Recent Developments in Computational Typology and Multilingual Natural Language Processing

February 12, 2021 · Issue #9

Editors: Ekaterina Vylomova and Ryan Cotterell

This is SIGTYP's ninth 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 Xinjian Li, Daan van Esch, Federico Bianchi, Johannes Bjerva, Vinit Ravishankar and Artur Kulmizev, Saliha Muradoğlu, Tiago Pimentel, Kemal Kurniawan describe their recent publications on linguistic typology and multilingual NLP.

Research Papers

Universal Phone Recognition with a Multilingual Allophone System

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

Summary by Xinjian Li

There is an increasing interest in building speech tools benefiting low-resource languages, in particular, we are interested in building multilingual speech recognition models by using rich resources from languages such as English and Mandarin. Those models can improve language processing, particularly for low resource situations, by sharing parameters across a variety of languages.

However, we find there is one critical issue with the traditional multilingual acoustic models: those models generally ignore the difference between 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). This can lead to serious performance degradation when combining more than 10 training languages, as identically annotated phonemes can actually correspond to several different underlying phonetic realizations.

In this work, we propose a joint model of both language-independent phone and language-dependent phoneme distributions. Our model first computes the phone distribution using a standard ASR encoder, then we map the phone distribution into the phoneme distribution by using the allophone layer, which is a layer associating the universal narrow phone set with the phonemes of each language. This architecture allows us to solve the performance issue by distinguishing phones and phonemes explicitly. In multilingual ASR experiments over 11 languages, we find that this model improves testing performance by 2% phoneme error rate absolute in low-resource conditions.

More interestingly, because we are explicitly modeling language-independent phones, we can build a (nearly-)universal phone recognizer that, when combined with the PHOIBLE large, manually curated database of phone inventories, can be customized into 2,000 language-dependent recognizers. Experiments on two low-resourced indigenous languages, Inuktitut and Tusom, show that our recognizer achieves phone accuracy improvements of more than 17%, moving a step closer to speech recognition for all languages in the world.

We have released two tools related to this project: Allovera (https://github.com/dmort27/allovera) contains all the phone-phoneme annotations of our training languages, Allosaurus (https://github.com/xinjli/allosaurus) is the pretrained universal phone recognizer mentioned in this work

Mining Large-Scale Low-Resource Pronunciation Data From Wikipedia

Tania Chakraborty, Manasa Prasad, Theresa Breiner, Sandy Ritchie, Daan van Esch

Summary by Daan van Esch

Accurately converting from spellings (graphemes) to pronunciations (phonemes) is a key task in speech technologies like automatic speech recognition and speech synthesis, known as "G2P". There are some solid G2P toolkits out there, like Epitran and Phonetisaurus, but they usually only cover a few dozen languages at most. In previous work, we have looked at ways to make it easier for linguists to create new G2P mappings by exploiting common correspondences across languages (https://research.google/pubs/pub48581/), but we also noticed Wikipedia articles describing human languages frequently contain tables that contain basic spelling-to-pronunciation mappings. In this paper, we tried to extract these mappings automatically to extend coverage of G2P systems. However, we found that this was a rather challenging task, with the main complexity being in the large differences we observed in table lay-out from one language to the next. Such diversity is not entirely unexpected, as these tables were designed for human interpretation, not machine processing, and even in spite of this diversity, we still managed to automatically extract some possibly useful mappings (which we posted to our GitHub for others to use if they come in handy). But our main conclusion was that some additional standardization (even if perhaps through manual/editorial curation) would be required for these tables to be used at scale. Still, given the relative scarcity of G2P mappings across the world's languages, investigating this area further could be an interesting opportunity to expand G2P coverage.

[EACL2021] Cross-lingual Contextualized Topic Models with Zero-shot Learning

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, Elisabetta Fersini

Summary by Federico Bianchi

Suppose we have a small set of documents in Portuguese that is not large enough to reliably run standard topic modeling algorithms. However, we have enough English documents in the same domain. With our cross-lingual zero-shot topic model, we can first learn topics on English and then predict topics for Portuguese documents (as long as we use pre-trained representations that account for both English and Portuguese).

Topic models allow us to extract meaningful patterns from text, making it easier to glance over textual data and better understand the latent distributions of topics that live underneath. However, this kind of models usually have to deal with two limitations:

1. Once trained, most topic models cannot deal with unseen words, this is due to the fact that they are based on Bag of Words (BoW) representations, which cannot account for missing terms.

2. It is difficult to apply topic models to multilingual corpora without combining the vocabulary of multiple languages, making the task computationally expensive and without any support for zero-shot learning.

Our new neural topic model, ZeroShotTM, takes care of both problems. ZeroShotTM is a neural variational topic model that is based on recent advances in language pre-training (for example, contextualized word embedding models such as BERT).

A pre-trained representation of the documents is passed to the neural architecture and then used to reconstruct the original BoW of the document. Once the model is trained, ZeroShotTM can generate the representations of the test documents, thus predicting their topic distributions even if the documents contain unseen words during training.

Moreover, if we use a multilingual pre-trained representation during training, we can get a significant advantage at test time. Using representations that share the same embedding space allows the model to learn topic representations that are shared by documents in different languages. A trained model can then predict the topics of documents in unseen languages during training.

We trained our model on English Wikipedia data and tested it on French, Italian, German, and Portuguese Wikipedia documents. Quantitative and qualitative evaluations show that our model

can effectively predict the topics of completely unseen documents in another language than the training data.

Our topic model can be easily run on common hardware; we release an installable pip package that can be used not only to replicate our experiments but also to run topic modeling on your own text: https://github.com/MilaNLProc/contextualized-topic-models.

[EACL2021] Does Typological Blinding Impede Cross-Lingual Sharing?

Johannes Bjerva, Isabelle Augenstein

Summary by Johannes Bjerva

We investigate whether typological information is a necessity for multilingual models, by introducing an analysis method based on blinding a model with respect to categories of typological features. Our architecture is based on multilingual BERT, with an adversarial learning objective, which has the goal of prohibiting a model from learning these features. On a sample of 4 NLP tasks, with at most 40 languages, we find that blinding the model to features which are relevant to the task at hand is detrimental to the model's performance. We further find that this type of blinding discourages the model from sharing between typologically similar languages.

[EACL2021] Attention Can Reflect Syntactic Structure (If You Let It)

Vinit Ravishankar, Artur Kulmizev, Mostafa Abdou, Anders Søgaard, Joakim Nivre

Summary by Vinit Ravishankar, Artur Kulmizev

Since the popularization of the Transformer as a general-purpose feature encoder for NLP, many studies have attempted to decode linguistic structure from its novel multi-head attention mechanism. However, much of such work focused almost exclusively on English -- a language with rigid word order and a lack of inflectional morphology. In this study, we present decoding experiments for multilingual BERT across 18 languages in order to test the generalizability of the claim that dependency syntax is reflected in attention patterns. We show that full trees can be decoded above baseline accuracy from single attention heads, and that individual relations are often tracked by the same heads across languages. Furthermore, in an attempt to address recent debates about the status of attention as an explanatory mechanism, we experiment with fine-tuning mBERT on a supervised parsing objective while freezing different series of parameters.

Interestingly, in steering the objective to learn explicit linguistic structure, we find much of the same structure represented in the resulting attention patterns, with interesting differences with respect to which parameters are frozen.

Modelling Verbal Morphology in Nen

Saliha Muradoğlu, Nicholas Evans, Ekaterina Vylomova

Summary by Saliha Muradoğlu

Diversity representation of languages in NLP is vital to test the generalisations of models. We present the first-ever neural network-based analysis of Nen, the first representation of the Yam language family and to the best of our knowledge, of a Papuan language. Nen provides an interesting case study as it exhibits non-monotonic morphological mapping: distributed exponence. We compare state-of-the-art models for morphological reinflection across various training sizes and two sampling methods: random and Zipfian. The results show no significant difference between sampling methods, and minor differences may be attributed to training set composition differences. In the Zipfian case, the prefixing verb types are over-represented as they are more frequent in natural speech. We provide extensive analysis of types of errors generated by each system and show that the most common error type is allomorphy errors; a misapplication of morphophonological rules, or feature category mappings. We introduce a new subcategory of error type: free variation, which is a consequence of the natural speech origins of the corpus. We further explore composition effects by generating training sets with incremental distributions for the three verb classes noted. As expected, we found that the models trained with one class had higher prediction accuracy for that class. Across homogeneous compositions, the prefixing verb class performed the best. This is likely due to a smaller E-complexity – or more simply – a smaller combination of feature tags for which the system must learn mappings. Finally, we explore the likelihood of learning syncretic behaviour and using this as a predictor for an unseen feature bundle – the second singular past perfective. Overwhelmingly, the system incorrectly predicts syncretism with over 80% for the Aharoni & Goldbery (2017) system and 90% for the Makarov & Clementide (2018) system. These results highlight that these systems can infer patterns from the data sets provided. Although in our case the prediction of syncretism mirrors that of a human learner, there may be underlying, unwanted properties learnt from the data given, which calls for careful preparation of data and observation of output.

[EACL2021] Disambiguatory Signals are Stronger in Word-initial Positions

Tiago Pimentel, Ryan Cotterell, Brian Roark

Summary by Tiago Pimentel

Most English speakers would agree that it is easier to identify the word "mathematics" from its initial segments ("mathe-") than from its final ones ("-atics"). In fact, psycholinguistic studies provide ample evidence of the preferred nature of word-initial versus word-final segments, e.g., listeners pay greater attention to word-initial segments, while they are more likely to reduce word-final ones. This has led to the conjecture—present, for instance, in van Son and Pols (2003) and Wedel et al. (2019)—that languages have evolved to provide more information earlier in words than later.

The methods previously used to establish such tendencies in lexicons, however, have suffered from several methodological shortcomings, the most critical being their (indiscriminate) use of the conditional entropy. The issue is that conditioning reduces uncertainty, i.e. $H(X) <= H(X \mid Y)$, so segments later in the words, which have larger contexts, will naturally carry less information under this definition. Consider a language where every word contains a copy of itself: foofoo, barbar, foobarfoobar, etc. First and second halves are the same, so one could disambiguate the word equally from them. Conditional surprisal, though, would be nearly zero for the second halves. We thus question if they were measuring a linguistic phenomena, or the trivial fact that conditioning reduces entropy.

In this paper, we point out the confounds in existing methods, and present several new measures for comparing the informativeness of segments early in the word versus later in the word. When controlling for these confounds, we still find evidence across hundreds of languages that indeed there is a cross-linguistic tendency to front-load information in words. Unlike prior work, however, we conclude this effect is not universal and a few languages actually seem to put more information in word-final segments.

[EACL2021] PPT: Parsimonious Parser Transfer for Unsupervised Cross-Lingual Adaptation

Kemal Kurniawan, Lea Frermann, Philip Schulz, Trevor Cohn

Summary by Kemal Kurniawan

Cross-lingual transfer is a leading technique for parsing low-resource languages in the absence of explicit supervision. Simple `direct transfer' of a learned model based on a multilingual input encoding has provided a strong benchmark. This paper presents a method for unsupervised cross-lingual transfer that improves over direct transfer systems by using their output as implicit supervision as part of self-training on unlabelled text in the target language. The method assumes minimal resources and provides maximal flexibility by (a) accepting any pre-trained arc-factored dependency parser; (b) assuming no access to source language data; (c) supporting both projective and non-projective parsing; and (d) supporting multi-source transfer. With English as the source language, we show significant improvements over state-of-the-art transfer models on both distant and nearby languages, despite our conceptually simpler approach. We provide analyses of the choice of source languages for multi-source transfer, and the advantage of non-projective parsing. Surprisingly, a pragmatic selection of source languages, which includes only exemplary high-resource languages (mostly Indo-European), performs better than a representative selection that covers more language families. The non-projective variant of our method performs similarly to the projective counterpart, but is twice as fast and consumes less memory, which makes it favourable for parsing languages that are predominantly non-projective. Our analysis also shows that the highest gain of our multi-source transfer comes from being multilingual and not ensembling nor larger data. Our code is available online.

Shared Tasks

SIGTYP 2021: Predicting Language IDs From Speech

This year, SIGTYP is hosting a **shared task on predicting language IDs from speech.** While language ID is a fundamental speech and language processing task, it remains a challenging task in many conditions, especially when expanding the set of languages past evaluation has focused on. Further, for many low-resource and endangered languages, only single-speaker recordings may be available, demanding a need for domain and speaker-invariant language ID systems.

We selected 16 languages from across the world, some of which share phonological features, and others where these have been lost or gained due to language contact, to perform what we call robust language ID: systems will be trained on largely single-speaker speech from one domain, but evaluated on data in other domains recorded from speakers under different recording circumstances, mimicking more realistic low-resource scenarios.

For training models, we provide participants with speech data from the CMU Wilderness Dataset, which contains read speech from the Bible in 699 languages, but usually recorded from a single speaker. This training data is released in the form of derived MFCCs---please contact the organizers if you want to use another representation instead.

The evaluation will be conducted on data from different sources, in particular data from the Common Voice project, several OpenSLR corpora (SLR24, SLR35, SLR36, SLR64, SLR66, SLR79), and the Paradisec collection, testing systems' capacity to generalize to new domains, new speakers, and new recording settings. We will also use these data sources to give participants validation data in all 16 languages to test their systems.

Please see the README in our data release for the specific languages and exact data size.

Participants will be invited to describe their system in a paper for the SIGTYP workshop proceedings. The task organizers will write an overview paper that describes the task and summarizes the different approaches taken, and analyzes their results.

Important Links:

Download the data: Google Drive or OneDrive
Register for the task: Registration Form
Additional details on submission: Shared Task site

Important Dates:

Training data Release: 1 February 2021

Test data Release:	15 March 2021
Submissions Due:	31 March 2021 (AoE)
Notification:	15 April 2021
Camera-ready Due:	26 April 2021
Workshop:	10 June 2021

Organizers:
Elizabeth Salesky, Badr Abdullah, Sabrina Mielke, Gabriella Lapesa, Edoardo Ponti
Elena Klyachko, Oleg Serikov, Ritesh Kumar, Ryan Cotterell, Ekaterina Vylomova

[Dialog-21] Low-resource Speech Evaluation

The participants can use the Lingvodoc project data provided by the organizers as well as any other available data. The source code of the solutions as well as the data used must be published. All files are UTF-8 (without BOM) encoded. Every participant can make up to 3 submissions.

Subtasks:

1. Language detection. The participants will detect the language, the genus and the family for an utterance. All genera and families will be specified in the training data. However, the test data will also have surprise languages. The participants should specify X for the surprise language utterances. The data can have repetitions as well as Russian stimuli pronounced within the utterances. We suppose that the language detection task has already been accomplished in "cleaner" conditions so it would be useful to see how the solutions will perform on "field" data.
2. Speech recognition. The participants will transcribe utterances or spell them. A test dataset without repetitions will be provided.
3. Automatic detection of Russian stimuli.

Important Dates:

01.02 — training data release
21.02 — test data release
06.03 — submissions due
12.03 — results published
20.03 — papers due

Organizers: Oleg Serikov, Elena Klyachko

Resources

A Digital Corpus of St. Lawrence Island Yupik

By Lane Schwartz, Emily Chen, Hyunji Hayley Park, Edward Jahn, Sylvia L.R. Schreiner

St. Lawrence Island Yupik (ISO 639-3: ess) is an endangered polysynthetic language in the Inuit-Yupik language family indigenous to Alaska and Chukotka. This work presents a step-by-step pipeline for the digitization of written texts, and the first publicly available digital corpus for St. Lawrence Island Yupik, created using that pipeline. This corpus has great potential for future linguistic inquiry and research in NLP. It was also developed for use in Yupik language education and revitalization, with a primary goal of enabling easy access to Yupik texts by educators and by members of the Yupik community. A secondary goal is to support development of language technology such as spell-checkers, text-completion systems, interactive e-books, and language learning apps for use by the Yupik community.

Pangloss

The Pangloss Collection hosts recordings of little-documented languages, which for the most part are currently endangered. These documents are painstakingly produced by professional linguists working to rescue the world's linguistic diversity, which is currently dwindling, parallel to the world's biodiversity.

The target languages are typically studied in the field, in their geographic and social context. Dialectologists like to say that each word has a history of its own (Jaberg 1908: 6); likewise, each linguistic document has a history of its own. Linguistic resources are a result of the collaboration between the author of the document (a native speaker) and the visiting linguist, a collaboration which often extends over many years. Thus, Georges Dumézil referred to the last speaker of the Ubykh language as my teacher and friend Tevfik Esenç.

The Pangloss Collection developed over more than twenty years of sustained work by researchers and specialized engineers at CNRS. It grows year after year, through contributions that come from French research centres and their partners in various places across the globe.

As of 2020, the collection contains some 780 hours of recordings in more than 170 languages. About a half of the resources (1530 out of 3600) are transcribed, annotated and translated, allowing listeners to access the contents.

HuggingFace

HuggingFace Datasets Hub now has 467 languages and dialects!

Talks

Abralin ao Vivo – Linguists Online

Abralin ao Vivo – Linguists Online has a daily schedule of lectures and panel sessions 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 afterwards.

James McElvenny: Typology and the History of Linguistics

Typological questions have played an important role in language scholarship since at least the beginning of disciplinary linguistics in the early 19th century. The first use of the term "typology" in a specifically linguistic sense, however, would seem to have come at the end of that century, in a posthumous 1894 paper by the German linguist and sinologist Georg von der Gabelentz (1840–1893). Gabelentz' paper represents a pivotal – and yet somewhat underappreciated – moment in the history of linguistics, which links traditional concerns of 19th-century language classification with innovative, hyper-modern proposals that seem to anticipate key features of later efforts in language typology from the early 20th century up to the present day. In this talk, I will examine Gabelentz' proposals in historical context and see what lessons this history might offer to us as practising linguists today.

Workshops

[EACL2021] AfricaNLP Workshop

The second AfricaNLP workshop solicits papers that treat languages indigenous to Africa, which are some of the least resources languages in the NLP.