Recent Developments in Computational Typology and Multilingual Natural Language Processing

March 19 2020 · Issue #2

Editors: Ekaterina Vylomova and Ryan Cotterell

This is SIGTYP's second 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 Johann-Mattis List, Takashi Wada, Johannes Bjerva, Shijie Wu, Antonis Anastasopoulos, Kyle Gorman describe their recent publications on linguistic typology and multilingual NLP.

Research Papers

Evolutionary Dynamics in the Dispersal of Sign Languages

Justin M. Power, Guido W. Grimm and Johann-Mattis List

Summary by Johann-Mattis List, Max Planck Institute for the Science of Human History

The evolution of spoken language has been studied since the mid-nineteenth century using the traditional comparative method from historical linguistics and, more recently, computational phylogenetic methods. By contrast, evolutionary processes resulting in the diversity of contemporary sign languages (SLs) have received much less attention, and scholars have been largely unsuccessful in grouping SLs into monophyletic language families using traditional methods. To date, no published study has attempted to use language data to infer relationships among SLs on a large scale. Here, we report the results of a phylogenetic analysis of 40 contemporary and 36 historical SL manual alphabets coded for morphological similarity. Our results support grouping SLs in the sample into six main European lineages, with three larger groups of Austrian, British and French origin, as well as three smaller groups centring around Russian, Spanish and Swedish. The results of our analysis for British and Swedish lineages support current knowledge of relationships among SLs based on extra-linguistic historical sources. With respect to other lineages, our results diverge from current hypotheses by indicating (i) independent evolution of Austrian, French and Spanish from Spanish sources; (ii) an internal Danish subgroup within the Austrian lineage; and (iii) evolution of Russian from Austrian sources.

The Database of Cross-Linguistic Colexifications, Reproducible Analysis of Cross-Linguistic Polysemies

Christoph Rzymski, Tiago Tresoldi, Simon J. Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm et al.

Summary by Johann-Mattis List, Max Planck Institute for the Science of Human History

Advances in computer-assisted linguistic research have been greatly influential in reshaping linguistic research. With the increasing availability of interconnected datasets created and curated by researchers, more and more interwoven questions can now be investigated. Such advances,

however, are bringing high requirements in terms of rigorousness for preparing and curating datasets. Here we present CLICS, a Database of Cross-Linguistic Colexifications (CLICS). CLICS tackles interconnected interdisciplinary research questions about the colexification of words across semantic categories in the world's languages, and show-cases best practices for preparing data for cross-linguistic research. This is done by addressing shortcomings of an earlier version of the database, CLICS2, and by supplying an updated version with CLICS3, which massively increases the size and scope of the project. We provide tools and guidelines for this purpose and discuss insights resulting from organizing student tasks for database updates.

Unsupervised Multilingual Word Embedding with Limited Resources using Neural Language Models

Takashi Wada, Tomoharu Iwata, and Yuji Matsumoto

Summary by Takashi Wada, the University of Melbourne

The paper proposes an unsupervised method for learning multilingual word embeddings with limited resources (e.g., 50k sentences in monolingual corpora). The method obtains multilingual word embeddings by jointly training a language model that shares LSTMs among multiple languages but has different look-up tables for word embeddings. The authors argue that if languages share word order (e.g., SVO, SOV), the shared LSTMs can extract the common structure and learn word embeddings in a shared space. They conduct several experiments simulating low-resource conditions and demonstrate that among syntactically similar languages, their approach outperforms baselines that map pre-trained word embeddings into a common space. They also show that monolingual word embeddings trained on small data sizes are far from isomorphic between different languages, which would explain the poor performance of the mapping methods.

Zero-Shot Cross-Lingual Transfer with Meta Learning

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, Isabelle Augenstein

Summary by Johannes Bjerva, University of Copenhagen

The paper addresses the problem of training on multiple different languages at the same time, with little or no data available for languages other than English. This is especially important when considering the fact that most languages in the world suffer from being under-resourced.
The authors show that this can be approached with meta-learning, where, in addition to training a source language model, another model learns to select which languages are the most beneficial. The

experiments are run using standard supervised, zero-shot cross-lingual, as well as few-shot cross-lingual settings for different natural language understanding tasks (natural language inference, question answering). An extensive experimental setup demonstrates the consistent effectiveness of meta-learning, on a total 16 languages. Furthermore, the results improve upon state-of-the-art on zero-shot and few-shot NLI and QA tasks on the XNLI and X-WikiRe datasets, respectively.
The paper includes a comprehensive analysis which indicates that correlation of typological features between languages can further explain when parameter sharing learned via meta learning is beneficial.

Emerging Cross-lingual Structure in Pretrained Language Models

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, Veselin Stoyanov

Summary by Shijie Wu, Johns Hopkins University

The paper offers an ablation study on bilingual BERT to understand why multilingual BERT-style models learn cross-lingual representation without any explicit cross-lingual signal. The authors consider the effects of domain similarity, single vocabulary with identical subwords shared across languages, single softmax layers and single transformer shared across languages. On English-French, English-Russian and English-Chinese, they show sharing a transformer across languages plays the most important role in learning cross-lingual representation and other factors only affect the representation slightly. Although the bilingual BERTs can learn cross-lingual representation without shared subwords across languages, more shared subwords, made possible by synthetic code-switching using bilingual dictionaries, is still beneficial. The paper then explores why sharing transformers across languages play the biggest role. It turns out that monolingual BERTs of different languages, despite being trained independently, exhibit surprising similarity. As a proof of concept, the authors align the representation of BERT in word-level, contextual word-level and sentence-level using only linear rotation and show surprisingly strong downstream performance. It is similar to how monolingual word embeddings can be aligned to produce cross-lingual word embeddings. Finally, the authors also present a similarity analysis with centered kernel analysis, a similarity measurement connected to CCA, and visualize how similar are monolingual BERTs of two languages compared to bilingual BERT of the same languages.

Universal Phone Recognition with a Multilingual Allophone System

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell et al.

Summary by Graham Neubig and Antonis Anastasopoulos, CMU

The authors create a multi-lingual ASR model that can do zero-shot phone recognition in up to 2,186 languages! Most multilingual ASR models conflate two concepts: phonemes (sounds that can support lexical contrasts in a *particular* language) and their corresponding phones (the sounds that are actually spoken, which are language *independent*). Li et al. create a model that first recognizes language-independent phones, and then converts these phones to language-specific phonemes. This makes the underlying representations of phones more universal and generalizable across languages. The authors create a database of phone-phoneme mappings, AlloVera, that contains such mappings for several languages, which were used for training the model. Special care was taken in order to ensure that the chosen training languages cover the phone inventory as much as possible. They also demonstrate how to customize the model for a specific language (by restricting the output space) for better zero-shot recognition, by using PHOIBLE, a large phonetic database with 2,186 languages. This leads to improvements of more than 17% Phone Error Rate in true zero shot settings for low-resource languages like Inuktitut and Tusom. Results show that this approach also helps both in standard multilingual recognition as well as recognition on entirely new languages. For future work, the authors are interested in incorporating these into real systems for low- or zero-resource ASR.

Massively Multilingual Pronunciation Mining with WikiPron

Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza et al.

Summary by Kyle Gorman, CUNY

WikiPron is a Python-based toolkit for scraping pronunciation data from Wiktionary. Using the command-line interface, users can specify language (using the English name or ISO-639 codes), dialect, whether they want to scrape "phonemic" pronunciations in angled brackets (//) or "phonetic" pronunciations in square brackets ([]), and various normalizations to be applied to the word-form or its transcription. Lee et al. (LREC, to appear) use this tool to produce a 165-language, 1.7 million-word pronunciation database, and validate this data set using models. They also train and evaluate two G2P models, a pair n-gram (Novak et al., 2016) and a neural sequence-to-sequence (Ott et al., 2019), on a set of typologically diverse languages. The results show that 1) the neural model outperforms the pair n-gram model on most (but not all) languages;

and 2) word error rates differ significantly ranging 2% in Hungarian (that has ``shallow'' orthography and one the largest training sets) upto 48% in English (with conservative and highly abstract orthography). Finally, the 2020 SIGMORPHON workshop will also use some of the WikiPron data for a shared task on grapheme-to-phoneme conversion.

Modeling Morphological Learning, Typology, and Change: What can the Neural Sequence-to-Sequence Framework Contribute?

Micha Elsner, Andrea D. Sims, Alexander Erdmann, Antonio Hernandez et al.

Summary by Ekaterina Vylomova, University of Melbourne

Although languages may substantially differ, all of them are learnable and are constrained by the learning mechanism. In the domain of morphology, the notion of learnability is often described in terms of morphological complexity. Are all morphological paradigms learnable? How can morphological complexity be estimated? Elsner and colleagues provide an extensive overview of recent approaches to address this question. They specifically focus on neural sequence-to-sequence models, the type of computational models that recently became popular in NLP.

In the first part, the authors discuss the notion of learnability in terms of two traditional models of morphology, Item-and-Arrangement (IA) and Word-and-Paradigm (WP).
IA models typically approach a learner's task as the acquisition of the morpheme inventory and the rules for composing the parts into words (and are limited to concatenative morphology). This makes the size of morphological systems the main focus in the evaluation of morphological complexity (E(numerative)-complexity in Ackerman and Malouf, 2013). E-complexity, therefore, grows rapidly with the increase of the number of allomorphs and paradigm cells. But, contrary to that, human languages have far fewer inflectional classes than they potentially could, given the number of allomorphs. Can it be explained by the fact that a large number of paradigm cells and inflectional classes make languages more difficult to learn? The authors list a number of principles (such as the Principle of Contrast (Clark, 1987) and its byproduct, the No Blur Principle (Carstairs-McCarthy, 1994)), that attempt to explain this.

The second model, Word-and-Paradigm (WP), takes words as basic units of morphological structure, and the inflectional meaning is instantiated via contrast between the (surface) forms filling paradigmatic cells. Importantly, WP models do not rely on a one-to-one correspondence between morphological form and the meaning. During acquisition, the learner needs to predict unobserved forms from a number of observed ones in the paradigm (the Paradigm Cell Filling Problem, PCFP). Morphological complexity is, therefore, taken as an amount of ambiguity in the prediction of the corresponding form's inflection class. Unlike IA models where complexity is

approached via the number of forms, the WP models focus on their predictability: the more forms are interpredictable, the lower complexity. This corresponds to I(ntegrative)-complexity in Ackerman and Malouf (2013) that is estimated as an average conditional entropy, i.e. average unpredictability of a target form (its affix) given another form of the same lexeme. The conditional entropy is typically low, and this is suggested to be typologically universal. This is then referred to as "the Low Entropy Conjecture" which states that E-complexity is effectively unrestricted, as long as I-complexity is low.

Then the authors proceed to another factor to be considered during language acquisition, the Zipfian distribution: most inflected forms are sparsely attested and speakers encounter the PCFP problem on a regular basis. On the other hand, high-frequency words might be stored in memory and are not subject to PCFP. This predicts cross-linguistic differences in the PCFP: some languages might be more challenging than others. The WP model also suggests language-internal differences: some parts of the paradigm might be easier to learn than others (Stump and Finkel, 2013).

The authors, therefore, conclude that all these factors suggest the need for fine-grained analysis and measure of learnability instead of (traditionally used) averaging. In particular, averaging fails to distinguish systems with a few highly irregular classes from systems where every form is slightly unpredictable. Moreover, the average pairwise entropy does not compute the joint entropy of the entire distribution (Cotterell, 2018).

The second part of the paper discusses computational tasks and methods to approach the PCFP, specifically focusing on the morphological reinflection task which is framed within the WP model. SIGMORPHON has previously organized a series of such tasks (2016, 2017, 2018, 2019) where the systems compete in their ability to generate a target (inflected) form from its lemma and target tags (morphosyntactic features). For instance, the input sequence might be ``understand'' + V;PST and the systems are required to predict ``understood''. The results of the shared tasks showed that neural sequence-to-sequence models adapted from machine translation achieved high prediction accuracies and outperformed most other types of models. Such models often do only rely on word segmentation and operate on a character level instead (fitting the WP model quite well) as well as do not expect one-to-one mapping between the meaning and the form. The results look promising from an engineering perspective, but can researchers use the data and models to approach language acquisition and estimate morphological complexity? The authors outline several drawbacks in the current state of the tasks. First of all, the models are compared in terms of percentage accuracy of the predicted output strings (i.e. a single number per language). Elsner and colleagues argue for a fine-grained analysis involving the notions of ``inflection class'' and ``exponent'' and provide an example for morphological reinflection analysis in Latin.
The authors also provide an insight into how the computational models can be used to estimate not only which system is easier to learn but also which forms, classes contribute to the morphological complexity. Second, sequence-to-sequence models were also used as cognitive models of infant language learners (Cotterell, 2018; Corkery, 2019). But is the data suitable for the task? Both lexical

items and paradigms should follow the Zipfian distribution. The datasets released by the SIGMORPHON do not reflect the Zipfian distribution and are biased towards rare forms and words.

Finally, the authors discuss the processes of language change as a combination of learnability and prestige (and other social factors). They outline why certain ``non-functional'' parts of the morphological system are not eliminated and which structural properties make morphological systems more conservative.

Resources

LIMA 3.0 beta: A Multilingual Analyser

LIMA, the LIbre Multilingual Analyzer, with state of the art performance and now fully multilingual with more than 60 language models available.

MILANLP: 30 BERT-based Models

A resource that (at this moment) collects 30 BERT-based models, 18 Languages and 28 Tasks for a total of 177 entries.

WikiPron: Pronunciation Data from Wiktionary

An open-source command-line tool for extracting pronunciation data from Wiktionary, a collaborative multilingual online dictionary.

TyDi QA: A Multilingual Question Answering Benchmark

“TyDi QA includes over 200,000 question-answer pairs from 11 languages representing a diverse range of linguistic phenomena and data challenges. Many of these languages use non-Latin alphabets, such as Arabic, Bengali, Korean, Russian, Telugu, and Thai. Others form words in complex ways, including Arabic, Finnish, Indonesian, Kiswahili, Russian. The authors collected questions from people who *wanted* an answer, but *did not* know the answer yet. To inspire questions, they showed people an interesting passage from Wikipedia written in their native language. They then had them ask a question, *any question*, as long as it was *not* answered by the passage and they *actually* wanted to know the answer. For each of these questions, the authors then performed a

Google Search for the best-matching Wikipedia article in the appropriate language and asked a person to find and highlight the answer within that article."
Paper: https://storage.cloud.google.com/tydiqa/tydiqa.pdf

New SIGs

SIGEL - ACL Special Interest Group for Endangered Languages

The purpose of SIGEL, as specified in its approved constitution, is "to foster computationally grounded research in all useful aspects in documenting, processing, revitalizing and supporting endangered languages, as well as minority, Indigenous and low-resource languages."

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.)
Stay tuned for the training data release, to happen shortly (expected date: 20 Mar 2020)! For more information on data format and important dates, please visit our website
https://sigtyp.github.io/st2020.html