SIGTYP

Recent Developments in Computational Typology and Multilingual Natural Language Processing

June 12 2020 · Issue #4

Editors: Ekaterina Vylomova and Ryan Cotterell

This is SIGTYP's fourth 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 Ethan A. Chi, Emanuele Bugliarello, William Lane, Dmitry Nikolaev and Omri Abend, Jens E. L. Van Gysel, and Tiago Pimentel describe their recent publications on linguistic typology and multilingual NLP.

Research Papers

Finding Universal Grammatical Relations in Multilingual BERT

Ethan A. Chi, John Hewitt, and Christopher D. Manning

Summary by Ethan A. Chi, Stanford University

Recent work has found evidence that Multilingual BERT (mBERT), a transformer-based multilingual masked language model, is capable of zero-shot cross-lingual transfer, suggesting that some aspects of its representations are shared cross-lingually. To better understand this overlap, we extend recent work on finding syntactic trees in neural networks' internal representations to the multilingual setting.

Specifically, we contribute the following:

- We extend the structural probe method of Hewitt and Manning (2019), which finds syntactic subspaces that recover syntax trees, to 10 languages other than English.
- Through zero-shot transfer experiments, we demonstrate that these syntactic subspaces overlap between languages, suggesting that Multilingual BERT stores certain syntactic features in a shared, cross-lingual space.
- We present an unsupervised analysis method that provides evidence mBERT learns representations of syntactic dependency labels, in the form of clusters which largely agree with the Universal Dependencies taxonomy. Unlike traditional probing, our method lacks supervision on UD labels, which allows us to gain insight into the latent syntactic distinctions drawn by mBERT itself. We visualize and perform a fine-grained linguistic analysis of these native distinctions.

Taken together, these results suggest that even without explicit supervision, multilingual masked language models learn certain linguistic universals, and that these universals are represented similarly across typologically divergent languages.

For more information, feel free to have a look at our blog post:
http://ai.stanford.edu/blog/finding-crosslingual-syntax/

General linguistics must be based on universals (or nonconventional aspects of language)

Martin Haspelmath

Summary by Martin Haspelmath, Leipzig University

This paper highlights the importance of the distinction between general linguistics (the study of Human Language) and particular linguistics (the study of individual languages), which is often neglected. The term "theoretical linguistics" is often used as if it entailed general claims. But I note that (unless one studies nonconventional aspects of language, e.g. reaction times as in psycholinguistics), one must study universals if one wants to make general claims. These universals can be of the Greenbergian type, based on grammatical descriptions of the speaker's social conventions, or they can be based on the natural-kinds programme, where linguists try to describe mental grammars as made up of universal building blocks of an innate grammar blueprint. The natural-kinds programme is incompatible with Chomsky's claims about Darwin's Problem, but it is indispensable for a general linguistics in the generative tradition. The Greenbergian programme, by contrast, can make use of framework-free descriptions because its comparisons are based on independently defined universal yardsticks.

It's Easier to Translate out of English than into it: Measuring Neural Translation Difficulty by Cross-Mutual Information

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, Naoaki Okazaki

Summary by Emanuele Bugliarello, University of Copenhagen

Is it easier to translate from English into Finnish or into Hungarian? And how much easier is it? Conversely, is it equally hard to translate Finnish and Hungarian into another language?

The performance of neural machine translation (NMT) systems is commonly evaluated in terms of BLEU scores. However, being a function of *n*-gram overlap between candidate and reference translations, the BLEU metric only allows for a fair comparison of the performance between models when translating into the *same* test set in the *same* target language.

In response, we propose cross-mutual information (XMI), a new metric towards cross-linguistic comparability in NMT. In contrast to BLEU, this information-theoretic quantity no longer explicitly depends on language, model, and tokenization choices. XMI hence enables the first systematic and controlled study of cross-lingual translation difficulties.

Our results show that XMI correlates well with BLEU scores when translating into the same language (where they are comparable), and that higher BLEU scores in different languages do not necessarily imply easier translations. In fact, we do find that translating *out of* English is always easier than translating *into* it!

Bootstrapping Techniques for Polysynthetic Morphological Analysis

William Lane and Steven Bird

Summary by William Lane, Charles Darwin University

Modern approaches to natural language processing are heavily influenced by English typology. As a well-resourced, analytic language, the assumptions that come with processing English are that there is plenty of data to learn from, and that morphological inflection is minimal. These assumptions simply do not hold for many of the world's languages.

A case-in-point is Kunwinjku, a polysynthetic language spoken by around 2,000 people in Northern Australia. A polysynthetic language is one where morphs combine to form a word with the semantic impact of an entire sentence in English. This morphological richness makes the vocabulary of polysynthetic languages large and sparse, by standard NLP definitions. Moreover, polysynthetic languages tend not to have the abundance of written corpora required for machine-based learning methods.

In our paper, we examine what it takes to go from few resources to a robust neural morphological analyser for a polysynthetic language like Kunwinjku. The starting point is a documentary grammar, and a Bible translation. From the grammar, we implement an FST which covers much of the morphotactics and morphophonology, but fails to account for reduplicative structure, orthographic variation, and a long tail of lexical and grammatical phenomena. In an effort to bridge this gap, we use the FST to generate morphotactically-valid surface forms, and use linguistically-motivated methods to hallucinate missing structures. Further, we create a scoring function to help rank the training data by likelihood, resampled from this ranking according to a Zipf distribution, and retrained the neural morph analyzer. The resulting system was more robust to spelling variation, was sometimes able to posit analyses for OOV lexical items, and successfully recognised all cases of reduplication. In sum, we show that even with very limited resources, it is possible to build useful models of morphology for low-resource, polysynthetic languages.

Fine-Grained Analysis of Cross-Linguistic Syntactic Divergences

Dmitry Nikolaev, Ofir Arviv, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Maria Saeboe, Omri Abend

Summary by Dmitry Nikolaev and Omri Abend

Quick adoption of the Universal Dependencies (UD) annotation scheme provided a basis for statistical in-depth studies of cross-linguistic syntactic divergences based on data from parallel corpora. This constitutes an improvement over traditional feature-based studies treating languages as vectors of categorical features (as languages are represented, e.g., in databases such as WALS or AutoTyp). However, these studies are mostly based on summary statistics over parallel corpora, such as relative frequencies of different word-order patterns, and do not reflect fine-grained cross-linguistic mappings that are very important both for linguistic typology and practical NLP applications. For example, this methodology cannot directly detect that English nominal compounds and nominal-modification constructions are often translated with Russian adjectival-modification constructions or that English adjectival-modification and nominal-modification constructions routinely give rise to Korean relative clauses.

In order to analyse such 'sub-typological' mappings we created a new cross-lingual resource—a manually word-aligned subset of the Parallel Universal Dependencies corpus collection—and conducted a quantitative and qualitative study based on it. We showed systematic correspondences between dominant syntactic patterns of English compared to other 'Standard Average European' languages (French, Russian) and more divergent languages of East Asia (Japanese, Korean, Mandarin Chinese). Ongoing work extends the analysis to further languages.

On a methodological note, the classification methodology we propose subsumes Bonnie Dorr's seminal classification of translation divergences, in the sense that Dorr's classes can largely be recast in UD terms.

We showed that our results can be largely reproduced without resorting to manual word alignment and annotation (by using off-the-shelf UD parsers and word aligners), which opens the door to productive cross-linguistic explorations. We further showed that syntactic mismatches uncovered using our method are predictive of performance of a zero-shot parser trained on English and using mBERT embeddings as input.

Cross-Lingual Semantic Annotation: Reconciling the Language-Specific and the Universal

Jens E. L. Van Gysel, Meagan Vigus, Pavlina Kalm,Sook-kyung Lee, Michael Regan, William Croft

Summary by Jens E. L. Van Gysel, University of New Mexico

The world's languages vary widely in the semantic distinctions they conventionally express. This complicates the development of multilingual semantic annotation schemes: regardless of which annotation categories a scheme proposes, there will be languages in which these semantic distinctions are not explicitly expressed, making it hard for annotators with little linguistic training to recognize them. This paper discusses how principled choices can be made as to the base level annotation categories, and how organizing annotation labels in a lattice structure can accommodate languages which defy these base level categories.

Two heuristics are proposed for establishing subdivisions of semantic domains with maximal cross-linguistic applicability. Firstly, annotation labels should be typologically informed in capturing semantic distinctions made by the majority of the world's languages. Secondly, conceptual categories chosen as base annotation labels should be highly salient in real-world experience, and definable in non-ambiguous ways, in order to allow annotators for languages which do not explicitly distinguish them to recognize them. Even if such precautions are taken, some languages will have more fine-grained categories (i.e. distinguish subcategories within the chosen annotation categories), more coarse-grained categories (i.e. lump together some of the chosen annotation categories), or cross-cutting categories (i.e. make distinctions in semantic space which overlap with those chosen for the annotation labels).

In order to solve these issues, we propose the organization of annotation labels in lattice structures, with a base level determined based on the heuristics described above, and equally typologically motivated levels of more fine-grained and more coarse-grained categories. In a lattice structure, base-level categories that are semantically intermediate between two other base-level categories may belong to multiple higher-level coarse-grained categories, reflecting the typological finding that many semantic domains are structured as hierarchical scales, where the middle value can in individual languages be categorized together with either end of the scale. Annotators will be recommended to use the base level categories, but when this is too hard because of language-specific categorizations, they will be able to choose broader, higher-level categories. If language-specific categorizations allow for more specific identification of meanings, annotators can use lower-level categories. The use of a specific lower-level value together with a more general label two levels higher is expected to facilitate annotation in languages with overlapping categories.

Because of the lattice architecture, annotations on different levels can be compared to each other, safeguarding cross-linguistic portability as shown in a small annotation pilot.

Climbing towards NLU:On Meaning, Form, and Understanding in the Age of Data

Emily M. Bender and Alexander Koller

Summary by Ekaterina Vylomova, University of Melbourne

Recent progress and success in many NLP fields led to hype in which neural models are said to ``understand'' language and capture ``meaning''. But what does it mean to ``understand'' language? In this position paper, the authors question the ability of  traditional language  models (i.e. those trained only on form) to capture any meaning.  The paper starts with examples of  misleading terminology (terms such as ``understand'' or ``comprehend'') in recent NLP papers that are overclaims if used in human-analogous sense. The authors further make a strict distinction between form (observable realization of language) and meaning (relation between communicative intent(s) (that are outside of language) and natural language expression). This way, ``understanding'' refers to retrieval of an intent(s) from an expression. By making an analogy with Searle's Chinese Room thought experiment, they outline that models should rather solve the symbol grounding problem (Harnad, 1990). The authors further propose a number of thought experiments to illustrate and strengthen their claims as well as outline possible counter-arguments.

Phonotactic Complexity and its Trade-offs

Tiago Pimentel, Brian Roark, Ryan Cotterell

Summary by Tiago Pimentel, University of Cambridge

Several measures of a language's phonological complexity exist, such as the size of its phoneme inventory (e.g. number of vowels or consonants allowed in it) and how marked it is, or the number of licit syllables in the language. Many researchers believe that as one aspect of phonological complexity becomes more complex, another should become simpler—a claim which relies on the compensation hypothesis. Consequently, researchers have looked for compensatory relationships between such complexity measures, showing for example that the vowel inventory size in a language correlates to its average word length (implicitly considering length as another complexity measure).

In this work, we present a new measure of phonotactic complexity—bits per phoneme—which permits a straightforward cross-linguistic comparison. This measure can be easily estimated for a language by relying on character-level language models, such as an LSTM-based one. These models' empirical cross-entropy in a held-out sample of wordforms upper bounds the phonotactic entropy of the language.

Using a collection of 1016 basic concept words across 106 languages (in NorthEuraLex), we demonstrate a very strong negative correlation of −0.74 between bits per phoneme and the average length of words across languages. By further controlling for language families, we show this compensatory relationship is robust to both cross- and intra-family analysis—unlike other measures which do not find compensatory relationships inside language families. Finally, we run several extra experiments to eliminate possible confounds such as phoneme position, also showing, for example, that phonemes in early positions (in a word) encode more bits than later ones.

Resources

XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning

Edoardo Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, Anna Korhonen

In order to simulate human language capacity, natural language processing systems must complement the explicit information derived from raw text with the ability to reason about the possible causes and outcomes of everyday situations. Moreover, the acquired world knowledge should generalise to new languages, modulo cultural differences. Advances in machine commonsense reasoning and cross-lingual transfer depend on the availability of challenging evaluation benchmarks. Motivated by both demands, the authors introduce Cross-lingual Choice of Plausible Alternatives (XCOPA), a typologically diverse multilingual dataset for causal commonsense reasoning in 11 languages. They benchmark a range of state-of-the-art models on this novel dataset, revealing that current methods based on multilingual pretraining and zero-shot fine-tuning transfer suffer from the curse of multilinguality and fall short of performance in monolingual settings by a large margin. Finally, the authors propose ways to adapt these models to out-of-sample resource-lean languages (such as Quechua or Haitian Creole) where only a small corpus or a bilingual dictionary is available, and report substantial improvements over the random baseline. XCOPA is available at https://github.com/cambridgeltl/xcopa.

MLSUM: The Multilingual Summarization Corpus

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano

MLSUM is the first Multilingual Summarization dataset. It is obtained from online newspapers and contains 1.5M+ article/summary pairs in five languages, namely French, Spanish, German, Russian, and Turkish. Using the dataset, the authors additionally evaluate and compare several abstractive and extractive summarization models. The authors also outline that the dataset can be useful in other areas such as multilingual question answering, news title generation, and topic detection.

Stanza: A Python Natural Language Processing Toolkit for Many Human Languages

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, Christopher D. Manning

Stanza is an open-source Python natural language processing toolkit supporting 66 human languages. Trained on 112 datasets, it can be used for text analysis, including tokenization, multi-word token expansion, lemmatization, part-of-speech and morphological feature tagging, dependency parsing, and named entity recognition. Source code, documentation, and pretrained models for 66 languages are available at https://stanfordnlp.github.io/stanza/.

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

Talks

Abralin ao Vivo – Linguists Online

Abralin ao Vivo – Linguists Online has a daily schedule of lectures and panel session with distinguished linguists from all over the world and from all subdisciplines. Most of the lectures and discussions will be in English. These activities will be broadcast online, on an open and interactive platform: abral.in/aovivo. The broadcasts will be freely available for later access on the platform aftewards.

SIGTYP 2020 (Online) – Second CFP

SIGTYP workshop is the first dedicated venue for typology-related research and its integration in multilingual NLP. The workshop is specifically aimed at raising awareness of linguistic typology and its potential in supporting and widening the global reach multilingual NLP. The topics of the workshop will include, but are not limited to:

- Language-independence in training, architecture design, and hyperparameter tuning
- Integration of typological features in language transfer and joint multilingual learning
- New applications
- Automatic inference of typological features
- Typology and interpretability
- Improvement and completion of typological databases

WE ACCEPT EXTENDED ABSTRACTS

These may report on work in progress or may be cross submissions that have already appeared in a non-NLP venue. The extended abstracts are of maximum 2 pages + references. These submissions are non-archival in order to allow submission to another venue. The selection will not be based on a double-blind review and thus submissions of this type need not be anonymized.

The abstracts should use EMNLP 2020 templates.

These should be submitted via softconf: https://www.softconf.com/emnlp2020/sigtyp/

Important Dates

— Submission Deadline: August 15th, 2020
— Retraction of workshop papers accepted for EMNLP: September 15th, 2020
— Notification of Acceptance: September 29th, 2020
— Camera-ready copy due from authors: October 10th, 2020
— Workshop: November 19th, 2020, Online