Recent Developments in Computational Typology and Multilingual Natural Language Processing

February 11 2020 · Issue #1

Editors: Ekaterina Vylomova and Ryan Cotterell

This is SIGTYP's first 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 Miryam de Lhoneux, Haim Dubossarsky, and Edoardo M. Ponti describe their recent publications on multilingual NLP. We also thank Elizabeth Salesky, Jennifer C. White, and Yulia Otmakhova for their summaries and contribution to preparation of the current issue.

Theses

Linguistically Informed Neural Dependency Parsing for Typologically Diverse Languages

Summary by Miryam de Lhoneux, Uppsala University

In her recent thesis, Miryam de Lhoneux studies neural dependency parsing for typologically diverse languages making use of treebanks from Universal Dependencies (UD) project. The high-level, overarching question she asks is *how can linguistics inform neural NLP?* At a time when NLP models are increasingly becoming inscrutable black boxes, it is not always obvious what the role of linguistics in NLP should be. Reviewing the literature, she identifies three approaches by which linguistic knowledge can be useful for neural NLP:

1. Informing neural architecture design,
2. analyzing what a model learns about language (think diagnostic classifiers), and
3. engineering a multi-task learning architecture.

In her thesis, Miryam considers each of these approaches in the case of dependency parsing.

In her thesis, her first step is to extend a neural parser to work well on a typologically diverse set of languages, including morphologically complex languages and languages whose treebanks have a large number of non-projective sentences. She additionally argues that we should be careful when evaluating our models to avoid biasing our models towards languages that are overrepresented in UD, such as Indo-European languages, and, to that end, defines criteria to sample a typologically representative subset of UD treebanks for parser development and evaluation.

Miryam next studies sequential and hierarchical neural models for parsing, with the intuition that hierarchical modelling should be helpful for syntactic parsing given the hierarchical nature of syntax. Neural parsers using BiLSTMs, which are sequential models, perform as well as the best reported scores using recursive neural network parsers, which are hierarchical models. This casts doubt on how crucial it is to directly incorporate a hierarchical inductive bias in a parsing model. She studies the incorporation of a recursive neural network layer in a BiLSTM-based parser; she

confirms that recursive networks are unnecessary, giving further evidence that sequential BiLSTMs are sufficient.

Then, Miryam investigates the transitivity and agreement information learned by a BiLSTM-based parser on auxiliary verb constructions (AVCs). She suggests that a parser should learn similar information about AVCs to the information it learns for finite main verbs. This is motivated by work in theoretical dependency grammar[1]. The parser learns different information about these two if it is not augmented with a recursive layer, but similar information if it is, indicating that there may be benefits from using that layer and we may not yet have found the best way to incorporate it in our parser.

Miryam finally investigates polyglot parsing where one model is trained on data from multiple languages. She considers a multi-task scenario where each language is a task; working in this framework, she asks the question of what model parameters should be shared. Training a polyglot model for related languages leads to substantial improvements in parsing accuracy over a monolingual baseline. This is true to a smaller extent when training a polyglot model for unrelated languages. Sharing parameters that partially abstract away from word order appears to be beneficial in both cases, but sharing parameters that represent words and characters is more beneficial for related than for unrelated languages. This makes linguistic sense but might also be a result of the use of BiLSTMs as a main block in the architecture. It seems intuitive that standard BiLSTMs are particularly good at learning word order patterns due to their sequential nature.

Research Papers

Lost in Embedding Space: Explaining Cross-Lingual Task Performance with Eigenvalue Divergence

Haim Dubossarsky, Ivan Vulić, Roi Reichart, Anna Korhonen

Summary by Edoardo M. Ponti, University of Cambridge

Bilingual Lexicon Induction (BLI), the task of mapping two languages' embedding spaces onto a shared space to produce a bilingual dictionary, varies quite a lot in performance across different language pairs. In order to account for this variance it was postulated that the similarity, or isomorphism, between the embedding spaces is critical to learn a good mapping. In a recent paper, Dubossarsky et al. propose a new method to measure such similarity that is based on PCA decomposition of the embedding space, which they call EigenValue Divergence (EVD). The authors test EVD on hundreds of language pairs, and demonstrate it predicts BLI scores far better than

[1] In Tesnière (1959), auxiliary verb constructions form a nucleus, which is dissociated into several words. Finite main verbs are their non-dissociated counterpart. Nuclei are the main units of syntax.

previous methods (e.g., Isospectrality and Gromov-Hausdorff distance), thus corroborating the isomorphic hypothesis on a massive scale for the first time. The paper further reports that EVD is highly correlated with performance in several downstream tasks (dependency parsing, machine translation, and POS tagging), which the authors find to be suggestive to the ability of EVD to capture more profound aspects of language similarity that go beyond superficial differences in embedding space. Finally, the authors emphasize that EVD captures information that is complementary to typologically driven language distance measures, as their combination yields even higher correlations with performance levels in all cross-lingual tasks.

Do We Need Word Order Information for Cross-lingual Sequence Labeling

Zihan Liu, Pascale Fung

Summary by Jennifer C. White, University of Cambridge

Word order can differ significantly between languages. It is natural to expect that this affects the performance of cross-lingual models when the word order of the source language differs from the word order of the target language. In this paper, Liu and Fung aim to test the hypothesis that the removal of word order information could improve the performance of cross-lingual models on sequence labeling.

They investigate two methods for creating an order-agnostic model: removing positional encoding in the transformer encoder and obtaining word-level predictions using a linear layer with softmax rather than using a CRF layer, which prevents the model from learning interactions in the predicted output structure. Additionally, they test the use of single-headed attention in the transformer, rather than multi-headed attention, in order to avoid breaking the alignment of cross-lingual embeddings. They evaluate their models on three tasks: POS tagging, NER and dialogue NLU. Their models are trained on English and evaluated on five languages for the POS task (Spanish, French Portuguese, Russian, Greek), three languages for the NER task (German, Spanish, Dutch) and two languages for the NLU task (Spanish, Thai), without using training data in these languages.

They find that transformer-based models *without* the positional encoding consistently outperform their counterparts that retain the positional encoding across languages and tasks. Simply removing the positional encoding provides an average improvement of 10% accuracy on across languages on the POS task. Surprisingly, the improvement is also observed for languages such as French whose word order is more similar to English's. (We note, however, that French adjectives tend to appear after the noun, whereas English adjectives tend to appear before the noun.) This suggests a natural baseline that the authors appear to have omitted: comparing the transformer with and without the positional encoding on languages with identical word order to English. Indeed, without such a baseline, it is difficult to conclude that the improvements found by removing the positional

encoding are, in fact, due to word order. Additionally, the authors find that the transformer with single-headed attention slightly outperform their counterparts with multi-headed attention. However, they find that performance is worsened by removing the CRF layer, which they posit is due to losing tag–tag interactions that the CRF captures. This work offers some insight into possible ways typology influences cross-lingual methods in NLP. It would be interesting to see this work expanding on further in future, in particular by performing an evaluation on a more diverse set of languages.

Parameter Space Factorization for Zero-Shot Learning across Tasks and Languages

Edoardo M. Ponti, Ivan Vulić, Ryan Cotterell, Marinela Parovic, Roi Reichart, Anna Korhonen

Summary by Haim Dubossarsky, University of Cambridge

Given the paucity of annotated data, most combinations of NLP tasks and language varieties lack in-domain examples for supervised training. How can neural models perform sample-efficient generalization on 'unseen task–language' combinations? In this paper, Edoardo M. Ponti and colleagues propose a possible solution: a generative model of the neural parameter space, factorized into variables for several languages and tasks. This enables zero-shot classification on unseen combinations at prediction time. For instance, given training data for named entity recognition (NER) in Vietnamese and for part-of-speech (POS) tagging in Wolof, the proposed model can perform accurate predictions for NER in Wolof. In order to assess the ability of the model to generalize and to re-create a realistic setting, the authors selected a sample of 33 languages from 4 continents and 11 families. The generative model yields comparable or better results than state-of-the-art, zero-shot cross-lingual transfer methods; it increases performance by 4.49 points for POS tagging and 7.73 points for NER on average compared to the strongest baseline.

Dialectal layers in West Iranian: a Hierarchical Dirichlet Process Approach to Linguistic Relationships

Chundra A. Cathcart

Summary by Ryan Cotterell, ETH Zurich

In his paper, Chundra A. Cathcart makes an interesting methodological contribution to historical linguistics. He investigates various non-regular sound changes in the evolution of the Western Iranian languages using statistical methods. We will first explain why this problem is interesting. The neo-grammarians, a 19th century group of German historical linguistics based at the University

of Leipzig, pioneered many methods in the field. Among their most strongly held beliefs was that "sound changes know no exceptions." So what is a sound change? Let's consider the case of German and English. There is a regular relation between initial /d/ and / θ / in cognates between the two languages. For instance, observe the relation in the following pairs of German and English words: *Dieb* and *thief*, *dick* and *thick*, and *Durst* and *thirst*. For those pairs of words that are cognates, i.e. reflexes of the same proto-Germanic source, this law is nearly completely regular, i.e. there exist no exceptions. Indeed, most such sound change laws are regular in this sense, hence the neo-grammarians' assertion. With this background, Cathcart studies a known problem in the Western Iranian languages where sound change rules are often violated. His primary research contribution, from our eyes, is the application of Bayesian statistics, endeavouring to provide a principled explanation as to why the sound changes are not regular. Non-regular sound changes often arise from borrowing and language contact. Methodologically, this work builds on Cathcart's previous work on testing the Indo-Aryan Inner–Outer hypothesis--also using Bayesian statistics.

Universals of word order reflect optimization of grammars for efficient communication

Michael Hahna, Dan Jurafsky, and Richard Futrell

Summary by Ekaterina Vylomova, University of Melbourne

The study continues a strand of work that evaluates a hypothesis that language is shaped by communication and computation efficiency. Hahna and colleagues focus on Greenberg's universals of word order (Greenberg word order correlations). The authors address two major research questions: 1) whether the grammars of human languages are optimized in terms of efficiency of communication; and 2) which properties make human languages efficient. First of all, the authors formalize grammatical structures using dependency trees. In particular, they use annotations for 51 languages provided in the Universal Dependencies. In order to address the first question, the authors compare grammars of the 51 languages with 1) a sample of randomly constructed baseline grammars (that have systematic word order rules similar to natural language but no correlations among the orderings of different syntactic relations) as well as 2) grammars for each of the 51 languages that are optimized in terms of their efficiency. The authors estimate efficiency of a language as a weighted combination of parseability and predictability (negative entropy) which are obtained from language-specific neural models. The results show that the majority of real language grammars present higher predictability and/or parseability than the baseline and are close to Pareto frontier defined by optimized grammars. In the second study, the authors provide evidence that efficiency optimization accurately predicts the Greenberg correlations, while optimizing only for one part (predictability *or* parseability) does not predict all of the correlations. Finally, consistent with previous studies, both real language grammars and the ones optimized for efficiency demonstrate another dependency-length minimization, i.e. word order minimizes the average distance between syntactically related words.

Multilingual Denoising Pre-training for Neural Machine Translation

Yinhan Liu*, Jiatao Gu*, Naman Goyal*, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, Luke Zettlemoyer

Summary by Elizabeth Salesky, Johns Hopkins University

This work brings recently-popular large, masked, multilingual denoising autoencoder pretraining to machine translation (MT). They pretrain large multilingual sequence-to-sequence models with two objectives: predicting masked subsequences, and predicting the order of permuted sentences. This pretraining method is in contrast to most previous work on 'pretraining' in MT where a large general or news domain translation model, or multilingual translation model, is trained with a translation objective for later fine-tuning on a low-resource domain or language pair of interest. It is most similar to MASS (Masked Sequence-to-Sequence pretraining), which used the first objective, though trained only on sequences from the targeted language pair. They provide significant (up to 12 BLEU) improvements for low- and medium-resourced machine translation tasks, and find the improvements to be additive with back-translation. They use a large shared subword vocabulary, and provide additional experiments and analysis covering document-level MT, unsupervised MT, improvements from the inclusion of which language pairs, and typically find the inclusion of more languages is beneficial even with greater typological distance (though not when there is sufficient monolingual or bilingual data: with sufficient data for supervised pretraining on the task at hand, the pretrained weights become unnecessary or even a slightly detrimental initialization -- <1 BLEU decrease for language pairs with >25M sentences). Perhaps most interestingly from a modeling perspective, they even find improvements for language pairs (Arabic-English) with minimal lexical overlap with the pretrained model, on par with those with high coverage (Dutch-English).

Morphological Word Segmentation on Agglutinative Languages for Neural Machine Translation

Yirong Pan, Xiao Li, Yating Yang and Rui Dong

Summary by Ekaterina Vylomova, University of Melbourne

Translation into and out of morphologically rich languages has been a challenging task due to the large number of word forms in these languages. Neural machine translators typically address the problem of larger vocabulary size by means of constructing character- or ngram-level representations of word forms to share parameters among words. In this paper, authors compare

several such approaches: byte pair encoding (BPE), morphological segmentation, and a combination of both. Experiments on Turkish-to-English and Uyghur-to-Chinese translation tasks demonstrate superior performance of the latter in terms of BLEU score, suggesting that morphological segmentation (Tursun et al., 2016 for Uyghur and Zemberek for Turkish) can still enrich BPE-based methods.

Emotion semantics show both cultural variation and universal structure

Joshua Conrad Jackson, Joseph Watts, Teague R. Henry, Johann-Mattis List, Robert Forkel, Peter J. Mucha, et al.

Summary by Ekaterina Vylomova, University of Melbourne

Do words that are used to express emotions have the same meanings across cultures? To address this question, Jackson and colleagues study colefixications (cases when similar concepts are expressed with the same word within a language) in almost 2,500 languages from 20 languages families using the CLICS data that comprises ~2,400 distinct concepts, including 24 emotion concepts such as “anger”, “sadness”, “anxiety”. They first construct a universal as well as several language family-specific colexification networks (using random walk procedure) where nodes correspond to concepts, edges – to colexifications, and weights of the edges – to the number of languages that exhibit such colexifications. The next step Jackson et al. take is to cluster networks into ``concept communities'' in which concepts are more tightly related (colexified) with one another and compare the structure of the communities across families in order to estimate global variability. The results show high variability in community structures across families, but geographic proximity is an important factor here. The authors additionally compare their results with colexification of color concepts and reveal that variation in emotion concepts is significantly greater than in colors.
Finally, the authors investigate which of six psychophysiological dimensions (valence, activation, dominance, certainty, approach-avoidance, and sociality) predict the structure of emotion across families and show that valence (followed by activation) has the highest predictive power.

Resources

LowResourceEval-2019: a shared task on morphological analysis for low-resource languages

Elena Klyachko, Alexey Sorokin, Natalia Krizhanovskaya, Andrey Krizhanovsky, Galina Ryazanskaya

Summary by Ekaterina Vylomova, University of Melbourne

The paper provides an exhaustive list of resources for languages of Russia as well as a summary of the first shared task on morphological analysis for low-resource languages of Russia (Evenki, Selkup, Veps, Karelian) that consisted of three tracks: morphological analysis, inflection, and segmentation. The results demostranstate superior performance of neural models (compared to rule-based systems).

LRE Map: A database containing around 6,000 language resources and tools published at LREC conferences

A mechanism intended to monitor the use and creation of Language Resources by collecting information on both existing and newly-created resources during the submission process. It is a collective enterprise of the LREC community, as a first step towards the creation of a very broad, community-built, Open Resource Infrastructure.

Pretrained Models

OPUS-MT: over 1,000 pre-trained translation models

Pre-trained models are available at http://opus.nlpl.eu/Opus-MT/

- Number of bilingual models: 1042
- Number of multilingual models: 43
- Number of supported source languages: 180
- Number of supported target languages: 173
- Number of supported language pairs: 1421