Recent Developments in Computational Typology and Multilingual Natural Language Processing

March 19, 2021 · Issue #10

Editors: Eleanor Chodroff, Ekaterina Vylomova, Pranav A, and Ryan Cotterell

This is SIGTYP's tenth 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 Xinyi Wang, Niels van der Heijden, Muhao Chen, Ben Zhou, Yi Zhu, Toms Bergmanis, Xutan Peng, Ozan Caglayan, Janaki Sheth, Hiroaki Ozaki, Bonaventure Dossou, Chris Emezue, Albina Khusainova, Go Inoue, Federico Bianchi, Liz Salesky, Tiago Pimentel and Kyle Gorman describe their recent research on linguistic typology and multilingual NLP.

Research Papers

[NAACL 2021] Multi-view Subword Regularization

Xinyi Wang, Sebastian Ruder, and Graham Neubig

Summary by Xinyi Wang

Multilingual pretrained representations generally rely on subword segmentation algorithms to create a shared multilingual vocabulary. However, standard heuristic algorithms often lead to sub-optimal segmentation, especially for languages with limited amounts of data.

In this paper, we take two major steps towards alleviating this problem. First, we demonstrate empirically that applying existing subword regularization methods (Kudo, 2018; Provilkov et al., 2020) during fine-tuning of pre-trained multilingual representations improves the effectiveness of cross-lingual transfer. Second, to take full advantage of different possible input segmentations, we propose Multi-view Subword Regularization (MVR), a method that enforces the consistency between predictions of using inputs tokenized by the standard and probabilistic segmentations. Results on the XTREME multilingual benchmark (Hu et al., 2020) show that MVR brings consistent improvements of up to 2.5 points over using standard segmentation algorithms.

We also conduct various analyses to verify the effectiveness of our method. We found that both subword regularization and MVR deliver more improvements to non-Latin languages when using English as the source language. We also identify two potential benefits of the consistency loss in MVR.

Code: https://github.com/cindyxinyiwang/multiview-subword-regularization

[EACL 2021] Bootstrapping Multilingual AMR with Contextual Word Alignments

Janaki Sheth, Young-Suk Lee, Ramon Fernandez Astudillo, Tahira Naseem, Radu Florian, Salim Roukos, and Todd Ward

Summary by Janaki Sheth

Abstract Meaning Representation (AMR) is a popular formalism of natural language that represents the meaning of a sentence as a semantic graph. It uses an inventory of concepts and

captures “who-is-doing-what-to-whom" in a propositional style logic, abstracting away from syntactical variation.

Additionally, since AMR is agnostic about how to derive meanings from text it lends itself very well to the encoding of semantics across languages. In fact work by Damonte, Cohen in 2018 showed that it may be possible to use the original AMR annotations devised for English as representation for equivalent sentences in other languages without any modification despite translation divergence. However, developing an AMR parser for non English languages has been difficult because the existing annotated training resources that are sufficiently large are available in English only. Furthermore, acquiring semantic annotations for a large number of sentences is well-known to be a slow and expensive process in NLP.

In our work we address these concerns by exploiting transformer-based multilingual word embeddings in particular those from XLM-RoBERTa. Besides using these contextual word embeddings as input token representations, we leverage them for annotation projection, where existing AMR annotations for English are projected to a target language by using contextual word alignments.

We also combine different techniques for text-to-concept alignments and for AMR parser training which significantly improve performance over our base models. For concept alignment, we combine our proposed annotation projection method with previously established alignment techniques utilizing matching rules tailored to AMR as well as machine translation aligners (Flanigan et al., 2014; Pourdamghani et al., 2014). For AMR parser training, we pre-train an AMR parser on the treebanks of different languages simultaneously and subsequently finetune on each language.

We show that our proposed approaches not only achieve a highly competitive performance for German, Spanish, Italian and Chinese but are also easily scalable via zero-shot learning to the 100 languages included in the training set of the XLM-R multilingual transformer.

[EACL 2021] Combining Deep Generative Models and Multi-lingual Pretraining for Semi-supervised Document Classification

Yi Zhu, Ehsan Shareghi, Yingzhen Li, Roi Reichart, and Anna Korhonen

Summary by Yi Zhu

Deep generative models (DGMs) such as variational autoencoder (VAE) are capable of capturing complex data distributions at scale with rich latent representations. Instead of mapping the input x into a fixed vector z, VAE generates a distribution for the latent representation q(z | x), and

randomly samples $\tilde{z}$ from the distribution to reconstruct x, so that any $z \sim q(z \mid x)$, even not seen during training, could potentially be the latent representation of x. Semi-supervised learning with deep generative models (SDGMs) offers a framework for leveraging large amount of unlabelled data to learn better label-aware latent representations guided by little labelled data.Extending VAE which only models x and z, SDGMs also include the label y into consideration, and thus have the ability to learn directly from semi-supervised data.

Multi-lingual pretraining is another line of research shown to effectively use unlabelled data through learning shared representations across languages that can be transferred to downstream tasks. Nonetheless, the lack of labelled data still leads to inferior performance of the same model compared to those trained in languages with more labelled data such as English.

In this work, we bridge the gap to form a pipeline framework by combining SDGMs and multi-lingual pretraining for multi-lingual document classification. The pretrained model serves as multi-lingual encoder, and SDGMs can operate on top of it independently of encoding architecture. To highlight such independence, we experiment with two pretraining settings: (1) our LSTM-based cross-lingual VAE, and (2) the multi-lingual BERT (mBERT).

Experiments on document classification in several languages show that our semi-supervised framework with different encoders outperforms strong (semi-)supervised baselines including supervised mBERT, verifying that the benefits of SDGMs are orthogonal to the encoding architecture or pretraining procedure. It opens up a new avenue for SDGMs in low-resource NLP by incorporating unlabelled data potentially from different domains and languages. Our preliminary results in cross-lingual zero-shot setting are also promising, and we will continue the exploration in this direction as future work.

[EACL2021] Cross-lingual Entity Alignment with Incidental Supervision

Muhao Chen, Weijia Shi, Ben Zhou, and Dan Roth

Summary by Muhao Chen and Ben Zhou

Knowledge graphs (KG) provide sources of actionable knowledge to many intelligent applications. However, constructing the structural knowledge representations is very expensive and has often relied on massive human efforts [1]. However, knowledge is never isolated. For example, suppose we want the machine to answer this question: what is the genre of The Tale of Genji (which is the earliest existing novel and is written in Japanese)? If we go to English DBPedia, we may only find one answer, “novel.” At the same time, a Japanese KG can present more information such as

"Monogatari", "Love Story", "Royal Family Story", "Realistic Novel" and "Ancient Literature". This tells us that knowledge can be complementary in KGs curated in different languages.

The entity alignment problem is key to finding the association of knowledge across KGs. This problem seeks to identify the same entity across multiple KGs (typically curated in different languages). In this way, we allow the knowledge representations to be combined, synchronized, and jointly verified. Recently, there have been a handful of studies to tackle this problem. Most of them rely on costly direct supervision signals that are internal to KGs, including seed alignment labels, or entity profile information (see recent survey [2]).

In contrast, this work leverages cheap incidental supervision signals from external free text. The proposed method (JEANS) connects entities with available mentions in free text, and uses contextual similarity and induced lexical alignment as indirect supervision to improve entity alignment.

JEANS comprises three steps. The first step is to perform a noisy grounding process based on either an off-the-shelf entity discovery and linking (EDL) model [3] or simple surface form matching, which combines the KG entities and free text of the same language into a shared corpus. Based on the shared corpus, the second step of embedding learning captures the KG and lexemes of a language by jointly training a graph neural network and a neural language model. After that, the embedding of each language is fixed, and the alignment induction step is based on self-learning and optimal transport optimization to induce alignment for both entities and lexemes. In this case, the induced lexical alignment between two embedding spaces serves as incidental supervision signals to improve entity alignment.

Experiments on two benchmark datasets indicate the effectiveness of the incidentally supervised method, which achieves promising performance without incurring any additional labeled data.

[1] Paulheim. How much is a Triple? ISWC, 2018.
[2] Sun, et al. A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs. PVLDB, vol. 13. 2020.
[3] Khashabi, et al. CogcompNLP: Your Swiss army knife for NLP. LREC, 2018.

[EACL 2021] Cross-lingual Visual Pre-training for Multimodal Machine Translation

Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia

Summary by Ozan Caglayan

Pre-trained language models have been shown to improve performance in many natural language tasks substantially. Although the early focus of such models was single language pre-training, recent advances have resulted in cross-lingual and visual pre-training methods. In this paper, we combine these two approaches to learn visually-grounded cross-lingual representations. Specifically, we extend the translation language modelling by Lample and Conneau (2019) with masked region classification and perform pre-training with three-way parallel vision and language corpora. We show that when fine-tuned for multimodal machine translation, these models obtain state-of-the-art performance. We also provide qualitative insights into the usefulness of the learned grounded representations. Our data, code and models will be publicly available at the project website: https://hucvl.github.io/VTLM

[EACL2021] Facilitating Terminology Translation with Target Lemma Annotations

Toms Bergmanis and Mārcis Pinnis

Summary by Toms Bergmanis

Most of the recent work on terminology integration in machine translation has assumed that terminology translations are already inflected in forms that are suitable for the target language sentence. In the day-to-day work of professional translators, however, it is seldom the case as translators work with bilingual glossaries where terms are given in their dictionary forms; finding the right target language form is part of the translation process. We argue that the requirement for apriori specified target language forms is unrealistic and impedes previous work's practical applicability to other than morphologically impoverished languages.

In this work, we propose to train machine translation systems using a source-side data augmentation method that annotates randomly selected source language words with their target language lemmas. Systems trained on such augmented data learn to inflect the target language lemmas, so they fit the morphosyntactic context of the target language sentence. As we show in this work, such systems are readily usable for terminology integration in real-life translation scenarios. Our experiments on terminology translation into the morphologically complex Baltic and Uralic

languages show an improvement of up to 7 BLEU points over baseline systems, which are trained on original parallel data without terminology integration, and an average improvement of 4 BLEU points over the previous work, which assumed that terminology translations are given already inflected. Likewise, the results of human evaluation indicate a 47.7% absolute improvement over the previous work in term translation accuracy when translating into Latvian.

Although this work's primary merit is a practical contribution to the field of neural machine translation, it also serves as a case in point to illustrate why methods in machine translation should be developed and tested on more than just a few typologically similar Western European languages.

[EACL 2021] Multilingual and Cross-lingual Document Classification: A Meta-Learning Approach

Niels van der Heijden, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova

Summary by Niels van der Heijden

With the rise of large multilingual language models, cross-lingual transfer learning methods have achieved great performance gains. Yet, these methods either do not close the performance gap with a monolingual classifier in the target language or are impractically computationally expensive. We investigate meta-learning as a solution for cross-lingual document classification in previously unseen languages as well as a multilingual joint training setting with a limited set of labeled data available.

We conduct a systematic comparison of optimization- and metric-based meta-learning methods and investigate a wide array of settings in terms of the number of documents available per language during training, as well as the number of available languages. We compare these methods against strong baselines based on standard supervised learning and show that meta-learning thrives when the task distribution is heterogeneous.

We introduce a simple, yet effective modification to existing meta-learning methods which allows for better and more stable learning. Our method, ProtoMAMLn, performs above or on-par with state-of-the-art methods while requiring up to 100 times less computing power in terms of GPU hours.

[EACL 2021] Project-then-Transfer: Effective Two-stage Cross-lingual Transfer for Semantic Dependency Parsing

Hiroaki Ozaki, Gaku Morio, Terufumi Morishita, and Toshinori Miyoshi

Summary by Hiroaki Ozaki

Cross-lingual dependency parsing attracted much attention for its powerful representational capability in grammatical and semantic lexical relations. Several remarkable contributions have been made in syntactic dependency parsing, especially on universal dependencies (UD).

However, cross-lingual semantic dependency parsing, which is a totally different dependency structure from syntactic dependencies, has not been fully explored. A reason for this is the lack of parallel graph banks that cover many languages with consistent annotation policies. One exception is the Prague Semantic Dependencies (PSD), which is a treebank of bi-lexical semantic graphs and contains over 30,000 pairs of parallel annotated sentences from the Wall Street Journal in English and Czech.

Considering these circumstances, we propose to train semantic dependency parsers by capturing commonalities across languages as a remedy for the absence of massive multilingual graphbanks. Our work draws on the intuition that cross-linguality exists in both superficial level and semantic level. Accordingly, we leverage a two-stage fashion involving treebank-based transfer and model-based transfer.

Treebank-based transfer, often called annotation projection, is a method of projecting source language annotations to a target language by using a mapping function such as word alignment. The annotation projection has been reported as a promising approach under truly low-resource settings for UD parsing. However, annotation projection often suffers from noise in word alignment. For the model-based transfer, several studies on transferring contextualized word vectors have reported that it improves the parsing performance.
Our experiments on PSD graphbank indicate that the optimal performance can be achieved by incorporating the two-stage transfer.

Surprisingly, we observed improvement even when the projected treebank was erroneous.
Furthermore, the two-stage transfer method achieved almost upper-bound performance, which was approximated by evaluating the cross-linguality of PSD annotation through the projection. We also provide detailed analyses from both perspectives of cross-linguality.

[EACL 2021] Summarising Historical Text in Modern Languages

Xutan Peng, Yi Zheng, Chenghua Lin, Advaith Siddharthan

Summary by Xutan Peng

Summarising historical text is a fundamentally important routine to historians, as it can help people collect, organize, and share knowledge. Due to cultural and linguistic changes, this task is challenging even for experts, let alone being automated. In this paper, we introduce the task of summarising documents in historical forms of a language in the corresponding modern language to the NLP community. With the help of eight linguistic experts, we compile a gold-standard text summarisation dataset, which consists of historical German (sampled from the GerManC Corpus, 1600s - 1800s) and Chinese (sampled from the Wanli Gazette, 1500s - 1600s) news summarised in modern German or Chinese. Based on cross-lingual transfer learning techniques, we propose a summarisation model that can be trained even with no cross-lingual (historical to modern) parallel data, and further benchmark it against state-of-the-art algorithms. We report automatic and human evaluations that distinguish the historical to modern language summarisation task from standard cross-lingual summarisation (i.e., modern to modern language), highlight the distinctness and value of our dataset, and demonstrate that our transfer learning approach outperforms standard Transformer-based cross-lingual benchmarks on this task. We release our code and data at https://github.com/Pzoom522/HistSumm.

[AfricaNLP 2021] Crowdsourced Phrase-Based Tokenization for Low-Resourced Neural Machine Translation: The Case of Fon Language

Bonaventure Dossou and Chris Emezue

Summary by Bonaventure Dossou and Chris Emezue

Motivation: Building effective neural machine translation (NMT) models for very low-resourced and morphologically rich African indigenous languages is an open challenge. Besides the issue of finding available resources for them, a lot of work is put into preprocessing and tokenization. Recent studies have shown that standard tokenization methods do not always adequately deal with the grammatical, diacritical, and tonal properties of some African languages. That, coupled with the extremely low availability of training samples, hinders the production of reliable NMT models. In this paper we introduced WEB (Word-Expressions-based Tokenization), a human-involved super-words tokenization strategy to create a better representative vocabulary for training.

Word-Expressions-based Tokenization: The core of the WEB is a recursive greedy search algorithm which seeks and optimally breaks down sentences into words and expressions, as token units used to encode and decode sentences both at the phase of training and testing. WEB has sufficiently proved to enhance the translation's quality, compared to standard tokenization methods like subword-units (SU), word-based (WB) and phrase-based tokenization (PhB). We trained our model, using WEB as the core tokenization method, on the FFR dataset.

Dataset details: The case study is on Fon, an African native Language spoken in Benin Republic, Togo and Nigeria. The dataset contains 25k parallel Fon-French sentences, all obtained from crowdsourcing. The crowdsourcing involved approaching native and bilinguals and requesting their respective 15-20 daily most used sentences.

Results: On the Fon -> French task, WEB performed better than WB, SU and PhB, respectively by 59.8%, 59%, 27.7% (on SacreBleu metric). Similar performances are reported for the French -> Fon task. The full reports across all metrics is stated in Table 1 of the paper.

[VarDial 2021] Hierarchical Transformer for Multilingual Machine Translation

Albina Khusainova, Adil Khan, Adín Ramírez Rivera, and Vitaly Romanov

Summary by Albina Khusainova

We present our work on utilizing linguistic trees to improve machine translation quality. When building a multilingual machine translation model, the important question is how to share parameters between different languages such that to get the most from the between-language similarities. While the simplest approach is to share model parameters for different languages, there could be more effective ways. We assume that the information from linguistic trees can help build better multilingual models. Namely, the suggestion is to organize a multilingual model in a hierarchical fashion according to how languages are connected in linguistic trees: the closer two languages are, the more parameters they share.

We tested this idea using the Transformer architecture and found out that this approach is indeed capable of improving machine translation quality. However, we could only achieve a stable improvement when we fixed the problem of overfitting which we found to be specific to training hierarchical models. Regularizing low-resource directions solved the problem, substantially improving the model's performance.

We showed that the hierarchical model greatly improves low-resource pairs' scores, however, at the expense of high-resource pairs. The comparison with the full sharing model provided positive evidence supporting the assumption about the usefulness of explicitly defining parameter sharing strategy in a multilingual model.

[WANLP 2021] The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash

Summary by Go Inoue

Pre-trained language models such as BERT and RoBERTa have shown significant success in a wide range of natural language processing tasks in various languages. Arabic has benefited from extensive efforts in building dedicated pre-trained language models, achieving state-of-the-art results in a number of NLP tasks, across both Modern Standard Arabic (MSA) and Dialectal Arabic (DA).

However, it is hard to compare these models to understand what contributes to their performances because of their different design decisions and hyperparameters, such as data size, language variant, tokenization, vocabulary size, number of training steps, and so forth.
Practically, one may empirically choose the best performing pre-trained model by fine-tuning it on a particular task; however, it is still unclear why a particular model is performing better than another and what design choices are contributing to its performance.

To answer this question, we pre-trained various language models as part of a controlled experiment where we vary pre-training data sizes and language variants while keeping other hyperparameters constant throughout pre-training. We started by scaling down MSA pre-training data size to measure its impact on performance in fine-tuning tasks. We then pre-trained three different variants of Arabic: MSA, DA, and classical Arabic (CA), as well as a mix of these three variants.

We evaluate our models along with eight other recent Arabic pre-trained models across five different tasks covering all the language variants we study, namely, named entity recognition (NER), part-of-speech (POS) tagging, sentiment analysis, dialect identification, and poetry classification, spanning 12 datasets.

Our contributions can be summarized as follows: (a) We create and release eight Arabic pre-trained models, which we name CAMeLBERT, with different design decisions, including one CAMeLBERT-Mix that is trained on the largest dataset to date. (b) We investigate the interplay of

data size, language variant, and fine-tuning task type through controlled experimentation. Our results show that variant proximity of pre-training data and task data is more important than pre-training data size. (c) We exploit this insight in defining an optimized system selection model.

Our pre-trained models are available at https://huggingface.co/CAMeL-Lab, and the fine-tuning code and models are available at https://github.com/CAMeL-Lab/CAMeLBERT.

[WASSA 2021] Universal Joy: A Dataset and Results for Classifying Emotions Across Languages

Sotiris Lamprinidis, Federico Bianchi, Daniel Hardt, and Dirk Hovy

Summary by Federico Bianchi

Emotions are fundamental to the human experience and are shared across languages and cultures. We introduce a novel dataset - Universal Joy -- containing more than 500,000 Facebook posts in 18 languages, labeled for emotions.

This dataset represents an advance over prior datasets regarding the number of posts it contains and its linguistic diversity. The languages come from different typological families, and they include Bengali, Chinese, German, English, French, Hindi, Indonesian, Italian, Khmer, Burmese, Dutch, Portuguese, Romanian, Spanish, Tagalog, Thai, Vietnamese, Malay.

We use multilingual BERT to predict emotions, both within and across languages. We find consistent evidence of cross-lingual learning, including in the zero-shot setting. This is good news for the possibility of effective emotion recognition, even with low resource languages. Moreover, these results raise an intriguing question -- what does a model learn about emotion in one language, that helps it to recognize emotion in a different language? One simple answer would involve code-switching, where a post in one language contains some text from another language, is in the following Dutch post, which contains some English words:

“We love you guys ... proud of you.. [PHOTO] jongens heel veel succes vanavond ..!!!!”

By comparing the mBERT model with a bag-of-words model, we show that code-switching does not explain the cross-lingual learning. Rather, it seems that mBERT is learning something more abstract about how emotions are expressed, and these ways of expressing emotion are to some extent constant across languages. In support of this view, we show that cross-lingual learning is correlated with typological closeness of languages -- we use selected metrics that are used to quantify the similarity of languages in terms of aspects of underlying word order. Furthermore, we saw that increased linguistic diversity in the training data also improves cross-lingual learning.

With its size and linguistic diversity, the Universal Joy dataset provides a rich empirical foundation for investigation of the linguistic and conceptual underpinnings of emotion, as expressed across the languages of the world.

Dataset: https://github.com/sotlampr/universal-joy/releases/tag/dataset-models

Shared Tasks

SIGTYP 2021: Predicting Language IDs From Speech

Summary by Liz Salesky

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

SIGMORPHON–UniMorph Shared Task on Generalization in Morphological Inflection Generation

Summary by Tiago Pimentel

The sixth installment of SIGMORPHON's inflection generation shared task is divided into two parts: (1) **Generalization Across Typologically Diverse Languages**, and (2) **Are We There Yet? A Shared Task on Cognitively Plausible Morphological Inflection**.

In both parts, participants will design a model that learns to generate morphological inflections from a lemma and a set of morphosyntactic features of the target form. Each language in the task has its own training, development, and test splits. Training and development splits contain triples, each consisting of a lemma, a target form, and a set of morphological features, provided in the UniMorph format. Test splits only provide lemmas and morphological tags: your model will need to predict the missing target forms.

The first part of the shared task aims at evaluating a model's generalization across typologically diverse languages. A model should be general enough to work for natural languages of any typological patterning. For example, Tagalog verbs exhibit circumfixation; thus, a model with a

strong inductive bias towards suffixing will likely not work well for Tagalog. Like last year, this task will proceed in three phases: the Development, the Generalization, and the Evaluation phases. In the initial Development Phase, we have provided training and development splits for 35 languages which should be used to develop your system. In the Generalization Phase, we will provide training and development splits for new languages where approximately half are genetically related (belong to the same family) and half are genetically unrelated (either isolates or belonging to a different family) to the development languages. In the Evaluation Phase, the participants' models will be evaluated on held-out forms from all of the languages from the previous phases. The languages from the Development Phase and the Generalization Phase will be evaluated simultaneously. The only difference is that there has been more time to construct a model for those languages released in the Development Phase

The second part of the task investigates the open question of to what degree these morphological inflection models resemble humans in how they generate language. With this in mind, we have created a large number of new nonce words in four languages: English, German, Portuguese and Russian. To the best of our knowledge, this will be the largest and most multilingual collection of nonce words in existence. The goal of the participants in the shared task is to design a model that morphologically inflects the nonce words according to the grammar of the given languages. As an example, consider the following nonce verb that obeys English phonotactics: flink /flɪŋk/. There is arguably more than one plausible way to inflect this verb, according to English grammar; the past tense of "flink" could be either "flinked" or "flank". For that reason, we have elicited human judgements (on Amazon's Mechanical Turk) that tell native speakers' preferences towards specific past tense inflections. Participants' models will be evaluated according to their correlation with these human judgements.

Please, see the full task description here. Participants will be invited to describe their system in a paper for the SIGMORPHON workshop proceedings, while the organizers will write an overview paper about the task.

Important Links:

Find more details about the task here, together with the released data.
Please join our Google Group to stay up to date.
Click here to register for the task!

Important Dates:

Training data release: 1 March 2021
Generalization data release: 20 April 2021
Test data release: 27 April 2021
Submissions due: 4 May 2021
Description papers due: 1 June 2021
Camera-ready due: 7 June 2021

Organizers:
Tiago Pimentel, Brian Leonard, Eleanor Chodroff, Maria Ryskina, Sabrina Mielke, Garrett Nicolai, Yustinus Ghanggo Ate, Francis Tyers, Edoardo M. Ponti, Coleman Haley, Niklas Stoehr, Ritesh Kumar, Kairit Sirts, Zoey Liu, Mans Hulden, David Yarowsky, Ryan Cotterell, Ekaterina Vylomova, Ben Ambridge

Annotators:
Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Zoey Liu, Richard J. Hatcher, Emily Prud'hommeaux, Maria Ryskina, Karina Mishchenkova, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew and Natalia Krizhanovsky, Ritesh Kumar, Clara Vania, Yustinus Ghanggo Ate, Witold Kieraś, Marcin Wolinski, Totok Suhardijanto, Zahroh Nuriah, Mohit Raj, Shyam Ratan

Second SIGMORPHON Shared Task on Grapheme-to-Phoneme Conversions

Summary by Kyle Gorman

The SIGMORPHON workshop at ACL 2021 will host a shared task on multilingual grapheme-to-phoneme conversion. For this task, participants will build computational models that map a sequence of "graphemes"—characters—representing a word to a transcription of that word's pronunciation. This task is an important part of many speech technologies including recognition and synthesis.

This is the second iteration of this task. The first, held at the 2020 SIGMORPHON meeting, received 23 submissions from 9 teams. This second iteration introduces a new, stronger baseline using the imitation learning paradigm, an ensembled variant of the second-best performing system in the previous iteration of the shared task. The second iteration also introduces new languages (for a total of 21 languages in all), and splits these into three subtasks—high-, medium-, and low-resource—based on the amount of training data available, and which external resources teams are permitted to use. Finally, the data used for the 2021 shared task has been vetted using novel quality-assurance procedures.

This shared task is organized by the Computational Linguistics Lab at the Graduate Center, City University of New York and the Institut für Computerlinguistik at the University of Zurich.

Those who are interested in participating can find instructions, data, and code at the shared task website. Participant teams should also register on the shared task mailing list linked there.

Important Dates:

March 1, 2021: Data released.
March 8, 2021: Baseline code and results released.
May 1, 2021: Participants' submissions due.
May 8, 2021: Participants' draft system description papers due.
May 15, 2021: Participants' camera-ready system description papers due.

Organizers:

The task is organized by members of the Computational Linguistics Lab at the Graduate Center, City University of New York and the Institut für Computerlinguistik at the University of Zurich.

Resources

MasakhaNER: Named Entity Recognition model for 10 African Languages

∀ et al, Masakhane

Summary by Pranav A

This is a named entity recognition model which uses XLM-Roberta as a base. This has been trained on 10 African languages (Amharic, Hausa, Igbo, Kinyarwanda, Luganda, Nigerian Pidgin, Swahilu, Wolof, and Yorùbá) on MasakhaNER dataset. The training dataset involves news articles from a specific period, hence this may not generalize well for all use cases in different domains. At the time of the release of this newsletter, this is currently the state-of-the-art for NER on these languages with F1 scores ranging from 0.7 to 0.91. The dataset details are here and the HuggingFace model link is available here. This paper will be presented in the AfricaNLP workshop at EACL 2021.

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.

Explaining Diverse Language Structures From Convergent Evolution of Linguistic Conventions by Martin Haspelmath

Grammar in Language Use: A Minimalist View by David Adger