SIGTYP 2021

The 3rd Workshop on Research in Computational Typology and Multilingual NLP

Proceedings of the Workshop

June 10, 2021

©2021 The Association for Computational Linguistics

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 978-1-954085-34-3

SIGTYP 2021 is the third edition of the workshop for typology-related research and its integration into multilingual Natural Language Processing (NLP). The workshop is co-located with the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2021), which takes place virtually this year. Our workshop includes a shared task on robust language identification from speech.

The final program of SIGTYP contains 4 keynote talks, 3 shared task papers, 10 archival papers, and 14 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 Claire Bowern, Miryam de Lhoneux, Johannes Bjerva, and David Yarowsky for kindly accepting our invitation as invited speakers. The workshop is generously sponsored by Google.

Please find more details on the SIGTYP 2021 website: `https://sigtyp.github.io/ws2021.html`

Organizing Committee:

Ekaterina Vylomova, University of Melbourne
Elizabeth Salesky, Johns Hopkins University
Sabrina Mielke, Johns Hopkins University
Gabriella Lapesa, University of Stuttgart
Ritesh Kumar, Bhim Rao Ambedkar University
Harald Hammarström, Uppsala University
Ivan Vulić, University of Cambridge
Anna Korhonen, University of Cambridge
Roi Reichart, Technion – Israel Institute of Technology
Edoardo M. Ponti, Mila Montreal and University of Cambridge
Ryan Cotterell, ETH Zurich

Program Committee:

Željko Agić, Corti
Emily Ahn, University of Washington
Isabelle Augenstein, University of Copenhagen
Emily Bender, University of Washington
Johannes Bjerva, University of Copenhagen
Claire Bowern, Yale University
Miriam Butt, University of Konstanz
Giuseppe Celano, Leipzig University
Agnieszka Falenska, University of Stuttgart
Richard Futrell, University of California, Irvine
Elisabetta Ježek, University of Pavia
Gerhard Jäger, University of Tubingen
John Mansfield, University of Melbourne
Paola Merlo, University of Geneva
Joakim Nivre, Uppsala University
Robert Östling, Stockholm University
Thomas Proisl, FAU Erlangen-Nurnberg
Michael Regan, University of New Mexico
Ella Rabinovich, University of Toronto
Tanja Samardžić, University of Zurich
Richard Sproat, Google Japan
Sabine Stoll, University of Zurich
Daan van Esch, Google AI
Giulia Venturi, ILC "Antonio Zampolli"
Nidhi Vyas, Apple
Ada Wan, University of Zurich
Eleanor Chodroff, University of York
Elizabeth Salesky, Johns Hopkins University
Sabrina Mielke, Johns Hopkins University
Edoardo M. Ponti, University of Cambridge
Damián Blasi, Harvard University
Adina Williams, Facebook
Ivan Vulić, University of Cambridge
Arturo Oncevay, University of Edinburgh
Koel Dutta Chowdhury, Saarland University

Elena Klyachko, National Research University Higher School of Economics
Alexey Sorokin, Moscow State University
Sylvain Kahane, Université Paris Nanterre
Taraka Rama, University of North Texas
Harald Hammarström, Max Planck Institute for the Science of Human History
Olga Lyashevskaya, National Research University Higher School of Economics
Kaushal Kumar Maurya, IIT Hyderabad
Johann-Mattis List, Max Planck Institute for the Science of Human History
Garrett Nicolai, University of British Columbia
Yevgeni Berzak, Technion – Israel Institute of Technology
Olga Zamaraeva, University of Washington
Zoey Liu, Boston College
Jeff Good, University at Buffalo
Priya Rani, National University of Ireland
Silvia Luraghi, University of Pavia
Beata Trawinski, University of Vienna
Miryam de Lhoneux, University of Copenhagen
Kemal Kurniawan, University of Melbourne
Andreas Shcerbakov, University of Melbourne
Ritesh Kumar, Bhim Rao Ambedkar University

Invited Speakers:

Claire Bowern, Yale University
Miryam de Lhoneux, Uppsala University / KU Leuven / University of Copenhagen
Johannes Bjerva, Aalborg University
David Yarowsky, Johns Hopkins University

Table of Contents

Non-archival Abstracts

Graph Convolutional Network for Swahili News Classification

Alexandros Kastanos and Tyler Martin

In this work, we demonstrate the ability of Text Graph Convolutional Network (Text GCN) to surpass the performance of traditional natural language processing benchmarks on the task of semi-supervised Swahili news categorisation. Our experiments highlight the more severely label-restricted context often facing low-resourced African languages. We build on this finding by presenting a memory-efficient variant of Text GCN which replaces the naive one-hot node representation with a bag of words representation.

Exploring Linguistic Typology Features in Multilingual Machine Translation

Oscar Moreno and Arturo Oncevay

We explore whether linguistic typology features can impact multilingual machine translation performance (many-to-English) by using initial pseudo-tokens and factored language-level embeddings. With 20 languages from different families or groups, we observed that the features of “Order of Subject (S), Object (O) and Verb (V)”, “Position of Negative Word with respect to S-O-V” and “Prefixing vs. Suffixing in Inflectional Morphology” provided slight improvements in low-resource language-pairs despite not overcoming the average performance for all languages.

Multilingual Slot and Intent Detection (xSID) with Cross-lingual Auxiliary Tasks

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi and Barbara Plank

Digital assistants are becoming an integral part of everyday life. However, commercial digital assistants are only available for a limited set of languages (as of March 2020, between 8 to around 20 languages). Because of this, a vast amount of people can not use these devices in their native tongue. In this work, we focus on two core tasks within the digital assistant pipeline: intent classification and slot detection. Intent classification recovers the goal of the utterance, whereas slot detection identifies important properties regarding this goal. Besides introducing a novel cross-lingual dataset for these tasks, consisting of 13 languages, we evaluate a variety of models: 1) multilingually pretrained transformer-based models, 2) we supplement these models with auxiliary tasks to evaluate whether multi-task learning can be beneficial, and 3) annotation transfer with neural machine translation.

Plugins for Structurally Varied Languages in XMG Framework

Valeria Generalova

This paper aims to suggest an XMG-based design of metagrammatical classes storing language-specific information on a multilingual grammar engineering project. It also presents a method of reusing the information from WALS. The principal claim is the hierarchy of features and the modular architecture of feature structures.

Modeling Linguistic Typology - A Probabilistic Graphical Models Approach

Xia Lu

In this paper, we propose to use probabilistic graphical models as a new theoretical and computational framework to study linguistic typology. The graphical structure of such a model represents a meta-language that consists of linguistic variables and the relationships between them while the parameters associated with each variable can be used to infer the strength of the relationships between the variables. Such models can also be used to predict feature values of new languages. Besides providing better solutions to existing problems in linguistic typology such a framework opens up to many new research topics that can help us to gain further insights into linguistic typology.

Unsupervised Self-Training for Unsupervised Cross-Lingual Transfer

Akshat Gupta, Sai Krishna Rallabandi and Alan W Black

Labelled data is scarce, especially for low-resource languages. This beckons the need to come up with unsupervised methods for natural language processing tasks. In this paper, we introduce a general framework called Unsupervised Self-Training, capable of unsupervised cross-lingual transfer. We apply our proposed framework to a two-class sentiment analysis problem of code-switched data. We use the power of pre-trained BERT models for initialization and fine-tune them in an unsupervised manner, only using pseudo labels produced by zero-shot predictions. We test our algorithm on multiple code-switched languages. Our unsupervised models compete well with their supervised counterparts, with their performance reaching within 1-7% (weighted F1 scores) when compared to supervised models trained for a two-class problem.

Let-Mi: An Arabic Levantine Twitter Dataset for Misogynistic Language

Hala Mulki and Bilal Ghanem

Misogyny is one type of hate speech that disparages a person or a group having the female gender identity; it is typically defined as hatred of or contempt for women. Online misogyny has become an increasing worry for Arab women who experience gender-based online abuse on a daily basis. Such online abuse can be expressed through several misogynistic behaviors which reinforce and justify underestimation of women, male superiority, sexual abuse, mistreatment, and violence against women. Misogyny automatic detection systems can assist in the prohibition of anti-women Arabic toxic content. Developing these systems is hindered by the lack of the Arabic misogyny benchmark datasets. In this work, we introduce an Arabic Levantine Twitter dataset for Misogynistic language (LeT-Mi) to be the first benchmark dataset for Arabic misogyny. The proposed dataset consists of 6,550 tweets annotated either as neutral (misogynistic-free) or as one of seven misogyny categories: discredit, dominance, cursing/damning, sexual harassment, stereotyping and objectification, derailing, and the threat of violence. We further provide a detailed review of the dataset creation and annotation phases. The consistency of the annotations for the proposed dataset was emphasized through inter-rater agreement evaluation measures. Moreover, Let-Mi was used as an evaluation dataset through binary, multi-class, and target classification tasks which were conducted by several state-of-the-art machine learning systems along with Multi-Task Learning (MTL) configuration. The obtained results indicated that the performances achieved by the used systems are consistent with state-of-the-art results for languages other than Arabic, while employing MTL improved the performance of the misogyny/target classification tasks.

Our dataset is available at `https://github.com/bilalghanem/let-mi`

Towards Figurative Language Generation in Afrikaans
Imke van Heerden and Anil Bas
This paper presents an LSTM-based approach to figurative language generation, which is an important step towards creative text generation in Afrikaans. Due to the scarcity of resources (in comparison to resource-rich languages), we train the proposed network on a single literary novel. This follows the same approach as Van Heerden and Bas (2021), however, we explicitly focus and expand on fully automatic text generation, centring on figurative language in particular. The proposed model generates phrases that contain compellingly novel figures of speech such as metaphor, simile and personification.

Improving Access to Untranscribed Speech by Leveraging Spoken Term Detection and Self-supervised Learning of Speech Representations
Nay San, Martijn Bartelds and Dan Jurafsky
We summarise findings from our recent work showing that a large self-supervised model trained only on English speech provides a noise-robust and speaker-invariant feature extraction method that can be used for a speech information retrieval task with unrelated low resource target languages. A qualitative error analysis also revealed that the majority of the retrieval errors could be attributed to the differences in phonological inventories between English and the evaluation languages. With a longer-term aim of leveraging typological information to better adapt such models for the target languages, we also report on work in progress which examines the phonetic information encoded in these representations.

On the Universality of Lexical Concepts
Bradley Hauer and Grzegorz Kondrak
We posit that lexicalized concepts are universal, and thus can be annotated cross-linguistically in parallel corpora. This is one of the implications of a novel theory that formalizes the relationship between words and senses in both monolingual and multilingual settings. The theory is based on a unifying treatment of the notions of synonymy and translational equivalence as different aspects of the relation of sameness of meaning within and across languages.

Quantitative Detection of Cognacy in the Predictive Structure of Inflection Classes: Romance Verbal Conjugations against the Broader Typological Variation
Borja Herce and Balthasar Bickel
In recent years, Information Theory (with its core notion of entropy) has provided the theoretical background for a lot of empirical research on inflectional systems, and has inspired various metrics to capture (different aspects of) their complexity. So far, however, entropy-based metrics have chiefly been used to assess synchronic states. Here we explore their potential for capturing patterns in language change and phylogenetic relatedness. Specifically, we probe different aspects of an inflectional system for their stability within one language family, Romance, and for the degree to which they distinguish this family from unrelated and less closely related languages. Based on most metrics, Romance appears to be different from the control sample in the mean, variance, or both. The difference in variance is particularly interesting because it might suggest differences in relative diachronic stability and as phylogenetic signals of relatedness.

Subword Geometry: Picturing Word Shapes
Olga Sozinova and Tanja Samardzic
In this work in progress, we are investigating the structural properties of subwords in 20 languages by extracting word shapes, i.e. sequences of subword lengths.

A Look to Languages through the Glass of BPE Compression

Ximena Gutierrez-Vasques, Tanja Samardzic and Christian Bentz

One of the predominant methods for subword tokenization is Byte-pair encoding (BPE). Originally, this is a data compression technique based on replacing the most common pair of consecutive bytes with a new symbol When applied to text, each iteration merges two adjacent symbols; this can be seen as a process of going from characters to subwords through iterations.

Regardless of the language, the first merge operations tend to have a stronger impact on the compression of texts, i.e., they capture very frequent patterns that lead to a reduction of redundancy and to an increment of the text entropy. However, the natural language properties that allow this compression are rarely analyzed, i.e., do all languages get compressed in the same way through BPE merge operations? We hypothesize that the type of recurrent patterns captured in each merge depends on the typology and even orthography and other corpus-related phenomena. For instance, for some languages, this compression might be related to frequent affixes or regular inflectional morphs, while for some others, it might be related to more idiosyncratic, irregular patterns or even related to orthographic redundancies.

We propose a novel way to quantify this, inspired by the notion of morphological productivity.

Information-Theoretic Characterization of Morphological Fusion

Neil Rathi, Michael Hahn and Richard Futrell

Traditionally, morphological typology divides synthetic languages into two broad groups (e.g. von Schlegel, 1808; von Humboldt, 1843). Agglutinative languages, such as Turkish, segment morphemes into independent features which can be easily split. On the other hand, fusional languages, such as Latin, “fuse” morphemes together phonologically (Bickel and Nichols, 2013). At the same time, there has long been recognition that the categories “agglutinative” and “fusional” are best thought of as a matter of degree, with Greenberg (1954) developing an “index of agglutination” metric for languages. Here, we propose an information-theoretic definition of the fusion of any given form in a language, which naturally delivers a graded measure of the degree of fusion. We use a sequence-to-sequence model to empirically verify that our measure captures typical linguistic classifications.