Recent Developments in Computational Typology and Multilingual Natural Language Processing

October 29, 2020 · Issue #7

Editors: Ekaterina Vylomova and Ryan Cotterell

This is SIGTYP's seventh 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 Erich R. Round, Arturo Oncevay, Taraka Rama, Jun Yen Leung, Ivan Vulić, Masakhane group, Aditi Chaudhary and Antonios Anastasopoulos, and Johannes Bjerva describe their recent publications on linguistic typology and multilingual NLP.

Research Papers

Comparability and Measurement in Typological Science: The Bright Future for Linguistics

Erich R. Round and Greville G. Corbett

Summary by Erich R. Round

Linguistic typology is fundamentally comparative, however good comparison can be deceptively intricate work. Here we tease out some knotty problems and highlight solutions in linguistics and allied fields. We make explicit what we and others see as the logical underpinnings of 'typological sciences', that is, sciences in which the very units and dimensions in which we measure our objects of enquiry are not yet settled, but are still under vigorous debate. In these circumstances, debates around definitions and operationalizations, etc. are not a sign of disarray, to be stopped. On the contrary, they're necessary ingredients of progress. What we do need, however, is debate that's productive. So how do we get it?

We focus on methods for dividing up spaces of empirical variation, and then situating that division among the many possible alternatives. This aids productive debate about those alternatives, and progress towards greater certainty. We highlight the perfectly reasonable 'typological discomfort' that typologists encounter as we find ourselves refining our tape measures (or choosing our coordinate system) at the same time as measuring our objects of inquiry. This is unnerving, but it's okay. It's what developing sciences do. Mature fields like physics reached points of settled agreement not by shunning or truncating such debate, but by letting it run productively.

As concrete illustration, we take up two subtly clever questions posed to linguists by mathematician Kolmogorov in 1956, and demonstrate how comparison and measurement can succeed within modern linguistic typology.

Though we don't touch on computational work directly, there are some rich veins that could be mined within computational typology. The computational community is well attuned to challenges around operationalization and definitions, which is an excellent start. However, what if the field in which we are working computationally is one whose definitions are in a state of productive flux, owing to ongoing, necessary debates, whose progress we want to support? Open questions include: How can we build into computational typology an orderly flexibility around units and definitions? What helpful concepts can computational typology contribute to the debate? (The reliable rigidity of names, but flexibility of alternative namespaces, perhaps?) And how can notions like gold standards in computational typology be formulated for maximal generality and utility, suited to our disciplinary circumstances of fluid debates over units, measures and metrics?

We're optimistic that more explicit discussion of these issues will speed progress in linguistic typology.

[EMNLP] Bridging Linguistic Typology and Multilingual Machine Translation with Multi-view Language Representations

Arturo Oncevay, Barry Haddow, Alexandra Birch

Summary by Arturo Oncevay

We propose to investigate and fuse the two most common approaches to obtain a language characterisation: sparse features from linguistic typology databases and learned embeddings from tasks like multilingual neural machine translation (NMT). We use singular vector canonical correlation analysis or SVCCA (Raghu et al., 2017), which first performs a singular vector decomposition that deals with the redundant information in both sources, and then project the two views into a shared space.

As each source encodes specific language relatedness knowledge, we asked: what do we represent in the shared space? By predicting typological features (like the SIGTYP Shared Task), we observed that SVCCA increased typological-awareness in the task-learned embeddings. We then reconstructed a language phylogeny, where the gold standard tree is calculated with lexicostatistics methods (Serva and Petroni, 2008) and is a proxy for lexical similarity. With a more transparent edit distance-based evaluation metric (Pawlik and Augsten, 2016), we noted that SVCCA infers the most similar phylogeny, even when the original embeddings did not include some of the languages, which we projected from the database features into the shared space.

Afterwards, we took advantage of the multi-view language vector on multilingual NMT tasks that require guidance from language relationships. For instance, language clustering (Tan et al., 2019), which identifies which groups of languages can build up smaller but still accurate multilingual NMT systems. Our SVCCA clusters achieved robust translation accuracy across different training sizes of a 53 language-pair dataset. Moreover, with a smaller SVCCA shared space of 23 entries, we projected the missing representations and still obtained clusters without compromising performance.

We also adopted a language ranking (LangRank; Lin et al., 2019) approach to identify the best candidates for multilingual transfer in NMT. Without the need to retrain a ranking model, SVCCA can choose related language-pair datasets to enhance performance in low-resource datasets similar to LangRank.

As a side outcome, we identified that using a factored neural architecture to input the language tag embedding, instead of an initial pseudo-token, encodes better language relationships.
Finally, we release our method as a tool, where everyone can use their language embeddings to compute an SVCCA shared space and perform tasks like clustering or ranking.

Testing Methods of Homeland Detection Using Synthetic Data

Taraka Rama and Søren Wichmann

Summary by Taraka Rama

There are two families of quantitative methods for inferring geographical homelands of language families: Bayesian phylogeography and the 'diversity method'. Bayesian methods model how populations may have moved using the backbone of a phylogenetic tree, while the diversity method, which does not need a tree as input, is based on the idea that the geographical area where linguistic diversity is highest likely corresponds to the homeland. No systematic tests of the performances of the different methods in a linguistic context are available, however. Here we carry out performance testing by simulating language families, including branching structures and word lists, along with speaker populations moving in areas drawn from real-world geography. We test five different methods: two versions of BayesTraits; the random walk model of BEAST; our own RevBayes implementations of a fixed rates and a variable rates random walk model; and the diversity method. Each method is tested on the same synthetic family of 20 languages, evolving in 1000 different random geographical locations. The results indicate superiority in the performance of BayesTraits and different levels of performance for the other methods, but overall no radical differences in performance.

[EMNLP] Investigating Cross-Linguistic Adjective Ordering Tendencies with a Latent-Variable Model

Jun Yen Leung, Guy Emerson, Ryan Cotterell

Summary by Jun Yen Leung

Most native speakers of a language would agree that certain adjective orderings are preferable to others. For instance, in English, "the big red dog" sounds natural while "the red big dog" sounds awkward. Similar ordering preferences have been found to apply widely across the languages in the world: for example, the adjective for "big" in most languages tends to be farther away from the noun, syntactically, than the word for "red".

Cross-linguistic adjective ordering preferences have been widely studied in the linguistics literature, and have been theorized to be universal across languages (e.g. Dixon, 1982; Sproat and Shih, 1991; Cinque, 1994, 2010; etc.). However, most of these studies have relied primarily on the

judgment of native speakers rather than on corpus data. While corpus-based models of adjective ordering do exist (e.g. Malouf, 2000; Mitchell, 2009; Hahn et al., 2018; Futrell et al., 2020; etc.), they have focused exclusively on English.

We bridge this gap by presenting a novel interpretable, multi-lingual, latent variable model of adjective ordering that directly enforces a hierarchy of semantic adjective classes and is trained entirely using corpus data. We empirically show that our model accurately orders adjectives across 24 different languages, even when tested on languages that it has not been trained on (a zero-shot setting!). In doing so, we provide converging evidence supporting the theory. While technical limitations stemming from a lack of data prevent us from being able to make a true universality claim, our methodology seems promising and can be easily applied to a more representative sample of languages in future work.

[EMNLP] Probing Pretrained Language Models for Lexical Semantics

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, Anna Korhonen

Summary by Ivan Vulić

While prior research focused on morphosyntactic, semantic, and world knowledge probing of neural language models pretrained on large corpora (e.g., monolingual and multilingual BERT, XLM-R), it remains unclear to which extent these models capture lexical type-level knowledge, how this knowledge can be extracted from their parameters, and if there are any differences across pretrained models, lexical tasks, as well as across typologically diverse languages. Therefore, we carried out a systematic empirical study on 6 languages and 5 lexical tasks, and found universally applicable strategies to extract high-quality type-level representations from pretrained language models. In general, these strategies exclude special tokens such [CLS] and [SEP] from the representations, the representations are more useful when we average lexical representations over their usage (i.e., over their different sentence-level contexts) instead of encoding them "in isolation", and our results also show that language-specific monolingual encoders outperform massively multilingual encoders. We also empirically verify that type-level lexical knowledge is distributed across multiple Transformer layers, but lower layers seem to carry more lexical knowledge. Further, we demonstrate that with some lexical knowledge extraction strategies monolingual BERTs (but not their massively multilingual counterparts) even outperform standard fastText vectors in different languages. Static word embeddings are getting undermined even in their last bulwark, lexical tasks. This begs the (provocative) question: do they still have a place in our ever-changing field? Finally, we find through Centered Kernel Alignment that translationally equivalent words receive similar representations in disjoint monolingual encoders from different languages. However, this approximate isomorphism holds only in proportion to language distance.

[EMNLP] Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages

∀ et al, Masakhane

Summary by Jade Abbott, Masakhane

Research in NLP lacks geographic diversity, and the question of how NLP can be scaled to low-resourced languages has not yet been adequately solved. “Low-resourced”-ness of a language is a complex problem that stretches far beyond mere data availability.

It may refer to overall low-usage density, non-usage in education, or endangerment, each with different implications (Cieri et al., 2016). In this complex definition, the “low-resourced”-ness is a symptom of a range of societal problems, e.g. authors oppressed by colonial governments have been imprisoned for writing novels in their languages impacting the publications in those languages (Wa Thiong’o, 1992), or that fewer PhD candidates come from oppressed societies due to low access to tertiary education (Jowi et al., 2018). There is also the problem of lack of geographic and language diversity, as well as Anglocentrism amongst NLP researchers.

We scope the study to Machine Translation (MT) and diagnose the problems of MT systems for low-resourced languages by reflecting on what agents and interactions are commonly omitted but necessary for a sustainable MT research process. We assess what impact the exclusion of {stakeholders, content creators, translators, curators, language technologists and evaluators} has on the research.

To unite the necessary agents and encourage the required interactions, we propose a participatory research to build sustainable MT research communities for low-resourced languages. The feasibility and scalability of this method is demonstrated with a case study on MT for African languages, where we present its implementation and outcomes, including novel translation datasets, benchmarks for over 30 target languages contributed and evaluated by language speakers, and publications authored by participants without formal training as scientists. Benchmarks, models, data, code, and evaluation results are released at https://github.com/masakhane-io/masakhane-mt.

- *Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. Selection criteria for low resource language programs. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 4543–4549, Portoroz, Slovenia. European Language Resources Association (ELRA).*
- *Ngugi Wa Thiong’o. 1992. Decolonising the mind: The politics of language in African literature. East African Publishers*
- *James Jowi, Charles Ochieng Ong’ondo, and Mulu Nega. 2018. Building PhD Capacity in Sub-Saharan Africa.*

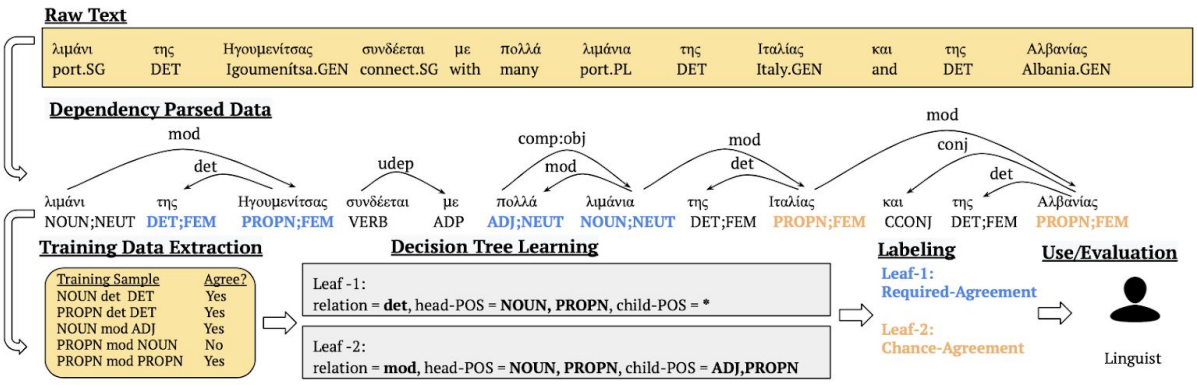
[EMNLP] Automatic Extraction of Rules Governing Morphological Agreement

Aditi Chaudhary, Antonios Anastasopoulos, Adithya Pratapa, David R. Mortensen et al.

Summary by Aditi Chaudhary and Antonios Anastasopoulos

All of the world's languages have one thing in common - an adherence to *grammar*. Creating a descriptive grammar of a language is an indispensable step for language documentation and preservation. Given the amazing diversity of languages, manually curating such grammar rules can be challenging and time-consuming, even more so when linguists are not native speakers of the language.

To assist in this process, we propose a framework to automatically extract a first-pass grammatical specification from raw text in a concise, human- and machine-readable format. In this work, we focus on extracting rules for morphological agreement.



We define *agreement rules* purely using syntactic criteria, making the simplifying assumption that a rule can be defined over a dependency edge. To extract such rules we perform syntactic analysis on raw text, predicting POS tags, morphological features and dependency trees, from which we then extract training data and train an agreement prediction model. We use decision trees for binary classification, as they are easy to interpret and allow us to easily extract the classification rules from the tree. Last, we perform *rule labeling* on the learnt tree to identify which leaves correspond to probable agreement rules.

We apply our framework to all languages from the Surface-Syntactic Universal Dependencies (SUD) treebanks, which serve as a source of gold-standard syntactic analyses, but we also experiment with zero- and few-shot parses for low-resource languages. We evaluate our framework using expert annotations and also propose an automated proxy evaluation when manual evaluation is infeasible.

The evaluation shows that our framework extracts good first-pass rules for languages/features which have well-documented agreement rules. For low-resource languages, we find that with as few as 50 expertly-annotated syntactic analysis and cross-lingual transfer our framework can produce decent first-pass agreement rules.

Explore the rules and the general framework at: https://neulab.github.io/lase/! We would love to receive any feedback and suggestions on existing or new additions to the interface.

Shared Tasks

[EMNLP] SIGTYP 2020 Shared Task: Prediction of Typological Features

Johannes Bjerva, Elizabeth Salesky, Sabrina J. Mielke, Aditi Chaudhary et al.

Summary by Johannes Bjerva

The SIGTYP 2020 Shared Task on prediction of typological features has now concluded! A big thanks to all five teams who participated, and the co-organisers who made this happen. Participating teams built systems for predicting typological features in WALS, with an evaluation set-up including controls for both phylogenetic relationships and geographic proximity. The evaluation focussed on 6 genera: Tucanoan, Madang, Mahakiranti, Nilotic, Mayan, and Northern Pama-Nyungan. These were chosen so as to give us one per macro-area in WALS, thus yielding a considerable typological spread across the world. This resulted in a challenging setting where the best system (ÚFAL, Vastl et al. (2020)) obtained a macro-averaged accuracy of 75%. The most difficult features to predict were, unsurprisingly, the ones which were the least frequent in the training data. Nonetheless, the top four systems achieved >65% accuracy on these features, whereas other systems achieved ~20% accuracy on these.

We allowed teams to participate in a constrained setting, using only the shared task data from WALS, or an unconstrained setting, in which any resources could be used. We expect that going beyond 75% macro-averaged accuracy might require the use of information from other sources beyond WALS itself, as it is challenging to infer unobserved feature values for held-out language families based on language--feature co-occurrences alone. In a potential future iteration of this shared task, we especially welcome teams to participate in this unconstrained setting, for instance investigating how information from massively multilingual texts can be used to infer typological properties.

Our shared task description paper is available on arXiv, and the workshop presentation will be available online shortly.

Blogs

Why You Should Do NLP Beyond English

Sebastian Ruder

Natural language processing (NLP) research predominantly focuses on developing methods that work well for English despite the many positive benefits of working on other languages. These benefits range from an outsized societal impact to modelling a wealth of linguistic features to avoiding overfitting as well as interesting challenges for machine learning (ML).

There are around 7,000 languages spoken around the world. Most of the world's languages are spoken in Asia, Africa, the Pacific region and the Americas.

While we have seen exciting progress across many tasks in natural language processing over the last years, most such results have been achieved in English and a small set of other high-resource languages.

This post highlights reasons to work on languages other than English from a societal, linguistic, machine learning, cultural and normative, and cognitive perspective.

Resources

Atlas of Endangered Alphabets

Indigenous and minority writing systems, and the people who are trying to save them.
Watch an introduction to the project online: https://www.youtube.com/watch?v=vGLqLQ2pc00

Note: The Endangered Alphabets Project needs two interns for the fall. Ideal for a grad student in linguistics. Contact: Tim Brookes <tim@endangeredalphabets.com>

CMU Course “Multilingual Natural Language Processing”

Taught by Graham Neubig, Yulia Tsvetkov, Alan Black. Lecture videos, slides, and homework assignments will be publicly available.

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.

Free Online course on East Caucasian languages

The East Caucasian languages form a deep-level family that is considered indigenous to the Caucasus. It consists of 30-50 distinct languages (according to different classifications), and is in fact largely responsible for the high rate of language density that the Caucasus as a linguistic area is famous for. The languages of the family feature a number of striking features, including rich consonant inventories, pervasive gender agreement with unusual targets, and complex systems of nominal spatial inflection, among other traits.