Recent Developments in Computational Typology and Multilingual Natural Language Processing

August 4, 2020 · Issue #5

Editors: Ekaterina Vylomova and Ryan Cotterell

This is SIGTYP's fifth 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 Richard Futrell, John Mansfield, Yushi Hu, Nathanael Erik Schweikhard, Saliha Muradoglu describe their recent publications on linguistic typology and multilingual NLP.

Research Papers

Dependency Locality as an Explanatory Principle for Word Order

Richard Futrell, Roger P. Levy, and Edward Gibson

Summary by Richard Futrell

Why are human languages the way they are? This paper adopts the Efficiency Hypothesis: languages are shaped by a pressure for efficiency, where efficiency means maximizing information transfer while minimizing the cognitive difficulty involved in producing and comprehending utterances. One source of difficulty in production and comprehension, with strong support from many behavioral experiments in psycholinguistics, is dependency length: when utterances contain words that are linked in syntactic dependencies but are far from each other in linear order, then those utterances are harder to produce and comprehend. Therefore, under the Efficiency Hypothesis, we can make a very simple prediction: languages should be structured so that, when words depend on each other syntactically, those words are also close together in linear order. This idea is called dependency locality (a.k.a. domain minimization, dependency length/distance minimization).

Dependency locality is an old idea that links together the fields of linguistic typology, psycholinguistics, corpus linguistics, parsing, and graph theory. Furthermore, dependency locality already has a lot of empirical support. In particular, it can explain many of Greenberg's universals of word order, and also certain exceptions to them. We review the philosophy and evidence behind dependency locality in our paper from a linguistics perspective. Our paper also makes several contributions to the study of dependency locality and dependency length, using Universal Dependencies datasets.

First, it provides a systematic comparison of dependency length in real sentences compared with a wide variety of new random baselines, using Universal Dependencies corpora of many languages; the result is that nearly all languages have shorter dependencies than nearly all baselines.

Second, it provides a comparison with an entirely new kind of baseline: grammatical re-orderings of corpus sentences, generated by a generative model of word order conditional on dependency trees (i.e., a surface realization model). This comparison reveals both that (1) the grammar of languages is shaped by dependency locality, and (2) beyond that, speakers' usage preferences within the parameters set by grammar are also shaped by dependency locality.

Third, it provides systematic evidence for an interesting and currently-unexplained asymmetry among languages: languages with more head-final dependencies (for example, Japanese, Korean, Turkish) have longer dependencies than languages with more head-initial dependencies (for example, English, Arabic, Indonesian). Although we offer some speculation in this paper, the real explanation for this head-final/head-initial asymmetry remains unknown.

Category Clustering: A Probabilistic Bias in the Morphology of Verbal Agreement Marking

John Mansfield, Sabine Stoll, and Balthasar Bickel

Summary by John Mansfield

It is often taken for granted that inflectional affixes of the same category will appear in the same position in the word. For example, verbs with subject and object agreement may be structured as STEM-S-O, and this is expected to be the case irrespective of which particular S and O affixes are selected. This expectation permeates much of morphological theory, and is also perhaps informally held by most linguists. We call this principle “category clustering”.

While category clustering has been widely assumed to be the norm, inspection of diverse documentary sources shows that many violations do in fact occur. One fairly common type is where affixes of the same category appear in different positions, e.g. STEM-1sgS but 2sgS-STEM. Another type, which has only been widely recognised in recent years, is the possibility of free variation in affix order, i.e. sequences of inflectional affixes that may permute without affecting meaning or grammaticality. This has been observed in several languages including Chintang, Rarámuri, Tagalog and Murrinhpatha. These morphological phenomena beg the question, can category clustering be shown to be a genuine preference in language structure?

In this study we present two convergent forms of evidence that suggest there is indeed a bias towards category clustering. Firstly, we conducted a corpus study of free prefix ordering in Chintang. The grammar allows any ordering of S, O and NEG prefixes, but do speakers prefer particular orders? We show that certain orders are indeed preferred, and crucially, these preferences are based on the category of the affix, but are not affected by its specific value. For example, 2S- and 3nsgS- show the same probabilistic biases.

Secondly, we conducted a typological study of verbal agreement paradigms from 136 languages in the AUTOTYP database. A large number of these paradigms show some degree of clustering violation, of the “ STEM-1sgS, 2sgS-STEM” type mentioned above. But when we compared the observed affix positions against a null hypothesis of random placement, we found a clear bias towards category clustering. Interestingly, although the sample showed a general bias towards clustering, three of the language families included (Algonquian, Berber, Kiranti) did not conform to

the broader pattern. Finally, we suggest that psycholinguistic research should be used to further investigate the cause and nature of the category clustering bias.

Multilingual Jointly Trained Acoustic and Written Word Embeddings

Yushi Hu, Shane Settle, Karen Livescu

Summary by Yushi Hu

Word embeddings are vector representations of words. Often these embeddings are intended to be semantic, such that proximity in the embedding space indicates similarity in word meaning, but they can also be trained to learn acoustic similarities between spoken words. Acoustic word embeddings (AWEs) are representations of spoken word segments. AWEs can be learned jointly with embeddings of their corresponding character or phone sequences to generate phonetically meaningful embeddings of written words, or acoustically grounded word embeddings (AGWEs). Such embeddings have been successfully used to improve speech retrieval, recognition, and spoken term discovery. In this work, we jointly train an AWE model and an AGWE model, using phonetically transcribed data from multiple low-resource languages. These trained models can then be applied to unseen zero-resource languages or fine-tuned further on data from low-resource languages. We find that multilingual pre-training offers significant benefits when we have only a small amount of (or no) labeled training data for the target language. We also investigate the use of distinctive features, as an alternative to phone labels, to better share cross-lingual information. We find this improves cross-lingual transfer by allowing phones unseen during training to share information with similar phones seen in the training set.

We test our models on twelve languages using two word discrimination tasks: (1) acoustic word discrimination is the task of determining whether a pair of acoustic segments correspond to the same word, while (2) cross-view word discrimination is the task of determining whether an acoustic segment and word label match.

When trained on eleven languages and tested on the remaining unseen language, our model significantly outperforms traditional unsupervised approaches like dynamic time warping. After fine-tuning the pre-trained models on one hour or even ten minutes of data from a new language, performance is typically much better than training on data from the target language alone. We also find that phonetic supervision improves performance over character sequences, and that distinctive feature supervision is helpful in handling unseen phones in the target language, especially for language with many unique phones or tones such as Cantonese and Lithuanian.

Developing an Annotation Framework for Word Formation Processes in Comparative Linguistics

Johann-Mattis List, Nathanael Erik Schweikhard

Summary by Nathanael Erik Schweikhard

Word formation plays a central role in human language. Yet computational approaches to historical linguistics often pay little attention to it. This means that the detailed findings of classical historical linguistics are often only used in qualitative studies, yet not in quantitative studies.

Based on human- and machine-readable formats suggested by the CLDF-initiative, we propose a framework for the annotation of cross-linguistic etymological relations that allows for the differentiation between etymologies that involve only regular sound change and those that involve processes of word formation.

We base our framework primarily on the distinction between linear and non-linear processes of word formation as they allow for different kinds of annotation. Etymological relations that only involve linear word formation processes (like affixation or compounding) can be easily represented by merely annotating the morpheme boundaries and noting which morphemes are cognate. This can also be applied to under-investigated languages and used to easily determine regular sound correspondences between them.

This paper introduces this approach by means of sample datasets from six different language families and a small Python library to facilitate annotation, as well as a tutorial included in the appendix.

Other types of etymological relations on the other hand necessitate a more elaborate annotation format. Therefore, as an outlook, we demonstrate how they can be visualized in derivation networks, which will be introduced in more detail in a forthcoming article. The datasets, scripts, and tutorial can be accessed on GitHub (https://github.com/digling/word-formation-paper) and Zenodo (https://zenodo.org/record/3889970).

To Compress or Not to Compress? A Finite-State Approach to Nen Verbal Morphology

Saliha Muradoglu, Nicholas Evans, Hanna Suominen

Summary by Saliha Muradoglu

Increased capacity and availability of technology has allowed for more data to be captured than ever before. This is particularly pertinent in language documentation efforts, shared amongst fieldworkers and descriptive linguists. The bottleneck of analysis in its various forms remains an issue.This paper focuses on the development of Finite-State architecture in aid of the glossing process for the language, Nen. Nen is a low-resourced language of the Morehead-Maro language family of Southern New Guinea. Approximately 300–350 people speak it in the village of Bimadbn in the Western Province of Papua New Guinea. The resources developed here feed directly into the efforts of documentation and corpus building. Aside from aiding the documentation process, the linguistic property of distributive exponence, a type of multiple exponence, makes Nen an intriguing case study for computational methods. Nen verbal morphology is particularly complex, with a transitive verb taking up to 1,740 unique features. This structural property presents interesting choices for analysis; we compare two possible methods of analysis: 'Chunking' and decomposition. 'Chunking' refers to the concept of collating morphological segments into one, whereas the decomposition model follows a more classical linguistic approach. The resultant architecture shows differences in size and structural clarity. While the 'Chunking' model is under half the size of the full decomposed counterpart, the decomposition displays higher structural order. We found no difference in the accuracies, both 80.3% overall. We present a preliminary FST for Nen.

Blogs

What's Universal in Phonetics?

A blogpost by Christian DiCanio

The blogpost attempts to outline universal and near universals in phonetics. More specifically, it discusses the following assumptions:

1) Dorsal stops (almost always) have longer VOT (voice onset time) than coronal or labial stops;
2) All languages have utterance-final lengthening;
3) Languages optimize the distance between vowels in articulation/acoustics;
4) Intrinsic F0 of high vowels;
5) Voiced stops are shorter in duration than voiceless stops;
6) (Not a universal) Word-initial strengthening;
7) (Not a universal) Native listeners of a tone language are better at pitch perception than native listeners of non-tonal languages.

Why do Linguists Talk about Universals all the Time when the Phenomena They Discuss aren't Truly Universal?

A conversation between Martin Haspelmath and Östen Dahl

The conversation follows Martin's recent paper "General linguistics must be based on universals (or nonconventional aspects of language)" and is mainly focused on the distinction between p-linguistics (particular; the study of individual languages) and g-linguistics (general; the study of Human Language). Speakers discuss terminology-related questions such as:

1) What should be referred to as "universal"? How is it different from "general"?
2) "Grams" and "morphs"

Shared Tasks on Multilinguality

Multilingual Grapheme-to-Phoneme Conversion

Kyle Gorman, Lucas F.E. Ashby, Aaron Goyzueta, Arya McCarthy, Shijie Wu, Daniel You

Participants of the shared task were asked to submit systems which take in a sequence of graphemes in a given language as input, then output a sequence of phonemes representing the pronunciation of that grapheme sequence. Nine teams submitted a total of 23 systems, at best achieving a 18% relative reduction in word error rate (macro-averaged over languages), versus strong neural sequence-to-sequence baselines.
The main results are as follows:

1) Unsurprisingly, the best systems all used some form of ensembling;

2) Many of the best teams performed self-training and/or data augmentation experiments, but most of these experiments were performance-negative except in simulated low-resource conditions. Maybe we'll do a low-resource challenge in a future year;
3) LSTMs and transformers are roughly neck-and-neck; one strong submission used a variant of hard monotonic attention;
4) Many of the best teams used some kind of pre-processing romanization strategy for Korean, the language with the worst baseline accuracy. We speculate why this helps in the task paper.

To facilitate error analysis, the complete outputs for all systems were publicly released.

Typologically Diverse Morphological Inflection

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu et al.

A broad goal in natural language processing (NLP) is to develop a system that has the capacity to process any natural language. Most systems, however, are developed using data from just one language such as English. The SIGMORPHON 2020 shared task on morphological reinflection aims to investigate systems' ability to generalize across typologically distinct languages, many of which are low resource. Systems were developed using data from 45 languages and just 5 language families, fine-tuned with data from an additional 45 languages and 10 language families (13 in total), and evaluated on all 90 languages. A total of 22 systems (19 neural) from 10 teams were submitted to the task. All four winning systems were neural (two monolingual transformers and two massively multilingual RNN-based models with gated attention). Most teams demonstrate utility of data hallucination and augmentation, ensembles, and multilingual training for low-resource languages. Non-neural learners and manually designed grammars showed competitive and even superior performance in some languages (such as Ingrian, Tajik, Tagalog, Zarma, Lingala), especially with very limited data. Some language families (Afro-Asiatic, Niger-Congo, Turkic) were relatively easy for most systems and achieved over 90% mean accuracy (for Niger-Congo this is partially due to the data sources that only listed most regular forms) while others were more challenging. For instance, some languages such as Evenki were challenging due to variations in orthography since there was little attempt at any standardization in the oral speech transcription.

Resources

Processing South Asian Languages Written in the Latin Script: the Dakshina Dataset

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke et al.

This is a new resource consisting of text in both the Latin and native scripts for 12 South Asian languages (Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Sindhi, Sinhala, Tamil, Telugu, Urdu). The dataset includes, for each language: 1) native script Wikipedia text; 2) a romanization lexicon; and 3) full sentence parallel data in both a native script of the language and the basic Latin alphabet.

https://github.com/google-research-datasets/dakshina

CoVoST: A Diverse Multilingual Speech-To-Text Translation Corpus

Changhan Wang, Juan Pino, Anne Wu, Jiatao Gu

A multilingual speech-to-text translation corpus from 11 languages (French, German, Dutch, Russian, Span-ish, Italian, Turkish, Persian, Swedish, Mongolian and Chinese) into English, diversified with over 11,000 speakers and over 60 accents.

ftfy 5.8

Robyn Speer

A tool that takes in bad Unicode (“schÃ¶n”) and outputs good Unicode (“schön”), for use in your Unicode-aware code.

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