Recent Developments in Computational Typology and Multilingual Natural Language Processing

April 30 2020 · Issue #3

Editors: Ekaterina Vylomova and Ryan Cotterell

This is SIGTYP's third newsletter on recent developments in computational typology and multilingual natural language processing. Each month, various members of SIGTYP will endeavour to summarize recent papers that focus on these topics. The papers or scholarly works that we review are selected to reflect a diverse set of research directions. They represent works that the editors found to be interesting and wanted to share. Given the fast-paced nature of research in our field, we find that brief summaries of interesting papers are a useful way to cut the wheat from the chaff.

We expressly encourage people working in computational typology and multilingual NLP to submit summaries of their own research, which we will collate, edit and announce on SIGTYP's website. In this issue, for example, we had Edoardo Ponti, Himanshu Yadav, Kazuya Kawakami, Elizabeth Salesky, Eleanor Chodroff, Pratik Joshi and Sebastin Santy, Pranav A, Tiago Pimentel, and Kartikay Khandelwal describe their recent publications on linguistic typology and multilingual NLP.

Research Papers

Multi-SimLex: A Large-Scale Evaluation of Multilingual and Cross-Lingual Lexical Semantic Similarity

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart and Anna Korhonen

Summary by Edoardo M. Ponti, University of Cambridge

Multi-SimLex is a large-scale multilingual resource for lexical semantics. The current version of Multi-SimLex provides human judgments on the semantic similarity of word pairs for as many as 12 monolingual and 66 cross-lingual datasets. The languages covered are typologically diverse and represent both major languages (e.g., Mandarin Chinese, Spanish, Russian) and less-resourced ones (e.g., Welsh, Kiswahili). Each language dataset is annotated for the lexical relation of semantic similarity and contains 1,888 semantically aligned concept pairs, providing a representative coverage of word classes (nouns, verbs, adjectives, adverbs), frequency ranks, similarity intervals, lexical fields, and concreteness levels. Moreover, we evaluate a wide array of recent state-of-the-art representation models as baselines for both the monolingual and cross-lingual benchmarks, and present a step-by-step annotation protocol for creating consistent datasets for additional languages. The dataset, baseline scores, and guidelines can be found at multisimlex.com.

Learning Robust and Multilingual Speech Representations

Kazuya Kawakami, Luyu Wang, Chris Dyer, Phil Blunsom, Aaron van den Oord

Summary by Kazuya Kawakami, DeepMind

Recently, unsupervised speech representation learning has shown remarkable success at finding representations that correlate with phonetic structures and improve downstream speech recognition performance. However, most research has been focused on evaluating the representations in terms of their ability to improve the performance of speech recognition systems on read English (e.g. Wall Street Journal and LibriSpeech). This evaluation methodology overlooks two important properties that speech representations should have: robustness to domain shifts and

transferability to other languages. Traditionally such invariances were hard-coded in feature extraction methods. For example, standard MFCC features are known to be sensitive to additive noise and many modifications have been proposed to overcome those limitations. In this paper we learn representations from up to 8000 hours of diverse and noisy speech data and evaluate the representations by looking at their robustness to domain shifts and their ability to improve recognition performance in many languages. We find that our representations confer significant robustness advantages to the resulting recognition systems: we see significant improvements in out-of-domain transfer relative to baseline feature sets and the features likewise provide improvements in 25 phonetically diverse languages including tonal languages and low-resource languages. Our results suggest we are making progress toward models that implicitly discover phonetic structure from large-scale unlabelled audio signals.

Word Order Typology Interacts With Linguistic Complexity: A Cross-Linguistic Corpus Study

Himanshu Yadav, Ashwini Vaidya, Vishakha Shukla, Samar Husain

Summary by Himanshu Yadav and Samar Husain

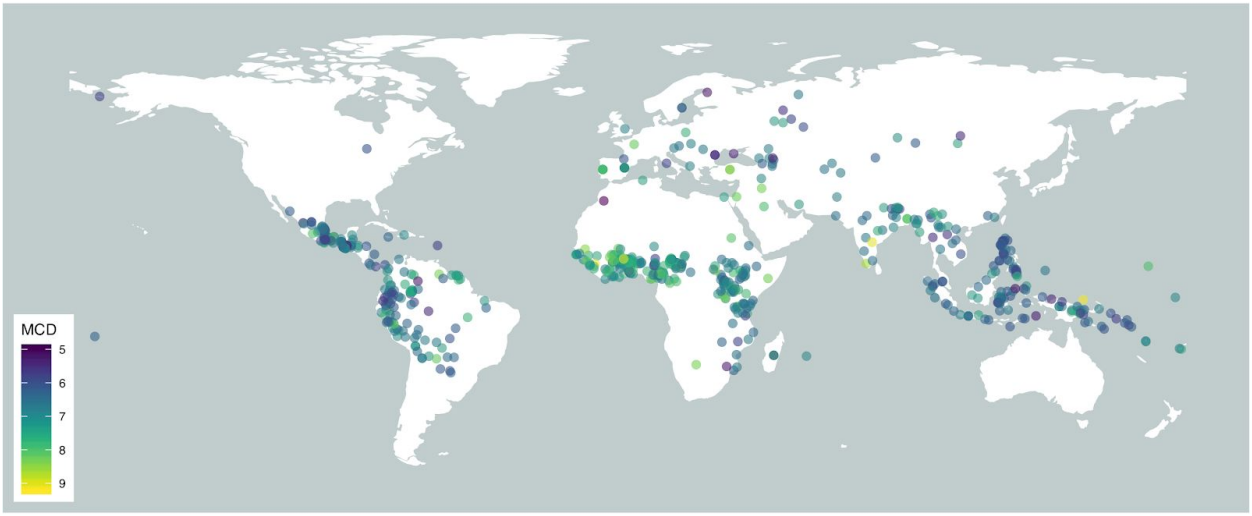
It has been argued that natural languages minimize the linear head-dependent distance (measured as the number of words that intervene a head and its dependent). In a cross-linguistic study, we found that languages allow for differences in head-dependent distance based on the directionality of a dependency, i.e., whether a head follows a dependent vs. whether it precedes a dependent. Critically, such an asymmetry in linear distance is driven by the typological word order of the language — SOV languages allow for longer dependencies when heads follow dependents, while SVO languages allow for longer dependencies when heads precede dependents. Interestingly, we find that compared to linear dependency distance, hierarchical distance (measured as the number of syntactic heads that intervene a head and its dependent) is less across languages with differing word orders. This suggests that cross-linguistically, constraints on hierarchical distance are stronger than that on linear distance. The pattern across various languages points to 'limited' adaptability with regard to the word order of a language and highlights the close interaction of linguistic exposure and working memory constraints in determining sentence complexity. We argue that processing adaptability has limits and working memory constraints cannot be overridden beyond a certain threshold.

A Corpus for Large-Scale Phonetic Typology (to be presented at ACL)

Elizabeth Salesky, Eleanor Chodroff, Tiago Pimentel, Matthew Wiesner, Ryan Cotterell, Alan W. Black, Jason Eisner

Summary by Elizabeth Salesky and Eleanor Chodroff

A major obstacle in data-driven research on typology is having sufficient data in a large number of languages to draw meaningful conclusions. We are excited to present the first large-scale corpus for phonetic typology, with aligned phonological segments and phonetic measures for 699 languages. At present, we provide durational and spectral measures of vowels and sibilants for over 150 languages, with several hundred more in the works. This corpus will allow investigation of phonetic typology across many languages, for which several had no pre-existing speech resources, as well as research into phonetic and phonological universals at a much larger scale than before. Previous research on phonetic and phonological typology has largely relied on type-level descriptions (e.g., phoneme inventories) or been limited in the number of languages for which phenomena can be investigated. The token-level measurements in our corpus enable further research into distributional trends within and across phonetic segments. Extending extraction procedures to new and especially low-resource languages is difficult and computationally-intensive, particularly without high-quality resources like pronunciation lexicons and transcribed speech: to create our corpus, we leverage the CMU Wilderness corpus (Black et al. 2019) to create and release phone alignments, vowel formants, and sibilant measurements, enabling the community to skip our 6-CPU year walk in the wilderness. We present our measurements with a phone-level confidence (cepstral distortion (CD): a measure of the distance between speech examples and synthesized speech using the generated phone alignments), so that researchers may choose their own trade-off between number of examples and possible corpus noise. We additionally present a series of case studies illustrating possible types of research enabled by this corpus: studies of phonetic dispersion, uniformity, and frequency effects. For example, across more than 150 languages, we find a correlation of 0.75 between the F1 of /e/ and /o/. This may reflect a universal 'uniformity' constraint on the phonetic realization of phonological segments. The corpus will be released by the time of ACL2020. Keep an eye to https://voxclamantisproject.github.io/ for the official data release; this page will be a stable platform for updates and announcements!



The State and Fate of Linguistic Diversity and Inclusion in the NLP World

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury

Summary by Pratik Joshi and Sebastin Santy, Microsoft Research Labs, India

Language technologies are becoming increasingly important in boosting multilingualism and diversity around the world. However, only a small fraction of over 7000 world-wide languages are supported by these rapidly transforming applications and technologies. In this work, we quantitatively investigate the relation between the types of languages, resources, and their representation in NLP conferences to understand the trajectory that different languages have followed over time. We start by proposing a taxonomy of languages based on their availability in terms of labeled and unlabeled data, and subsequently conduct analyses through the lens of these taxonomy classes.

We perform the following quantitative analyses: 1: Assessing the resource disparities over individual repositories with respect to languages from each taxonomy class, 2: Measuring typological diversity and representation in various standard NLP resources, 3: Calculating statistical metrics (entropy and MRR) for language occurrence in iterations of NLP conferences, and 4: Using entity (author, conference, language) embeddings to capture subtle trends, then visualizing and deriving insights of conference trajectories with respect to the taxonomy classes.

The findings show that the taxonomy is evident throughout the various analyses, highlighting the disparity between support for different languages. Observations are made on how some typological features are underrepresented in standard NLP resources, possibly making their use in transfer learning less effective. Further, we note that some venues, such as LREC and Workshops, are more inclusive than others, observed through the entropy plots and embeddings visualization. The embeddings analysis indicate a chronological and technological shift in NLP. Finally, through the MRR calculations, it is observed that there are focused communities working on low-resource languages, but many yet are still in need of support.

2kenize: Tying Subword Sequences for Chinese Script Conversion (to be presented at ACL)

Pranav A, Isabelle Augenstein

Summary by Pranav A

Simplified Chinese to Traditional Chinese script conversion is a common preprocessing step in Chinese NLP. A significant issue in script conversion is that a simplified Chinese character can correspond to multiple traditional characters (Halpern et al.). Due to this, we find that current off-the-shelf script converters give 55-85% sentence accuracy. Our further investigations show that advanced neural models like neural language model character disambiguation and neural sequence models result in sentence accuracy of about 84-85%, mainly due to the false positives. We speculate that this is because these models are not able to determine subword boundaries correctly, leading to an incorrect conversion.

Hence, we propose 2kenize, a subword segmentation model which jointly takes Simplified Chinese and 'lookahead'-ing Traditional Chinese constructions, into consideration. We achieve this by constructing a joint Simplified Chinese and Traditional Chinese language model based Viterbi tokenizer. Mapping disambiguation based on this tokenization gives a result of 91-95% sentence accuracy on a challenging dataset. Our qualitative error analysis reveals that our method's particular strengths are in dealing with code-mixing and named entities.

We conduct an extrinsic evaluation on topic classification tasks and find that the dataset converted using our model outperforms other converters. Additionally, our results on topic classification show that subword tokenizers outperform character and word-based models; and subword regularization methods like BPE-Drop and Unigram outperform BPE. We then tweaked 2kenize by tokenizing it only on Traditional Chinese sentences, calling it as 1kenize. Our experiments show that 1kenize performs at par with other subword tokenizers in formal-style datasets and outperforms in informal-style datasets. From this result, we deduce that the performance of these tokenizers is highly correlated with the skewness of token distribution.

Our contributions in terms of resources are:

1. 2kenize: Simplified Chinese to Traditional Chinese script converter
2. Character conversion evaluation datasets: Spanning Hong Kong and Taiwanese literature and news genres.
3. Topic Classification datasets: Formal style (*zh-hant*) and informal style (*zh-hant* and *zh-yue*) traditional Chinese spanning genres like news, social media discussions, and memes.

Information-Theoretic Probing for Linguistic Structure

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, Ryan Cotterell

Summary by Tiago Pimentel, University of Cambridge

This paper looks at what information theory has to say about probing. We analysed the question "how much information about linguistic structure is encoded in some specific contextual embeddings" and found that, under a weak assumption, the embeddings contain as much information as the original sentence. As such, the quest to answer this question only tells us something about linguistics and not about the embeddings themselves. We then propose the use of control functions to compare contextual and uncontextual (type-based) embeddings. Our experiments on Basque, Czech, English, Finnish, Tamil, and Turkish suggest that BERT embeddings encode only at most 5% more information about POS tagging than fastText ones.

Unsupervised Cross-lingual Representation Learning at Scale

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek et al.

Summary by Kartikay Khandelwal, Facebook AI

In this work, we introduce XLM-R, a state-of-the-art multilingual model pre-trained on 100 languages, that significantly outperforms previous work across a variety of benchmarks. We evaluate our model on cross-lingual natural language inference (XNLI), multilingual question answering (MLQA) and multilingual NER. Specifically,

- On XNLI, XLM-R obtains an average accuracy of 80.9%, outperforming the XLM-100 and multilingualBERT open-source models by 10.2% and 14.6%. XLM-R performs particularly well on low-resource languages, improving 15.7% in XNLI accuracy for Swahili and 11.4% for Urdu over previous XLM models.
- On MLQA, XLM-R outperforms the previous SOTA by 9.1% F1-score and multilingualBERT by 13.0%.

- Apart from these impressive gains on cross-lingual benchmarks, XLM-R shows strong monolingual performance on GLUE, where it is competitive with SOTA English only models.

The goal for this work is to not only provide strong models with high performance on a range of benchmarks, but also to provide analysis and intuitions for what makes these models work. Through carefully designed ablation studies, we highlight the limitations of previous multilingual models (multilingualBERT and XLM), especially in modeling low resource languages. We also present a comprehensive study of different factors that are important to pre-training large scale multilingual models and show for the first time the possibility of multilingual modeling without sacrificing per-language performance. Specifically, we investigate:

- The trade-offs between the positive transfer from high resource to low resource languages and the dilution in per-language capacity (interference), as we scale the number of languages during pre-training.
- The importance of different parameters, including the scale of data, vocabulary construction and language sampling, on the effectiveness of the model to trade-off performance between high and low resource languages.

Our paper is the first to provide comprehensive experiments that help understand this transfer/interference trade-off in the context of unsupervised cross-lingual representation learning. Our results which show multilingual models out-performing monolingual ones is an important result for the XLU/NLU community and has significant consequences on how these models are deployed in industry, namely the ability to deploy a single model for all languages. We open-sourced all of our models and code and hope the research community builds on top of our learnings.

Resources

Multilingual Culture-Independent Word Analogy Datasets

Matej Ulčar, Kristiina Vaik, Jessica Lindström, Milda Dailidėnaitė, Marko Robnik-Šikonja

Summary by Ekaterina Vylomova, University of Melbourne

Word analogy task is a typical benchmark to evaluate and compare quality of word embeddings. Such tasks consist of two word pairs such as (“king”, ”man”) and (“queen”, “woman”). The models are then provided with three words and asked to predict the missing part, e.g, (“king”, ”man”), (“?”, “woman”). Majority of models are only evaluated on English, partially due to lack of corresponding datasets in English. Here the authors introduce new (culturally independent) word analogy datasets in Croatian, English, Estonian, Finnish, Latvian, Lithuanian, Russian, Slovenian, and Swedish. Similar to its English counterpart, the dataset contains encyclopedic (e.g., “capital -- country”), morphosyntactic (e.g., “present tense -- past tense” or superlative/comparative

adjectives), and morphosemantic (e.g., adjective--adverb) relations but is also augmented with some extra relations such as genitive--dative in order to address more complex morphologies. Evaluation on FastTest embeddings indicate significant differences across languages and types of relations and suggest that there is a substantial room for further improvement.

Shared Task

SIGTYP 2020 - Prediction of Typological Features

We are pleased to announce its first shared task on Typological Feature Prediction this year! The shared task covers around 2,000 languages, with typological features taken from the World Atlas of Language Structures. Submitted systems will be tested on held-out languages balanced for both genetic relationships and geographic proximity. Two sub-tasks will be present: 1) Constrained: only provided training data can be employed. 2) Unconstrained: training data can be extended with any external source of information (e.g. pre-trained embeddings, texts, etc.)
For more information on data format and important dates, please visit our website https://sigtyp.github.io/st2020.html

SIGTYP 2020 Workshop - Online!

We would like to announce that SIGTYP 2020 will be held ***online***. The workshop is preliminary scheduled for **Nov, 19th.**