SIGTYP 2020

Second Workshop on Computational Research in Linguistic Typology

Proceedings of the Workshop

November 19, 2020
Online

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 978-1-952148-73-6

Introduction

SIGTYP 2020 is the second edition of the workshop for typology-related research and its integration into multilingual Natural Language Processing (NLP), which was inaugurated last year at ACL 2019. The workshop is co-located with the 2020 Conference on Empirical Methods in Natural Language Processing, which takes place virtually this year. Our workshop includes a shared task on typological feature prediction, which saw the participation of 5 teams for a total of 8 systems submitted.

The final program of SIGTYP contains 5 keynote talks, 6 shared task papers, 11 extended abstracts, selected among a large number of non-archival submissions, and 3 papers from "Findings of ACL: EMNLP 2020". This workshop would not have been possible without the hard work of its program committee, to whom we would like to express our gratitude. We should also thank our invited speakers, Miriam Butt, Yulia Tsvetkov, Richard Sproat, Bill Croft, Harald Hammarström, 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).

Please find more details on the SIGTYP 2020 website: `https://sigtyp.github.io/ws2020.html`

Organizing Committee:

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

Steering Committee:

Anna Korhonen, University of Cambridge
Eitan Grossman, Hebrew University of Jerusalem
Roi Reichart, Technion - Israel Institute of Technology
Ryan Cotterell, ETH Zurich and University of Cambridge
Yevgeni Berzak, MIT

Program Committee:

Johannes Bjerva, University of Copenhagen
Shuly Wintner, University of Haifa
Giuseppe Celano, Leipzig University
John Mansfield, University of Melbourne
Robert Östling, Stockholm University
Jörg Tiedemann , University of Helsinki
Željko Agić, Corti
Daan van Esch, Google AI
Tanja Samardžić, University of Zurich
Ella Rabinovich, University of Toronto
Barend Beekhuizen, University of Toronto
Nidhi Vyas, Carnegie Mellon University
Kilian Evang, University of Düsseldorf
Emily Ahn, Carnegie Mellon University
Annebeth Buis, The University of Colorado
Giulia Venturi, ILC "Antonio Zampolli"
Ada Wan, University of Zurich
Mark Ellison, Australian National University
Richard Futrell, University of California, Irvine
Michael Regan, University of New Mexico
Elisabetta Ježek, University of Pavia
Kyle Gorman, City University of New York
Joakim Nivre, Uppsala University
Chris Dyer, DeepMind
Emily Bender, University of Washington
Silvia Luraghi, Università di Pavia
Ehsan Asgari, UC Berkeley
Eleanor Chodroff, University of York
Elizabeth Salesky, Johns Hopkins University

Sabrina Mielke, Johns Hopkins University
Georgia Loukatou, Ecole Normale Supérieure

Invited Speakers:

Miriam Butt, University of Konstanz
Yulia Tsvetkov, Carnegie Mellon University
Richard Sproat, Google
Bill Croft, University of New Mexico
Harald Hammarström, Uppsala University

Table of Contents

Conference Program

Thursday, November 19, 2020

8:30–8:40 Opening Session

8:40–10:20 Keynote Session 1

8:40–9:30 *Invited Talk*
Richard Sproat

9:30–10:20 *Invited Talk*
Miriam Butt

10:20–10:30 Coffee Break

10:30–11:35 Shared Task Session

10:30–10:45 *SIGTYP 2020 Shared Task: Prediction of Typological Features*
Johannes Bjerva, Elizabeth Salesky, Sabrina J. Mielke, Aditi Chaudhary, Celano Giuseppe, Edoardo Maria Ponti, Ekaterina Vylomova, Ryan Cotterell and Isabelle Augenstein

10:45–10:55 *KMI-Panlingua-IITKGP @SIGTYP2020: Exploring rules and hybrid systems for automatic prediction of typological features*
Ritesh Kumar, Deepak Alok, Akanksha Bansal, Bornini Lahiri and Atul Kr. Ojha

10:55–11:05 *NEMO: Frequentist Inference Approach to Constrained Linguistic Typology Feature Prediction in SIGTYP 2020 Shared Task*
Alexander Gutkin and Richard Sproat

11:05–11:15 *Predicting Typological Features in WALS using Language Embeddings and Conditional Probabilities: ÚFAL Submission to the SIGTYP 2020 Shared Task*
Martin Vastl, Daniel Zeman and Rudolf Rosa

11:15–11:25 *Imputing typological values via phylogenetic inference*
Gerhard Jäger

11:25–11:35 *NUIG: Multitasking Self-attention based approach to SigTyp 2020 Shared Task*
Chinmay Choudhary

Thursday, November 19, 2020 (continued)

11:40–12:30 Keynote Session 2

11:40–12:30 *Invited Talk*
Harald Hammarström

12:20–12:30 Coffee Break

12:30–13:30 Oral Session 1

12:30–12:40 *DEmA: the Pavia Diachronic Emergence of Alignment database*
Sonia Cristofaro and Guglielmo Inglese

12:40–12:50 *A dataset and metric to evaluate lexical extraction from parallel corpora*
Barend Beekhuizen

12:50–13:00 *Keyword Spotting: A quick-and-dirty method for extracting typological features of language from grammatical descriptions*
Harald Hammarström

13:00–13:10 *SNACS Annotation of Case Markers and Adpositions in Hindi*
Aryaman Arora and Nathan Schneider

13:10–13:20 *Information from Topic Contexts: The Prediction of Aspectual Coding of Verbs in Russian*
Michael Richter and Tariq Yousef

13:20–13:30 *The role of community size and network structure in shaping linguistic diversity: experimental evidence*
Limor Raviv, Antje Mayer and Shiri Lev-Ari

Thursday, November 19, 2020 (continued)

13:30–14:05 Lunch Break

14:05–14:30 Findings Session 1

14:05–14:20 *SimAlign: High Quality Word Alignments without Parallel Training Data using Static and Contextualized Embeddings*
Masoud Jalili Sabet, Philipp Dufter, François Yvon and Hinrich Schütze

14:20–14:30 *Parsing with Multilingual BERT, a Small Corpus, and a Small Treebank*
Ethan C. Chau, Lucy H. Lin and Noah A. Smith

14:30–15:30 Oral Session 2

14:30–14:45 *Is Typology-Based Adaptation Effective for Multilingual Sequence Labelling?*
Ahmet Üstün, Arianna Bisazza, Gosse Bouma and Gertjan van Noord

14:45–15:00 *Multilingual BERT Learns Abstract Case Representations*
Isabel Papadimitriou, Ethan A. Chi, Richard Futrell and Kyle Mahowald

15:00–15:10 *Towards Induction of Structured Phoneme Inventories*
Alexander Gutkin, Martin Jansche and Lucy Skidmore

15:10–15:20 *Uncovering Typological Context-Sensitive Features*
Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni and Giulia Venturi

15:20–15:30 *Multilingual Jointly Trained Acoustic and Written Word Embeddings*
Yushi Hu, Shane Settle and Karen Livescu

Thursday, November 19, 2020 (continued)

15:30–15:45 Findings Session 2

15:30–15:45 *Cross-lingual Alignment Methods for Multilingual BERT: A Comparative Study*
Saurabh Kulshreshtha, José Luis Redondo García and Ching-Yun Chang

15:45–16:20 Afternoon Break

16:20–18:00 Keynote Session 3

16:20–17:10 *Invited Talk*
Yulia Tsvetkov

17:10–18:00 *Invited Talk*
Bill Croft

18:00–18:10 Closing Session

Non-archival Abstracts

Information from topic contexts: the prediction of aspectual coding of verbs in Russian
Michael Richter and Tariq Yousef
Based on Shannon's coding theorem, we predict that aspectual coding asymmetries of verbs in Russian can be predicted by the verbal feature Average Information Content. We employ the novel Topic Context Model that calculates the verbal information content from the number of topics in the target words' larger discourses and their local discourses. In contrast to a previous study, TCM yielded disappointing results in this study which is, as we conclude, mainly due to the small number of local contexts we utilized.

Uncovering Typological Context-Sensitive Features
Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni and Giulia Venturi
This contribution presents the results of a method for typological feature identification in multilingual treebanks. The results are exemplified on Italian and English subject relations. Applications of the method for multilingual dependency parsing evaluation are discussed.

Towards Induction of Structured Phoneme Inventories
Alexander Gutkin, Martin Jansche and Lucy Skidmore
This extended abstract provides a summary of our past and ongoing work on assessing the quality of multilingual phoneme inventories derived from typological resources, inducing phonological inventories using distinctive feature representations from the speech data and the important role phonological typology plays in these approaches.

Is Typology-Based Adaptation Effective for Multilingual Sequence Labelling?
Ahmet Üstün, Arianna Bisazza, Gosse Bouma and Gertjan van Noord
Recent work has shown that a single multilingual model with typologically informed parameter sharing can improve the performance in dependency parsing on both high-resource and zero-shot conditions. In this work, we investigate whether such improvements are also observed in the POS, NER and morphological tagging tasks.

SNACS Annotation of Case Markers and Adpositions in Hindi
Aryaman Arora and Nathan Schneider
The use of specific case markers and adpositions for particular semantic roles is idiosyncratic to every language. This poses problems in many natural language processing tasks such as machine translation and semantic role labelling. Models for these tasks rely on human-annotated corpora as training data.

There is a lack of corpora in South Asian languages for such tasks. Even Hindi, despite being a resource-rich language, is limited in available labelled data. This extended abstract presents the in-progress annotation of case markers and adpositions in a Hindi corpus, employing the cross-lingual scheme proposed by Schneider et al. (2017), Semantic Network of Adposition and Case Supersenses (SNACS). The SNACS guidelines we developed also apply to Urdu. We hope to finalize this corpus and develop NLP tools making use of the dataset, as well as promote NLP for typologically similar South Asian languages.

Multilingual Jointly Trained Acoustic and Written Word Embeddings
Yushi Hu, Shane Settle and Karen Livescu
Acoustic word embeddings (AWEs) are vector representations of spoken word segments. AWEs can be learned jointly with embeddings of character sequences, to generate phonetically meaningful embeddings of written words, or acoustically grounded word embeddings (AGWEs). Such embeddings have been used to improve speech retrieval, recognition, and spoken term discovery. In this work, we extend this idea to multiple low-resource languages. We jointly train an AWE model and an AGWE model, using phonetically transcribed data from multiple languages. The pre-trained models can then be used for unseen zero-resource languages, or fine-tuned on data from low-resource languages. We also investigate distinctive features, as an alternative to phone labels, to better share cross-lingual information. We test our models on word discrimination tasks for twelve languages while varying the amount of target language training data, and find significant benefits to the proposed multilingual approach.

Multilingual BERT learns abstract case representations
Isabel Papadimitriou, Ethan A. Chi, Richard Futrell and Kyle Mahowald
We investigate how Multilingual BERT (mBERT) encodes grammar by examining how the high-order grammatical feature of morphosyntactic alignment (how different languages define what counts as a "subject") is manifested across the embedding spaces of different languages. To understand if and how morphosyntactic alignment affects contextual embedding spaces, we train classifiers to recover the subjecthood of mBERT embeddings in transitive sentences (which do not contain overt information about morphosyntactic alignment) and then evaluate them zero-shot on intransitive sentences (where subjecthood classification depends on alignment), within and across languages. We find that the resulting classifier distributions reflect the morphosyntactic alignment of their training languages. Our results demonstrate that mBERT representations are influenced by high-level grammatical features that are not manifested in any one input sentence, and that this is robust across languages. Further examining the characteristics that our classifiers rely on, we find that features such as passive voice, animacy and case strongly correlate with classification decisions, suggesting that mBERT does not encode a purely syntactic subjecthood, but a continuous subjecthood as is proposed in much of the functional linguistics literature. Together, these results provide insight into how grammatical features manifest in contextual embedding spaces, at a level of abstraction not covered by previous work.

Keyword Spotting: A quick-and-dirty method for extracting typological features of language from grammatical descriptions
Harald Hammarström

DEmA: the Pavia Diachronic Emergence of Alignment database
Sonia Cristofaro and Guglielmo Inglese
This paper describes a workflow to impute missing values in a typological database, a subset of the World Atlas of Language Structures (WALS). Using a world-wide phylogeny derived from lexical data, the model assumes a phylogenetic continuous time Markov chain governing the evolution of typological values. Data imputation is performed via a Maximum Likelihood estimation on the basis of this model. As back-off model for languages whose phylogenetic position is unknown, a k-nearest neighbor classification based on geographic distance is performed.

A dataset and metric to evaluate lexical extraction from parallel corpora
Barend Beekhuizen
This work presents a novel dataset and a metric for evaluating methods for automated extraction of translation equivalent expressions in massively parallel corpora for the purposes of lexical semantic typology. Patterns in the annotation and the evaluation of the extraction methods were discussed, and directions for future research were indicated.

The role of community size and network structure in shaping linguistic diversity: experimental evidence
Limor Raviv, Antje Mayer and Shiri Lev-Ari
Why are there so many different languages in the world? How much do languages differ from each other in terms of their linguistic structure? And how do such differences come about?

One possibility is that linguistic diversity stems from differences in the social environments in which languages evolve. Specifically, it has been suggested that small, tightly knit communities can maintain high levels of linguistic complexity, while bigger and sparser communities tend to have languages that are structurally simpler, i.e., languages with more regular and more systematic grammars. However, to date this hypothesis has not been tested experimentally. Moreover, community size and network structure are typically confounded in the real-world, making it hard to evaluate the unique contribution of each social factor to this pattern of variation.

To address this issue, we used a novel group communication paradigm. This experimental paradigm allowed us to look at the live formation of new languages that were created in the lab by different micro-societies under different social conditions. By analyzing the emerging languages, we could tease apart the causal role of community size and network structure, and see how the process of language evolution and change is shaped by the fact that languages develop in communities of different sizes and different social structures.

During the group communication game, participants' goal was to communicate successfully about different novel scenes, using only invented nonsense words. A 'speaker' would see one of four shapes moving in some direction on a screen, and would type in a nonsense word to describe the scene (its shape and direction). The 'listener' would then guess which scene their partner was referring to by selecting one of eight scenes on their own screen. Participants received points for every successful interaction (correct guesses), and also feedback to allow them to learn for future interactions. Participants paired up with a different person from their group at every new round, taking turns producing and guessing words.

At the start of the game, people would randomly guessed meanings and make up new names. Over the course of several hours, participants started to combine words or part-words systematically, creating an actual mini-language. For instance, in one group, 'wowo-ik' meant that a specific shape was going up and right, whereas 'wowo-ii' meant that the same shape was going straight up. With such a 'regular' system, it becomes easier to predict the meaning of new labels ('mop-ik' meant a different shape going up and right).

In the first experiment, we examined the role of community size by having participants play in either 'small' groups of four participants or 'large' groups of eight participants. Would the large groups invent more structured languages than the small groups? Results showed that larger groups created languages with more systematic grammars, and did so faster and more consistently than small groups. This finding suggested that the number of people in the community can affect the grammar of languages. We suggest that larger groups are under a stronger pressure to create systematic languages because members of larger groups are typically faced with more input variability, and have less shared history with each member of their group.

In contrast, in the second experiment we found no evidence for a similar role of network structure. When we compared the performance of three network conditions (i.e., fully connected networks, small-world networks, scale-free networks) that varied in their degree of connectivity while group size constant was kept constant, we found that all groups developed languages that were highly systematic, communicatively efficient, stable, and shared across members, with dense and sparse groups reaching similar levels of linguistic structure over time. Although there were no significant differences between networks with respect to their degree of systematic grammar, we found that small-world networks showed the most variance in their behaviors. This result suggests that small-world networks may be more sensitive to random events (i.e., drift).

Together, the findings from the two experiments reported above show that factors in the social environment, and specifically community size, can affect patterns of language diversity and shape the nature and structure of languages.