and Sabine Stoll, for their irreplaceable contribution to our program. The workshop is generously sponsored by Google and by the European Research Council (ERC) Consolidator Grant LEXICAL (no. 648909).

Find more details on the TyP-NLP 2019 website: `https://typology-and-nlp.github.io/`

Organizing Committee:

Haim Dubossarsky, University of Cambridge
Arya D. McCarthy, Johns Hopkins University
Edoardo M. Ponti, University of Cambridge
Ivan Vulić, University of Cambridge
Ekaterina Vylomova, University of Melbourne

Steering Committee:

Yevgeni Berzak, MIT
Ryan Cotterell, University of Cambridge
Manaal Faruqui, Google
Eitan Grossman, Hebrew University of Jerusalem
Anna Korhonen, University of Cambridge
Thierry Poibeau, Centre national de la recherche scientifique (CNRS)
Roi Reichart, Technion - Israel Institute of Technology

Program Committee:

Željko Agić, University of Copenhagen
Ehsaneddin Asgari, University of California, Berkeley
Barend Beekhuizen, University of Toronto
Christian Bentz, University of Zürich
Richard Futrell, UC Irvine
Edward Gibson, MIT
Giuseppe Celano, Leipzig University
Mans Hulden, University of Colorado
Gerhard Jäger, University of Tuebingen
Katharina Kann, NYU
Douwe Kiela, Facebook
Silvia Luraghi, University of Pavia
John Mansfield, University of Melbourne
David R. Mortensen, Carnegie Mellon University
Phoebe Mulcaire, University of Washington
Jason Naradowsky, Johns Hopkins
Joakim Nivre, Uppsala University
Robert Östling, Stockholm University
Ella Rabinovich, University of Toronto
Michael Regan, University of New Mexico
Tanja Samardzic, University of Zurich
Hinrich Schütze, University of Munich
Sabine Stoll, University of Zurich
Jörg Tiedemann, University of Helsinki
Reut Tsarfaty, Open University of Israel
Yulia Tsvetkov, Carnegie Mellon University

Daan van Esch, Google
Shuly Wintner, University of Haifa

Invited Speakers:

Isabelle Augenstein, University of Copenhagen
Emily M. Bender, University of Washington
Balthasar Bickel, University of Zurich
Jason Eisner, Johns Hopkins University
Sabine Stoll, University of Zurich

Conference Program

Thursday, August 1, 2019

7:30–8:45 ***Breakfast***

8:45–9:00 ***Opening Remarks***

9:00–9:45 *Invited Talk*
Emily M. Bender

9:45–10:30 *Invited Talk*
Jason Eisner

10:30–11:00 ***Coffee Break***

11:00–12:30 **Poster Session**

Unsupervised Document Classification in Low-resource Languages for Emergency Situations
Nidhi Vyas, Eduard Hovy and Dheeraj Rajagopal

Polyglot Parsing for One Thousand and One Languages (and Then Some)
Ali Basirat, Miryam de Lhoneux, Artur Kulmizev, Murathan Kurfalı, Joakim Nivre and Robert Östling

Dissecting Treebanks to Uncover Typological Trends. a Multilingual Comparative Approach
Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni and Giulia Venturi

What Do Multilingual Neural Machine Translation Models Learn about Typology?
Ryokan Ri and Yoshimasa Tsuruoka

Syntactic Typology from Plain Text Using Language Embeddings
Taiqi He and Kenji Sagae

Typological Feature Prediction with Matrix Completion
Annebeth Buis and Mans Hulden

Feature Comparison across Typological Resources
Tifa de Almeida

Using Typological Information in WALS to Improve Grammar Inference
Youyun Zhang, Tifa de Almeida, Kristen Howell and Emily M. Bender

Thursday, August 1, 2019 (continued)

Cross-linguistic Robustness of Infant Word Segmentation Algorithms: Oversegmenting Morphologically Complex Languages
Georgia R. Loukatou

Cross-linguistic Semantic Tagset for Case Relationships
Ritesh Kumar, Bornini Lahiri and Atul Kr. Ojha

AfricaSign - a Crowd-sourcing Platform for Lexical Documentation of African Sign Languages
Abdelhadi Soudi, Kristof Van Laerhoven and Elmosta Bou-Souf

Predicting Continuous Vowel Spaces in the Wilderness
Emily Ahn and David R. Mortensen

Towards a Multi-view Language Representation: A Shared Space of Discrete and Continuous Language Features
Arturo Oncevay, Barry Haddow and Alexandra Birch

Cross-lingual CCG Induction: Learning Categorial Grammars via Parallel Corpora
Kilian Evang

Towards a Computationally-Relevant Linguistic Typology for Polyglot/Multilingual NLP
Ada Wan

Transfer Learning for Cognate Identification in Low-Resource Languages
Eliel Soisalon-Soininen and Mark Granroth-Wilding

Contextualization of Morphological Inflection
Ekaterina Vylomova, Ryan Cotterell, Timothy Baldwin, Trevor Cohn and Jason Eisner

Towards Unsupervised Extraction of Linguistic Typological Features from Language Descriptions
Søren Wichmann and Taraka Rama

12:30–14:00 ***Lunch***

14:00–14:45 *Invited Talk*
Balthasar Bickel

14:45–15:30 *Invited Talk*
Sabine Stoll

Thursday, August 1, 2019 (continued)

15:30–16:00 ***Coffee Break***

16:00–16:45 *Invited Talk*
Isabelle Augenstein

16:45–17:30 ***Panel Discussion***

17:30–17:45 ***Best Paper Announcement and Closing Remarks***

Biography of the Speakers

Isabelle Augenstein is a tenure-track assistant professor at the University of Copenhagen, Department of Computer Science since July 2017, affiliated with the Copenhagen NLP group and the Machine Learning Section, and work in the general areas of Statistical Natural Language Processing and Machine Learning. Her main research interests are weakly supervised and low-resource learning with applications including information extraction, machine reading and fact checking.

Emily M. Bender's primary research interests are in multilingual grammar engineering, the study of variation, both within and across languages, and the relationship between linguistics and computational linguistics. She is the LSA's delegate to the ACL. Her 2013 book *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax* aims to present linguistic concepts in an manner accessible to NLP practitioners.

Jason Eisner works on machine learning, combinatorial algorithms, probabilistic models of linguistic structure, and declarative specification of knowledge and algorithms. His work addresses the question, "How can we appropriately formalize linguistic structure and discover it automatically?"

Balthasar Bickel aims at understanding the diversity of human language with rigorously tested causal models, i.e. at answering the question what's where why in language. What structures are there, and how exactly do they vary? Engaged in both linguistic fieldwork and statistical modeling, he focuses on explaining universally consistent biases in the diachrony of grammar properties, biases that are independent of local historical events.

Sabine Stoll questions how children can cope with the incredible variation exhibited in the approximately 6000–7000 languages spoken around the world. Her main focus is the interplay of innate biological factors (such as the capacity for pattern recognition and imitation) with idiosyncratic and culturally determined factors (such as for instance type and quantity of input). Her approach is radically empirical, based first and foremost on the quantitative analysis of large corpora that record how children learn diverse languages.