SIGTYP

Recent Developments in Computational Typology and Multilingual Natural Language Processing

April 13, 2021 · Issue #11

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

This is SIGTYP's eleventh 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 Olga Zamaraeva, Gerhard Jäger and Johannes Wahl, Mika Hämäläinen, David R. Mortensen, Bingzhi Li, Guillaume Wisniewski, Isabel Papadimitriou, Ibrahim Sharaf ElDen, Mans Hulden, Miikka Silfverberg, Alexandra Chronopoulou, Nora Hollenstein, Vilém Zouhar, Aaron Mueller, Xutan Peng,Emily P. Ahn, Aleco Kastanos, Salomey Osei, Bonaventure Dossou, Chris Emezue, Aymen Ben Elhaj Mabrouk, Nay San, Martijn Bartelds, Aviral Joshi, Mana Ashida, Seunghun J. Lee, Genta Indra Winata and Imke van Heerden describe their recent research on linguistic typology and multilingual NLP.

Dissertations

Assembling Syntax: Modeling Constituent Questions in a Grammar Engineering Framework

Olga Zamaraeva

My dissertation is concerned with a cross-linguistic account of constituent (aka *wh*-) questions integrated into the Grammar Matrix grammar engineering framework. The main goal is to present a fully implemented analysis which was automatically tested with data from multiple typologically diverse languages and in interaction with analyses of other syntactic phenomena. The implementation is in DELPH-IN HPSG which is a fully explicit syntactic formalism. In linguistics, such implemented grammars are a way of determining a clear area of applicability of hypotheses and a tool for increasing rigor and consistency in testing. In natural language processing, large grammars of this kind serve in particular to create banks of semantic annotations consistently and automatically. Large-to-medium size grammars are used also in education applications (grammar coaching). The Grammar Matrix is a system which uses stored analyses (such as the one I develop for constituent questions) to automatically output smaller, starter grammars customized to a typological and lexical specification. This means, grammars can be created quickly and for a wide typological range---and can be brought to broader coverage faster. As such, expanding the range of phenomena that the Matrix can handle potentially extends the reach of language technology to more languages of the world.

The main research question of the dissertation is: What, in formal grammar terms, comprises an analysis of the various attested ways to form constituent questions which is demonstrably compatible with analyses of other phenomena that also vary typologically? (By "varying typologically"' I mean that as the analyses I offer were driven by a review of typological literature on constituent questions, the interacting analyses that are part of the Grammar Matrix were driven by the typologies of other phenomena.) This research question is related to a big question in linguistics: What is the range of possible variation of human languages? Specifically, this work aims to contribute to this big question by providing a set of analyses which are (i) driven by typological surveys; (ii) demonstrably integrated with existing analyses; and (iii) rigorously tested. Thus, while not a claim about possibilities and impossibilities, this is a step towards establishing a range of specific syntactic analyses which are consistently successful across languages.

The scope of the analyses I offer includes: question phrase fronting, question particles, morphological marking of questions, and relevant lexical types (such as question pronouns). As part of the development, I contribute sample grammar fragments of Russian [rus] (Indo-European), English [eng] (Indo-European), Japanese [jpn] (Japonic), Chukchi [ckt] (Chukotko-Kamchatkan), and Yukaghir [yux] (Isolate), all capable of parsing and generating a range of interrogative constructions. I additionally evaluate the system on five “held-out”' languages, all from different language families which I did not consider during development. In most typological combinations which came up during evaluation, my analysis of constituent questions works as expected, including in interaction with other phenomena, such as adnominal possession, evidentials, and clausal modifiers. Some of the shortcomings are associated with the lack of support for non-verbal predicates in the Matrix as well as with the need in a more sophisticated support for morphological phenomena such as found in e.g. Wakashan languages. On the other hand, there are examples discovered in the evaluation process which highlight the fact that implemented grammars make correct predictions which a human may overlook.

Research Papers

Phylogenetic typology

Gerhard Jäger, Johannes Wahle

Summary by Gerhard Jäger and Johannes Wahle

The search for cross-linguistic correlations between the distributions of typological variables has been marred by the problem of adequate language sampling since its inception by Greenberg in 1960. Sampling only one language per family or stock limits the amount of available data, while more dense sampling violates independence assumptions of statistical tests.

In a seminal paper, Maslova (2000, “A dynamic approach to the verification of distributional universals”, Linguistic Typology) proposed to estimate diachronic transition rates between typological categories and to treat the equilibrium distribution of the resulting Markov process as an unbiased distribution.

In our paper, we present an implementation of this idea, drawing on recent advances in Bayesian phylogenetic linguistics. Our pipeline consists of the following steps. After identifying a language sample for which both typological and lexical data are available (the latter, e.g., from the Automated Similarity Judgment Program which covers more than 5,000 languages):

- obtain binary characters from the lexical data via automatic cognate detection
- infer a posterior distribution of phylogenies for each language family involved using Bayesian phylogenetic inference on lexical characters,
- estimate the transition rates between the states of the typological variables of interest via Bayesian inference, using a continuous time Markov chain model and the outcome of the previous step as prior distribution over phylogenies, where proto-languages of families, as well as isolates, are assumed to draw their value from the equilibrium distribution,
- compute the posterior distribution over equilibrium distributions.

We demonstrate this program in a study considering all 28 pairings of eight word-order variables taken from WALS. First, we demonstrate via Bayesian model comparison that a model where all lineages share the same transition rates is strongly supported by the data in comparison with a model where each lineage has its own rates. (Such a model has been suggested by Dunn et al., 2011, “Evolved structure of language shows lineage-specific trends in word-order universals”, Nature) Second, we find evidence for a correlation between the features *adposition-noun, verb-subject, verb-object, noun-genitive* and *noun-relative clause,* as well as between *noun-demonstrative, noun-numeral, noun-adjective* and *noun-relative clause*.

Endangered Languages are not Low-Resourced!

Mika Hämäläinen

Summary by Mika Hämäläinen

In this paper I identify that the term low-resourced is very problematic in the field of NLP, as there are research papers calling many big languages such as Chinese, Arabic, Hindi and Japanese low-resourced. This means that the term low-resourced has very little semantic value since it seems to be synonymous with “non-English”. Therefore, calling endangered languages low-resourced is quite an overstatement as their resources cannot, by any means, compared to the resources available for larger languages.

I refuse to believe that my native Finnish with its 5 million speakers is a low-resourced language. Although the number of speakers seems small on a global scale, it is still big enough for us to have plenty of cultural artefacts of our own from a centuries old literary tradition to modern music and movies. I argue that many of the resources do exist already for big languages; they might not come pre-annotated, but they can be crawled in a savvy way online. In fact, many resources have already been annotated by linguists, but they might not be easily available for NLP research.

When conducting NLP research on endangered languages, one will immediately find oneself dealing with very different types of problems that are not at all the same as problems solved in a “simulated low-resourced scenario”. In this paper, I will describe some of the problems starting from the lack of consistent character encoding or characters being absent from the Unicode standard to more high-level issues such as very loosely defined normative language that might not be at all internalized by native speakers. I also describe the problem that quite often NLP for endangered languages and NLP in general get polarized into two extremes: rule-based approaches and neural approaches. I outline some ways in the paper to use both of these NLP research paradigms together for endangered languages.

[EACL 2021] Cross-Cultural Similarity Features for Cross-Lingual Transfer Learning of Pragmatically Motivated Tasks

Jimin Sun, Hwijeen Ahn, Chan Young Park, Yulia Tsvetkov, David R. Mortensen

Summary by David R. Mortensen

Much work in cross-lingual transfer learning explored how to select better transfer languages for multilingual tasks, primarily focusing on typological and genealogical similarities between languages. We hypothesize that these measures of linguistic proximity are not enough when working with pragmatically-motivated tasks, such as sentiment analysis. As an alternative, we introduce three linguistic features that capture cross-cultural similarities that manifest in linguistic patterns and quantify distinct aspects of language pragmatics: language context-level, figurative language, and the lexification of emotion concepts. Our analyses show that the proposed pragmatic features do capture cross-cultural similarities and align well with existing work in sociolinguistics and linguistic anthropology. We further corroborate the effectiveness of pragmatically-driven transfer in the downstream task of choosing transfer languages for cross-lingual sentiment analysis.

[EACL 2021] Are Neural Networks Extracting Linguistic Properties or Memorizing Training Data? An Observation with a Multilingual Probe for Predicting Tens

Bingzhi Li and Guillaume Wisniewski

Summary by Bingzhi Li and Guillaume Wisniewski

There have been many recent works performing analysis to better understand what neural networks learn. Many of these works rely on a method called "*probing*" to understand which linguistic features are encoded in a sentence embedding. A probe is a supervised classifier trained to predict linguistic properties such as syntactic relations or morphological information using the neural network representation of a sentence as sole features. Probing relies on the (intuitive) hypothesis that a "high" prediction accuracy of the property (e.g. part-of-speech) from the representation (e.g. Bert) implies the property was encoded in the representation. In this work we introduce a generalization of this approach, *multilingual probing*, that consists in comparing the performance achieved by a probe on two different languages in which the studied linguistic property is expressed differently. We believe that, building on the differences between languages identified and studied by typologists, contrasting the performance of probes on different languages and comparing in which case a feature can or cannot be encoded will shed a new light on the internal working of neural network models.

As a proof of concept, we evaluate, in this work, the ability of Bert embeddings to represent tense information, taking French and Chinese as a case study. In French, the tense information is expressed by verb morphology and can be captured by simple surface information. On the contrary, tense interpretation in Chinese is driven by abstract, lexical, syntactic and even pragmatic information. We show that while French tenses can easily be predicted from sentence representations, results drop sharply for Chinese, which suggests that Bert is more likely to memorize shallow patterns from the training data rather than uncover abstract properties. In addition to its contribution to our understanding of BERT internal working, our work also highlights the interest of comparing linguistic probes across well-chosen languages which opens up a new avenue for research.

[EACL 2021] Deep Subjecthood: Higher-Order Grammatical Features in Multilingual BERT

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, Kyle Mahowald

Summary by Isabel Papadimitriou

We examine how multilingual contextual embeddings treat the property of morphosyntactic alignment (or, ergativity): what is counted as a "subject" in different languages. Nominative languages (like English) treat intransitive subjects like subjects, while ergative languages (like Basque) treat them like objects. We find that the morphosyntactic alignment of a language influences the way that contextual embeddings in that language are organized -- and this high-order information is robustly transferred cross-lingually.

We train classifiers to distinguish transitive subjects from transitive objects in the mBERT representation space for 24 different languages. We test how they classify intransitive subjects, which they've never seen, and whether that depends on the morphosyntactic alignment of the training language. We find classifiers trained on the mBERT representations of Nominative languages classify intransitive subjects in all other languages as a subject, while classifiers trained on Ergative languages are significantly less likely to do so.

Interestingly, the classifier can transfer across languages. That is, a classifier trained to predict subjects and objects based on Japanese word embeddings can reliably predict subjects and objects in French or Basque or Indonesian. We argue that this result suggests a language-general representation of grammatical role in mBERT.

How does this language-general representation of grammatical role work? We might think of a role like subjecthood as a purely syntactic concept, but the linguistics literature shows us that across languages, discourse and semantic features are important in determining subjecthood. Our classifiers can show us that mBERT's notion of subjecthood is graded and that the features discussed by linguists are also at play. Passive voice, animacy, and case all play a role in classification, even when we take out the effect of syntactic information.

Our results shed light on the way multilingual models treat higher-order grammatical features at the representation level, and they show how we can use cross-linguistic variation to understand deep neural models of language.

[NAACL 2021] From Masked-Language Modeling to Translation: Non-English Auxiliary Tasks Improve Zero-shot Spoken Language Understanding

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Ustun, Marija Stepanovic, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi and Barbara Plank

Summary by Ibrahim Sharaf ElDen

The lack of publicly available evaluation data for low-resource languages limits progress in Spoken Language Understanding (SLU). As key tasks like intent classification and slot filling require abundant training data, it is desirable to reuse existing data in high-resource languages to develop models for low-resource scenarios. We introduce **xSID** (cross-lingual slot & intent detection), a new benchmark for intent and slot prediction in 13 languages from 6 language families, including a very low-resource dialect. To tackle the challenge, we propose a joint learning approach, with English SLU training data and non-English auxiliary tasks from raw text, syntax and translation for transfer. We study two setups that differ by type and language coverage of the pre-trained embeddings.

Contributions:

1. We provide **xSID**, a new cross-lingual Spoken Language Understanding (SLU) evaluation dataset covering Arabic, Chinese, Danish, Dutch, English, German, Indonesian, Italian, Japanese, Kazakh, Serbian, Turkish and an Austro-Bavarian German dialect, South Tyrolean.

2. We experiment with new non-English auxiliary tasks for joint cross-lingual transfer on slots and intents: Universal Dependencies (UD) parsing, machine translation, and masked language modeling.

3. We compare our approach to strong baselines: two multilingual pre-trained models (mBERT and XLM-R), and a strong machine translation model (Qin et al., 2020). The pre-trained models differ as

mBERT contains most of our target languages, whereas XLM-R is pre-trained on five languages only to evaluate a real low-resource setup.

4. The dataset and code are available in this repository.

Results:

1. For the slot filling target task, Masked Language Modeling (MLM) auxiliary task has yielded the most stable performance improvements. However, when a language is not seen during pre-training (in XLM-R), UD parsing led to an even larger performance increase.

2. For intent classification target task, generating target language training data using machine translation (nmt-transfer) was outperforming all our proposed models, however, at a much higher computational cost.

3. Our analysis further shows that nmt-transfer struggles with slot filling. Given training time and data availability trade-off, MLM multi-tasking is a viable approach for SLU.

[NAACL 2021] Do RNN States Encode Abstract Phonological Processes?

Miikka Silfverberg, Francis Tyers, Garrett Nicolai, Mans Hulden

Summary by Mans Hulden and Miikka Silfverberg

In this paper, we shed light on what kind of phonological generalizations neural networks are capable of learning from data. We report on an investigation into how consonant gradation, a set of morphophonological processes which are common in Finnish and other Uralic languages, is encoded in the hidden states of an LSTM encoder-decoder model trained to perform word inflection. Specifically, we train character-based sequence-to-sequence models for inflection of Finnish nouns from the nominative into the genitive case, an inflection type which commonly triggers consonant gradation.

The process of gradation either elides consonants or strengthens them going from the nominative to the genitive, for example "katto" ~ "katon" ('roof') or "saapas" ~ "saappaan" ('boot'). Various consonants are targeted, but the locus of gradation is always the onset of the final syllable in the word. Whether a word undergoes gradation is lexically determined - some words do and others do not. Our data set consists of lexemes in Finnish that cover 17 different types of gradation processes.

We find evidence for a generalized representation of consonant gradation in our models. In many of the randomly initialized and trained models we investigated, we found particular states that activate strongly whenever gradation happens, regardless of the actual consonant ("k", "p" or "t") undergoing gradation or of the direction of gradation (strengthening or weakening). Nevertheless, we find that this unified representation is sometimes absent, indicating that the model either fails to learn a generalized representation of gradation or that gradation is encoded in a more distributed fashion among the states.

We also show that by intervening during inflection prediction by scaling the activations in the relevant dimensions we can control both whether consonant gradation occurs and the direction in which it occurs---i.e. we can artificially turn gradation 'off' for words that normally undergo it, or turn it 'on' for words that do not.

Interestingly, whether a trained model learns the linguistically 'elegant' representation where gradation is encoded in a single state seems to have no bearing on its performance.

[NAACL 2021] Improving the Lexical Ability of Pretrained Language Models for Unsupervised Neural Machine Translation

Alexandra Chronopoulou, Dario Stojanovski and Alexander Fraser

Summary by Alexandra Chronopoulou

Translation between two languages without access to parallel data is an intriguing and challenging research field that can be addressed with unsupervised neural machine translation (UNMT). There are many low-resource languages in the world, for which there is no bitext available. Unsupervised translation is a step towards removing language barriers and bringing communities closer together. However, most papers in this field focus on languages for which large amounts of monolingual data exist online, ignoring low-resource settings.

A popular pretraining method by Lample and Conneau (2019) first trains a Transformer encoder with a cross-lingual masked language modeling objective on the two languages. Then the Transformer encoder is transferred to an encoder-decoder model, which is trained for translation in an unsupervised way. In low-resource settings, the cross-lingual ability of this pretraining method is limited. This seems to be happening because the representations of the two languages are not sufficiently aligned. Motivated by this, we enhance the bilingual masked language model pretraining with lexical-level information from type-level cross-lingual subword embeddings. Specifically, we build subword monolingual embeddings with fastText for each of the two languages. We map them in a common space using VecMap with identical tokens. The embeddings are used to initialize the embedding layer of the masked language model, which is then trained on the two languages of interest.

We evaluate our approach on two language pairs, English-Macedonian and English-Albanian, both in terms of translation and lexical quality (via bilingual lexicon induction), and find out that our approach outperforms strong baselines in UNMT. Our work shows that static embeddings capture a bilingual signal at the lexical level that is complementary to the learning process of a bilingual masked language model and their use in our approach leads to a higher translation quality.

[NAACL 2021] Do Multilingual Language Models Accurately Predict Human Reading Behavior?

Nora Hollenstein, Federico Pirovano, Lena Jäger, Ce Zhang, Lisa Beinborn

Summary by Nora Hollenstein

When processing language, humans selectively attend longer to the most relevant elements of a sentence. This ability to seamlessly evaluate relative importance is a key factor in human language understanding. It remains an open question how relative importance is encoded in computational language models. In human language processing, phenomena of relative importance can be approximated indirectly by tracking eye movements and measuring fixation duration.

In this paper, we analyze if large language models are able to predict patterns of human reading behavior. We compare the performance of language-specific and multilingual pretrained transformer models to predict reading time measures reflecting natural human sentence processing on Dutch, English, German, and Russian texts. Multilingual models represent multiple languages in a joint space and aim at a more universal language understanding. As eye tracking patterns are consistent across languages for certain phenomena, we hypothesize that multilingual models might provide cognitively more plausible representations and outperform language-specific models in predicting reading measures.

We find that pretrained transformer models are surprisingly accurate at predicting reading time measures in four all Indo-European languages. Multilingual models show an advantage over language-specific models, especially when fine-tuned on smaller amounts of data. Our results indicate that transformer models implicitly encode relative importance in language in a way that is comparable to human processing mechanisms. As a consequence, it should be possible to adjust the inductive bias of neural models towards more cognitively plausible outputs without having to resort to large-scale cognitive datasets.

Both monolingual and multilingual models achieve surprisingly high accuracy in predicting a range of eye tracking features across four languages. Compared to the XLM models, multilingual BERT is more robust in its ability to generalize across languages, without being explicitly trained for it. In contrast, the XLM models perform better when fine-tuned on less eye tracking data. We also observe that the models learn to reflect characteristics of human reading such as the word length effect and higher accuracy in more easily readable sentences.

[NAACL 2021] Backtranslation Improves Users Confidence in MT, Not Quality

Vilém Zouhar, Michal Novák, Matúš Žilinec, Ondřej Bojar, Mateo Obregón,Robin L. Hill, Frédéric Blain, Marina Fomicheva, Lucia Specia, and Lisa Yankovskaya

Summary by Vilém Zouhar

Most research in machine translation (MT) focuses on inbound translation, where the recipients of the translation are aware of the MT process, and thus it is largely their responsibility to interpret and understand the translated content correctly. For outbound translation, it is the other way round: the responsibility to create the content in a way that it is correctly interpreted by a recipient lies on its authors. The main issue is that the target language might be entirely unknown to them. Prototypically filling in foreign forms in a foreign language or communicating with an IT support. The focus of MT is placed not only on producing high-quality translations but also on reassuring the author that the MT output is correct. Basic machine translation facilities are often not best suited for such a scenario.

In this paper, we propose three ways of affecting authors' confidence during translation as well as final translation quality: backward translation, quality estimation (with alignment) and source paraphrasing. Using a web-based environment and a dataset focused on the e-commerce domain, we examined the effects of each proposed feedback module in experiments translating English into Czech and Estonian.

Our results showed that backward translation proves to be a powerful means to enhance user confidence in MT. Backward translation simultaneously neither significantly increases nor decreases the translation quality. The fact that backward translation has a marginal effect on objective quality but greatly increases user confidence is surprising because it is the most intuitive low-effort approach to outbound translation scenarios even with publicly available MT systems. Showing paraphrases seems to increase user confidence less with only slightly negative or no impact on translation quality. Without a better method to generate diverse and still adequate paraphrases, employing this cue is questionable. The effect of word-level quality estimation appears to be even more questionable. We attribute it mainly to the underlying word-level models, which may not be mature enough for user-facing applications.

We further focused on how the quality of MT systems influenced these findings and the users' perception of success. We mainly compared a strong but slow model and its speed-optimized student version. Despite a lower quality of translations, using the student model resulted in slightly higher average self-reported user confidence. We hypothesize it either managed to maintain its teacher's high trustworthiness or compensated for it by its speed.

[NAACL 2021] Fine-tuning Encoders for Improved Monolingual and Zero-shot Polylingual Neural Topic Modeling

Aaron Mueller, Mark Dredze

Summary by Aaron Mueller

Neural topic models can augment or replace bag-of-words inputs with the learned representations of deep pre-trained transformer-based word prediction models. One added benefit when using representations from multilingual models is that they facilitate zero-shot polylingual topic modeling. However, while it has been widely observed that pre-trained embeddings should be fine-tuned to a given task, it is not immediately clear what supervision should look like for an unsupervised task such as topic modeling.

Thus, we propose several methods for fine-tuning encoders to improve both monolingual and zero-shot polylingual neural topic modeling. We consider fine-tuning on existing auxiliary tasks, bootstrapping direct supervision a new topic classification task, integrating the topic classification objective directly into topic model training, and continued pre-training. We find that fine-tuning encoder representations on topic classification and integrating the topic classification task directly into topic modeling improves monolingual topic quality, and that fine-tuning encoder representations on any task is the most important factor for facilitating cross-lingual transfer.

We also perform qualitative analyses to better understand why fine-tuning induces better representations for polylingual neural topic modeling. In particular, we find that performance on the existing Semantic Textual Similarity (STS) benchmark after fine-tuning correlates strongly with cross-lingual transfer, but does not correlate strongly with monolingual topic quality.

[NAACL 2021] Cross-Lingual Word Embedding Refinement by $\ell1$ Norm Optimisation

Xutan Peng, Chenghua Lin, Mark Stevenson

Summary by Xutan Peng

Cross-Lingual Word Embedding (CLWE) techniques have recently received significant attention as an effective means to support Natural Language Processing applications for low-resource languages. Mainstream CLWE methods train orthogonal mappings by minimising the topological dissimilarity between source and target embeddings based on $\ell2$ loss (aka. Frobenius loss or squared error). This learning strategy takes advantage of a very elegant closed-form solution (Schönemann, 1966) and thus greatly simplifies the computation required. However, the $\ell2$ goodness-of-fit criterion has been demonstrated to be sensitive to outliers in fields such as Computer Vision and Data Mining. Empirically we observe similar outlier issues which harm CLWE quality. In this paper, we develop a simple yet effective post-processing technique based on the more robust Manhattan norm (aka. $\ell1$

norm) optimisation objective. This approach is fully agnostic to the training process of the original CLWEs and can therefore be applied widely. Extensive experiments are performed involving 10 diverse languages and embeddings trained on different corpora. Evaluation results on Bilingual Lexicon Induction and cross-lingual Natural Language Inference tasks show that the proposed ℓ1 refinement substantially outperforms 4 state-of-the-art baselines in both supervised and unsupervised settings. It is therefore recommended that this algorithm be adopted as a standard for CLWE methods. We will soon release our code at https://github.com/Pzoom522/L1-Refinement

[ComputEL] Developing a Shared Task for Speech Processing on Endangered Languages

Gina-Anne Levow, Emily P. Ahn, Emily M. Bender

Summary by Emily P. Ahn

Advances in speech and language processing have enabled the creation of applications that could, in principle, accelerate the process of language documentation, as speech communities and linguists work on urgent language documentation and reclamation projects. However, such systems have yet to make a significant impact on language documentation, as resource requirements limit the broad applicability of these new techniques. We aim to exploit the framework of shared tasks to focus the technology research community on tasks which address key pain points in language documentation.

Here we present initial steps in the implementation of these new shared tasks, through the creation of data sets drawn from endangered language repositories and baseline systems to perform segmentation and speaker labeling of these audio recordings---important enabling steps in the documentation process. This paper motivates these tasks with a use case, describes data set curation and baseline systems, and presents results on this data. We then highlight the challenges and ethical considerations in developing these speech processing tools and tasks to support endangered language documentation.

Data sets used in this work are drawn from 8 different languages: Cicipu, Effutu, Mocho', Northern Prinmi, Sakun, Upper Napo Kichwa, Toratán, and Ulwa. We plan to launch this shared task in the near future.

[AfricaNLP 2021] Graph Convolutional Network for Swahili News Classification

Aleco Kastanos and Tyler Martin

Summary by Aleco Kastanos

African languages are underrepresented in both the academic field of natural language processing (NLP) as well as the industry setting. This leads to a shortage of annotated datasets, purpose-built software tools, and limited literature comparing the efficacy of techniques developed for high-resource languages in a low-resource context. Swahili, being the most widely spoken African language, is no exception to this trend.

Our work attempts to address this disparity by providing a set of accessible traditional NLP benchmarks for the task of semi-supervised Swahili news classification. These baseline models include TF-IDF representations, pre-trained fastText embeddings, and Distributed Bag of Words representations each followed by a logistic regression layer. We draw particular attention to the semi-supervised context as this most accurately exemplifies the annotation constraints facing many Swahili text classification tasks.

Graph Neural Networks, a family of model architectures that can leverage implicit inter-document and intra-document relationships, has demonstrated remarkable performance in semi-supervised text classification tasks. As a result, we apply a Text Graph Convolution Network (Text GCN) to our news classification task. To our knowledge, this is the first time a Graph Neural Network has been for text classification for any African language. The experiments demonstrate that Text GCN outperforms the previously described baselines, especially when the proportion of labelled training set documents is reduced below 5% of the full training set. Furthermore, we present a Text GCN variant that is successfully able to maintain similar predictive performance whilst reducing the memory footprint and cloud cost of the model by factors of 5 and 20 respectively. This is achieved by replacing the naive one-hot representation of the nodes in the Text GCN graph with an appropriate bag of words representation.

The empirical results demonstrate the ability of graph-based models to outperform traditional baseline models for this task. We hope that the experimental results and freely available code contribute to addressing the shortage of accessible resources for semi-supervised text classification in Swahili.

[AfricaNLP 2021] English-Twi Parallel Corpus for Machine Translation

Paul Azunre, Salomey Osei, Salomey Afua Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac Kojo Essel Ampomah, Joseph Otoo, Reindorf

Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James Ben Hayfron-Acquah

Summary by Salomey Osei

We present a parallel machine translation training corpus for English and Akuapem Twi of 25,421 sentence pairs. We used a transformer-based translator to generate initial translations in Akuapem Twi, which were later verified and corrected where necessary by native speakers to eliminate any occurrence of translationese. In addition, 697 higher quality crowd-sourced sentences are provided for use as an evaluation set for downstream Natural Language Processing (NLP) tasks. The typical use case for the larger human-verified dataset is for further training of machine translation models in Akuapem Twi. The higher quality 697 crowd-sourced dataset is recommended as a testing dataset for machine translation of English to Twi and Twi to English models. Furthermore, the Twi part of the crowd-sourced data may also be used for other tasks, such as representation learning, classification, etc. We fine-tune the transformer translation model on the training corpus and report benchmarks on the crowd-sourced test set.

All our projects are open source and can be found at GitHub repo and our website.

The corpus is available here. Neural Machine Translation (Khaya App) Translators for English to Twi, Ewe and Ga are available at: Web app, Android app and IOS.

[AfricaNLP 2021] Contextual Text Embeddings for Twi

Paul Azunre, Salomey Osei, Salomey Afua Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac Kojo Essel Ampomah, Joseph Otoo, Reindorf Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James Ben Hayfron-Acquah

Summary by Salomey Osei

Transformer-based language models have been changing the modern Natural Language Processing (NLP) landscape for high-resource languages such as English, Chinese, Russian, etc. However, this technology does not yet exist for any Ghanaian language. In this paper, we introduce the first of such models for Twi or Akan, the most widely spoken Ghanaian language. The specific contribution of this research work is the development of several pretrained transformer language models for the Akuapem and Asante dialects of Twi, paving the way for advances in application areas such as

Named Entity Recognition (NER), Neural Machine Translation (NMT), Sentiment Analysis (SA) and Part-of-Speech (POS) tagging. Specifically, we introduce four different flavours of ABENA -- A BERT model Now in Akan that is fine-tuned on a set of Akan corpora, and BAKO - BERT with Akan Knowledge only, which is trained from scratch. We open-source the model through the Hugging Face model hub and demonstrate its use via a simple sentiment classification example.

All our projects are open source and can be found at GitHub repo and our website.
The demo link is available here.

[AfricaNLP 2021] NLP for Ghanaian Languages

Paul Azunre, Salomey Osei, Salomey Afua Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac Kojo Essel Ampomah, Joseph Otoo, Reindorf Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James Ben Hayfron-Acquah

Summary by Salomey Osei

NLP Ghana is an open-source non-profit organisation aiming to advance the development and adoption of state-of-the-art NLP techniques and digital language tools to Ghanaian languages and problems. In this paper, we first present the motivation and necessity for the efforts of the organisation; by introducing some popular Ghanaian languages while presenting the state of NLP in Ghana. We then present the NLP Ghana organisation and outline its aims, scope of work, some of the methods employed and contributions made thus far in the NLP community in Ghana.

Current work of NLP Ghana is focused on Ghanaian Languages. There are over 75 local languages spoken in Ghana with Government sponsored languages being Fante, Akuapem Twi, Asante Twi, Ewe, Dagaare, Dagbani, Dangme, Ga, Gonja, Kasem, Nzema etc. We plan to develop better data sources to train state-of-the-art (SOTA) NLP techniques for these Ghanaian languages, contribute to adapting SOTA techniques to work better in a lower resource setting and build functional systems for local applications, e.g., a “Google Translate” for Twi, Ewe, Frafra, etc.

Some of the targets that we have: create datasets for NLP tasks in low-resourced languages, train and benchmark algorithms for NER, POS and sentiment analysis, unsupervised spell-checking methods, automatic speech recognition, virtual assistant (e.g. Siri, Alexa).

All our projects are open source and can be found at GitHub repo and our website.

[AfricaNLP 2021] OkwuGbé: End-to-End Speech Recognition for Fon and Igbo

Bonaventure Dossou and Chris Emezue

Summary by Bonaventure Dossou & Chris Emezue

African languages have recently been the subject of research in natural language processing. However, there are few works being done on speech for these African languages, as more emphasis is being placed on their text. Due to the largely acoustic nature of African languages (mostly tonal, diacritical, etc), a careful speech analysis of African languages could provide better insight for text-based NLP involving African languages, as well as supplement the textual data needed for machine translation or language modelling. This is what inspired OkwuGbé, a step towards building speech recognition systems for African low-resourced languages. OkwuGbé, the union of two words from Igbo (Okwu, which means speech) and Fon (Gbé, which means languages) signifies studying and integrating automatic speech recognition to several African languages in an effort to unify them.

In this study, we give an in-depth linguistic analysis of Fon and Igbo languages. Then for both languages, we present our best model architecture made up of 5-blocks of residual convolutional neural networks and 3-blocks each of Bidirectional Long Short Term Memory (BiLSTMs) and Bidirectional Gated Recurrent Units (BiGRUs). The unique features of our model are 1) the exploration of the combination of BiLSTMs and BiGRUs and 2) the integration of the Bahdanau attention mechanism. Throughout our experiments, we explored different model architectures with various convolutional and bidirectional recurrent layers. Our experiments show that implementing attention mechanisms reduced the WER by 5% for Fon language, outperforming the baseline model.

A Multilingual African Embedding for FAQ Chatbots

Aymen Ben Elhaj Mabrouk, Moez Ben Haj Hmida, Chayma Fourati, Hatem Haddad, Abir Messaoudi

Summary by Aymen Ben Elhaj Mabrouk

Searching for an available, reliable, official, and understandable information is not a trivial task due to scattered information across the Internet, and the lack of governmental communication channels with African dialects and languages.

In this paper, we introduce an Artificial Intelligence Powered chatbot for crisis communication that would be omnichannel, multilingual and multi dialectal. We present our work on modified StarSpace embedding tailored for African dialects for the question-answering task along with the architecture of the proposed chatbot system and a description of the different layers. English, French, Arabic, Tunisian, Igbo, Yorùbá, and Hausa are used as languages and dialects. Quantitative and qualitative evaluation results are obtained for our real deployed Covid-19 chatbot.

This solution solves digital communication with citizens using local dialects, avoiding fake news, understanding and replying using the relevant language or dialect of the user without any predefined scenarios. Our solution digitalizes African institutions Services and hence, illiterate, and vulnerable citizens such as rural women and people that need social inclusion and resilience would have easier access to reliable, official, and understandable information 24/7.

Results show that users are satisfied and the conversation with the chatbot is meeting customer needs. Indeed, statistics demonstrate the high interaction rate with the chatbot. The architecture used can be tailored to any tokenized African dialect. Hence, the ability to make other chatbots for other underrepresented African languages and dialects.

[Interspeech 2021] Leveraging neural representations for facilitating access to untranscribed speech from endangered languages

Nay San, Martijn Bartelds, Mitchell Browne, Lily Clifford, Fiona Gibson, John Mansfield, David Nash, Jane Simpson, Myfany Turpin, Maria Vollmer, Sasha Wilmoth, Dan Jurafsky

Summary by Nay San and Martijn Bartelds

Language documentation efforts often yield a sizeable amount of untranscribed speech, which is difficult to index and search. These difficulties have a direct impact on how easily such resources may be used for language maintenance and revitalisation activities by many interested parties. For languages with insufficient resources to train speech recognition systems, query-by-example spoken term detection (QbE-STD) offers a way of accessing an untranscribed speech corpus by helping identify regions where spoken query terms occur. Yet retrieval performance can be poor when the query and corpus are spoken by different speakers and produced in different recording conditions.

In this paper, we evaluated whether QbE-STD performance on these languages could be improved by leveraging representations extracted from the pre-trained English wav2vec 2.0 model, using data selected from a variety of speakers and recording conditions from 7 Australian Aboriginal languages and a regional variety of Dutch (Gronings), all of which are endangered or vulnerable. Compared to the use of Mel-frequency cepstral coefficients and bottleneck features, we find that representations from the middle layers of the wav2vec 2.0 Transformer offer large gains in task performance. For example, in the worst of cases, retrieval performance improved 56–86% from only 27–28% of queries being retrievable to 42–52% when using wav2vec 2.0 features — a tolerable operating range, given the alternative is browsing untranscribed audio in near real-time.

While features extracted using the pre-trained English model yielded improved detection on all the evaluation languages, better detection performance was associated with the evaluation language's phonological similarity to English. Due to the consonantal inventory of English, the representations of the w2v2 English model do not appear to be sufficiently fine-grained to differentiate between place of articulation contrasts within certain classes (e.g. dental vs. retroflex in nasal consonants), yielding erroneous retrievals in QbE-STD template matching for low-resource Australian languages. Given this finding, it may be possible to reduce these types of errors by using a w2v2 model trained on a relatively higher resource language with a similar consonantal inventory (e.g. a Dravidian language such as Tamil).

Code, Data and experiment artefacts

Zero-Shot Language Transfer vs Iterative Back Translation for Unsupervised Machine Translation

Aviral Joshi, Chengzhi Huang, Har Simrat Singh

Summary by Aviral Joshi

This work focuses on comparing different solutions for machine translation on low resource language pairs, namely, zero-shot transfer learning and unsupervised machine translation. We discuss how the data size affects the performance of both unsupervised MT and transfer learning. Additionally, we also look at how the domain of the data and typological language similarity affects the quality of translation.

We conducted the following three experiments. Unsupervised MT with iterative BT using XLM model to demonstrate the feasibility of UMT for low-resource language pairs (Ro <-> En and Ne <-> En) and how data size will affect the translation quality; Zero-shot transfer learning with mBART model to illustrate how the size of data affects both fine-tuning scores for (Ro, Hi) -> En and transferring scores for (Cs, Hi) -> En and finally we finetune a pre-trained XLM model for (Fr, De and Ro) with training data from different domains and test on a completely different domain to observe the translation scores.
Iterative Back translation shows promising results when translating between similar language pairs and the performance scales with the available monolingual data(En<->Ro). However, on the dissimilar language pair (En<->Ne) no amount of monolingual data produced a translation of reasonable quality.

When performing Zero-Shot transfer with a transfer language (Hi) similar to the low resource source language (Ne) to a high resource target language (En) we observed that the translation quality improves with the amount of training data. However, when using a dissimilar transfer language (Ro) to translate from (Cs->En) we observe that the quality degrades as we increase the training data size beyond a certain point.

Our experiments on domain mismatch observed a consistent decrease of about 5-8 BLEU points for (Fr,De<->En). This reinforces the fact that for high resource languages the domain of the dataset is an essential factor to determine the overall translation quality, however translation is still feasible.We observe that for a low resource scenario (Ro<->En), the performance decrease is much more profound. The translation fails entirely with BLEU scores between 0-2.

We conclude that zero-shot transfer learning suffers from “curse of data size”, which means simply increasing the bi-text data size between the transfer and target language does not improve the quality of translation, especially when the transfer language is dissimilar to the source language. In this case, we suggest the use of Iterative BT over Zero-Shot language transfer given large enough monolingual training data is available. Whereas, when the source language and the target language are dissimilar, the presence of high quality bi-text data makes language transfer a more viable alternative.

[AACL SRW 2020] Building a Part-of-Speech Tagged Corpus for Drenjongke (Bhutia)

Mana Ashida, Seunghun J. Lee, Kunzang Namgyal

Summary by Mana Ashida and Seunghun J. Lee

Drenjongke (Bhutia) is a Tibeto-Burman language with the status of one of the official languages in Sikkim, India. The estimated number of Drenjongke speakers is around 40 000.This paper reports results of the creation of the first Drenjongke corpus based on texts taken from a phrasebook for beginners written in the Tibetan script. A corpus of sentences was created after correcting scanning errors in characters that appeared after the optical character recognition (OCR) process. A total of 34 Part-of-Speech (PoS) tags were defined based on manual annotation performed by the three authors, one of whom is a native speaker of Drenjongke. This first corpus of the Drenjongke language comprises 275 sentences and 1379 tokens, which we plan to expand with other materials to promote further studies of this language. The entire corpus is publicly available at ICULingLab/drenjongke.

NLP methods mostly target languages with writing systems that use a common set of Unicode characters and grammatical relations. In Drenjongke, this is not the case. Three observations are made based on issues that emerged due to a Drenjongke-specific writing system. Addressing the following points will make the current NLP methods more multilingually robust:

1. The presence of diacritics that are not recognized by any pre-existing OCR model for the Tibetan script,

2. The definition of a word or a morpheme: Drenjongke has productive monosyllabic morphemes that are not directly translatable into English, and
3. The creation of a set of Part-of-Speech tags for annotation: Drenjongke contains several modality markers that do not correspond to common set of Part-of-Speech tags

The Drenjongke corpus also contains the romanization of pronunciation of Drenjongke words and the English translation, which provides an effective tool in documenting the threatened language. Active collaboration between the native speaker, a linguist and an NLP researcher is another noteworthy part. The corpus is community-driven with colleagues in the NLP area.

Are Multilingual Models Effective in Code-Switching?

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, Pascale Fung

Summary by Genta Indra Winata

Multilingual language models have shown decent performance in multilingual and cross-lingual natural language understanding tasks. However, the power of these multilingual models in code-switching tasks has not been fully explored. Code-switching is a common phenomenon in which a person speaks more than one language in a conversation, and its usage is prevalent in multilingual communities.

In this paper, we study the effectiveness and efficiency of multilingual language models, such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), to understand their capability and adaptability to the mixed-language setting by considering the inference speed, performance, and the number of parameters to measure their practicality. We conduct experiments in three language pairs on named entity recognition and part-of-speech tagging and compare them with existing methods, such as using bilingual embeddings and multilingual meta-embeddings. Then, we further analyze the memory footprint required by each model over different sequence lengths in a GPU.

Our findings suggest that pre-trained multilingual models do not necessarily guarantee high-quality representations on code-switching. Interestingly, we found that XLM-R (large) achieves the best performance, but with a substantial cost in the inference time and 13x more parameters than the meta-embedding model for only around 2% average performance difference.

[HCI+NLP 2021 & AfricaNLP 2021] AfriKI: Machine-in-the-Loop Afrikaans Poetry Generation

Imke van Heerden and Anil Bas

Summary by Imke van Heerden

We propose a generative language model called AfriKI – Afrikaanse Kunsmatige Intelligensie (Afrikaans Artificial Intelligence). With the aim of promoting human creativity, we use the model as an authoring tool to explore machine-in-the-loop poetry generation. To our knowledge, this is the first study to attempt creative text generation in Afrikaans.

In comparison to resource-rich languages, NLP research in Afrikaans is limited. Instead of training on comprehensive poetry datasets or modeling poetic qualities (due to lack of resources), we train the network on a small corpus of contemporary fiction, i.e. on prose. Accordingly, the proposed model works as a text generator that produces individual lines rather than stanzas of verse. Experimenting with various networks, we obtain best results with a two-layer LSTM architecture.

Whereas NLG tends to prioritise fully automatic systems, we treat poetry generation as a co-creative system. Machine-in-the-loop frameworks emphasise human creativity through computational assistance, as opposed to human-in-the-loop pipelines. Furthermore, the study encourages human-centred design in low-resource languages, enabling the generation of high-quality poetic text with very limited data.

Our process consists of two stages. First, the model generates a set of unique individual lines. The generated phrases are distinct from the dataset, with hardly any repetition of word order. Second, the author responds by choosing phrases at will. To create the final artefact, the author arranges the selected lines vertically into short poems without modification. This collaborative writing system results in minimalist free verse poetry that is rich in imagery and figurative language, and draws inspiration from the literary movements Imagism and Surrealism.

A promising future direction to pursue would be the involvement of poets and writers to investigate whether this approach could inform and improve their creative writing practices.

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

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

SIGTYP Youtube Channel

SIGTYP now has a Youtube channel where we will be posting videos from our keynote speakers and other presenters who consent. This should allow us to reach a larger audience with our content! We will also post recordings from the SIGTYP speaker series (held outside of regular conferences), which is still in development.

For our community in China: we created a channel on Bilibili
https://space.bilibili.com/1055445444/channel/detail?cid=178350

FLORES101: A large evaluation benchmark for Multilingual Machine Translation in over 100 languages

Summary by Flores 101 team

Translation is a key technology to connect people and ideas together across language barriers. However, current translation technology works very well mostly in a few languages and it covers only few domains. Many people around the world still lack access, partly, due to the lack of compute and data resources to create translation models.

A prerequisite for developing new modeling techniques is having reliable evaluation. As a baby step in this direction back in 2019, we started FLORES, which came with two evaluation datasets for Nepali-English and Sinhala-English, that we later expanded to include Pashto and Khmer.

Today, we announce the FLORES101 evaluation benchmark: a full Many-to-Many evaluation dataset across over 100 languages, most of which are low-resource. True to the original multi-domain spirit of FLORES, this dataset consists of 3000 English sentences across several domains (news, books, and travel) all taken from Wikipedia, maintaining document-level context as well as document metadata, such as topics, hyperlinks, etc. These sentences are then professionally translated though several rounds of thorough evaluation.

We are making the entire dev and devtest splits of FLORES101 available to the research community (2000 sentences total, aligned Many to Many), in June 2021, along with a tech report describing the dataset in detail.

To ensure robust and fair evaluation, we'll keep the test split blind and not publicly accessible. Instead, we'll host an evaluation server based on open-source code, which will enable us to track the progress of the community in low-resource languages. Importantly, such a setup will enable comparison of models on several axes besides translation quality, such as compute and memory efficiency. The evaluation server will also be available starting June 2021.

Finally, we want to continue encouraging the research community to work on low-resource translation. As part of this, we are launching a WMT multilingual machine translation track and encourage people to apply for compute grants so that GPU compute is less of a barrier for translation research. You can see more detailed information and apply for the compute grant [here].

We propose two small tracks --- one low-resource European languages and another low-resource Southeast Asian languages --- along with the full track of 100+ languages.

Note that compute grants are available here.

NAACL 2021 Diversity & Inclusion subsidy

If you need funding for NAACL registration, we suggest you apply for the D&I funding. Colleagues from underrepresented communities are highly encouraged to apply. Please indicate in the form that you have an accepted SIGTYP paper and state the reasons that you need for funding.

First Languages Australia

First Languages Australia is working toward a future where Aboriginal language communities and Torres Strait Islander language communities have full command of their languages and can use them as much as they wish to. We recommend you to check out their resources section which consists of various toolkits, articles and corpora on Indigenous languages.

Queer Reads Lexicon

Queer Reads Lexicon is a project started by Queer Reads Library in 2019 to explore queerness in the context of local languages, primarily in Cantonese.

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.

Speaker-centered NLP with Finer-Grained Multilingual Model Proficiency

Yulia Tsvetkov at SIGTYP 2020

Building Resources: Language Comparison and Analysis

Miriam Butt at SIGTYP 2020

Taxonomy of Writing Systems

Richard Sproat at SIGTYP 2020

Crosslingual Syntactic and Semantic Annotation:Typological and Ethical Precepts

William Croft at SIGTYP 2020

Wikitongues

Vocal Fries Podcast

Carrie and Megan talk with Daniel Bogre Udell, co-founder of Wikitongues, about language documentation and revitalization.

Non Binary Linguistic Activism

Ártemis López at Queer in AI 2020

Their talk is about non-binary expression in Spanish, which has binary grammatical gender, and how this binary can be challenged beyond non-binary expression to engage in radical linguistic activism.