SIGTYP 2020 Shared Task: Prediction of Typological Features

Bjerva, Johannes; Salesky, Elizabeth; Mielke, Sabrina J.; Chaudhary, Aditi; Giuseppe, Celano; Ponti, Edoardo Maria; Vylomova, Ekaterina; Cotterell, Ryan; Augenstein, Isabelle

Published in:
Proceedings of the Second Workshop on Computational Research in Linguistic Typology

DOI:
10.18653/v1/2020.sigtyp-1.1

Publication date:
2020

Document version
Publisher's PDF, also known as Version of record

SIGTYP 2020 Shared Task: Prediction of Typological Features

Johannes Bjerva[•,◁] **Elizabeth Salesky**[◇] **Sabrina J. Mielke**[◇] **Aditi Chaudhary**[□]
Giuseppe G. A. Celano[⊙] **Edoardo M. Ponti**[‡] **Ekaterina Vylomova**[♯]
Ryan Cotterell[⋆] **Isabelle Augenstein**[◁]
[•]Aalborg University [◁]University of Copenhagen [◇]Johns Hopkins University
[♯]University of Melbourne [□]Carnegie Mellon University [⊙]Leipzig University
[⋆]ETH Zürich [‡]University of Cambridge
jbjerva@cs.aau.dk, augenstein@di.ku.dk

Abstract

Typological knowledge bases (KBs) such as WALS (Dryer and Haspelmath, 2013) contain information about linguistic properties of the world's languages. They have been shown to be useful for downstream applications, including cross-lingual transfer learning and linguistic probing. A major drawback hampering broader adoption of typological KBs is that they are sparsely populated, in the sense that most languages only have annotations for some features, and skewed, in that few features have wide coverage. As typological features often correlate with one another, it is possible to predict them and thus automatically populate typological KBs, which is also the focus of this shared task. Overall, the task attracted 8 submissions from 5 teams, out of which the most successful methods make use of such feature correlations. However, our error analysis reveals that even the strongest submitted systems struggle with predicting feature values for languages where few features are known.

1 Introduction

Linguistic typology is the study of structural properties of languages (Comrie, 1988; Croft, 2002; Velupillai, 2012). Approaches to the categorisation of the languages of the world according to their linguistic properties are represented by, e.g., typological features in databases such as WALS (Dryer and Haspelmath, 2013), URIEL (Littell et al., 2017), and AUTOTYP (Nichols et al., 2013), e.g. in terms of their syntax, morphology, and phonology. One example of such a typological feature is the basic word order feature in WALS. For instance, English is best described as a subject-verb-object (SVO) language, whereas Japanese is best described as a subject-object-verb (SOV) language.

Once a relatively niche topic in the NLP community, studying typological features has recently risen in popularity and importance for a number of reasons. The field has seen considerable advances in cross-lingual transfer learning, whereby stable cross-lingual representations can be learned on massive amounts of data in an unsupervised way, be it for words (Ammar et al., 2016; Wada et al., 2019) or, more recently, sentences (Artetxe and Schwenk, 2019; Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020). This naturally raises the question of what these representations encode, and some have turned to typology for potential answers (Choenni and Shutova, 2020; Zhao et al., 2020). In a similar vein, research has shown that these learned representations can be fine-tuned for supervised tasks, then applied to new languages in a few- or even zero-shot fashion with surprisingly high performance. This has raised the question of what causes such surprisingly high results, and to what degree typological similarities are exploited by such models (Bjerva and Augenstein, 2018a; Nooralahzadeh et al., 2020; Zhao et al., 2020).

In addition to using typology for diagnostic purposes, prior work has also found that typology can guide cross-lingual sharing (de Lhoneux et al., 2018). Finally, the relationship between typological knowledge bases (KBs) such as WALS (Dryer and Haspelmath, 2013) and language representations has been studied, which has shown that knowledge base population methods can be used to complete typological KBs (Malaviya et al., 2017; Murawaki, 2017; Bjerva and Augenstein, 2018a; Bjerva et al., 2019c), and that implications can be discovered in typological KBs (Daumé III and Campbell, 2007; Bjerva et al., 2019b).

The latter stream of work has provided the inspiration for this shared task on typological feature prediction. As knowledge bases are notoriously incomplete and require manual labour from (in this case, linguistic) domain experts to create, populate and maintain, high-performance methods for auto-

Proceedings of the Second Workshop on Computational Research in Linguistic Typology, pages 1–11

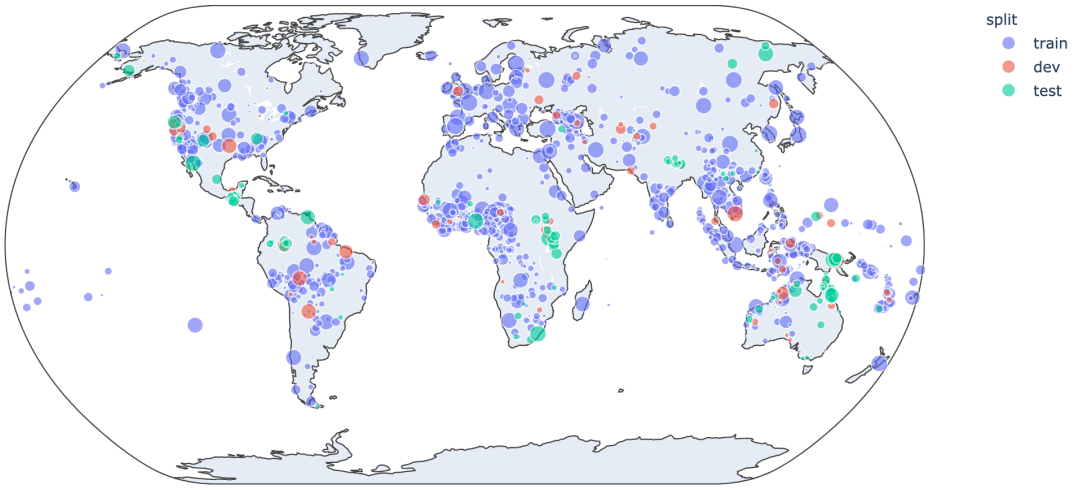

Figure 1: Shared task WALS data superimposed on a map showing one point per language with train, dev, and test splits; relative point sizes representing number of features for that language.

matic knowledge base population are highly desirable. While past approaches have shown the feasibility of typological feature prediction, the considered evaluation setups have some flaws which led to overestimated performance. Some papers control for phylogenetic relationships between languages, e.g. not both training and testing on Slavic languages, but little-to-no work has considered controlling for geographical proximity. This is corrected for in this shared task.

The shared task attracted 8 system submissions from 5 teams for two subtasks (constrained and unconstrained resources). In general, the systems which make use of correlations between features, and exploit observed features during inference, perform better, whereas those that do not make use of observed features perform similarly to our baselines.

2 Task Description

The SIGTYP 2020 shared task is concerned with predicting typological features from the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013). For the task, participants were invited to build systems to predict features for languages unseen at training time. The shared task consisted of two subtasks: 1) the *constrained* setting, for which only the provided training data may be used; 2) the *unconstrained* setting, for which training data may be extended with any external source of information (e.g. pre-trained embeddings, additional text, etc.)

Data Format For each instance, the following information is provided: the language code, name, latitude, longitude, genus, family, country code, and features. At training time, both the feature names and feature values are given, while at test time, submitted systems are required to fill values for the requested features. An example of a test instance is given in Table 1.

2.1 Dataset

WALS comprises 2679 languages and a total of 192 feature categories (Dryer and Haspelmath, 2013). However, the database is quite sparse in that many language-feature combinations lack annotation. Furthermore, it is a skewed database, in that a handful of languages have annotations for a large number of features, and some features are annotated for almost all languages, whereas some have very little coverage. In order to alleviate data sparsity in the shared task, only the subset of the languages in WALS with more than 3 features available are considered. Furthermore, of all the features of the languages so selected only those present in more than 9 languages have been retained. Most feature categories in WALS can take several feature values. For instance, the feature `Tone` can take one of the values: `No tones`, `Simple tone system`, or `Complex tone system`. This dataset has been divided into train set (90%), dev set (5%), and test set (5%).

	Lang code	Name	Lat	Long	Genus	Family	Count Code	Features
Input	mhi	Marathi	19.0	76.0	Indic	Indo-European	IN	order_of_subject,_object,_and_verb=? | number_of_genders=?
	jpn	Japanese	37.0	140.0	Japanese	Japanese	JP	case_syncretism=? | order_of_adjective_and_noun=?
Output	mhi	Marathi	19.0	76.0	Indic	Indo-European	IN	order_of_subject,_object,_and_verb=SOV | number_of_genders=three
	jpn	Japanese	37.0	140.0	Japanese	Japanese	JP	case_syncretism=no_case_marking | order_of_adjective_and_noun=demonstrative-Noun

Table 1: Data format for two test instances of the SIGTYP 2020 shared task dataset

3 Evaluation Setup

While a substantial amount of previous work deals with feature prediction in typological databases such as WALS (e.g. Malaviya et al. (2017); Murawaki (2017); Bjerva and Augenstein (2018a); Bjerva et al. (2019c)), most such work does not take into account that both phylogenetic and geographic proximity should be controlled for. Languages which have shared common ancestry will often have similar typological features, hence training and evaluating on the same language family will tend to inflate the expected performance of the model (Bjerva et al., 2019a). In the data for this shared task, we make sure to control for both of these factors.

Our evaluation setup is constructed as follows. We evaluate on a set of languages from small languages spread across the world, as defined by the WALS macroareas: Mayan (North America), Tucanoan (South America), Madang (Papuanesia), Mahakiranti (Eurasia), Northern Pama-Nyungan (Australia), and Nilotic (Africa). In addition, we include a subset of languages spoken around the world, by randomly sampling 10% of the available data in WALS. This yields two evaluation set-ups: one in which we evaluate on unobserved languages, controlling for both phylogenetic and geographic relationships, and one in which we perform a random evaluation as is common in previous work.

The languages in the test data vary in the number of removed and present feature values so that the *blanking ratios* are spread uniformly between 5% and 95%. This will allow our analysis to investigate whether some approaches benefit from observing a large number of features and whether some are robust to situations where only a small number of features are observed (subsection 5.4).

In order to control for phylogenetic and geographic effects, we remove all languages from the same language genus as the aforementioned languages from the training set, as well as all languages which are spoken within 1,000km of any of these languages.[1] This reduces the number of languages in the training set to 1250. The task had participants run their systems on the partial feature information for our held-out languages and send us the outputs of their systems, i.e., the imputed features.

3.1 Evaluation Metrics

We report macro-averaged accuracies, meaning that we first compute the average accuracy for each language, i.e., the fraction of to be imputed features correctly predicted by the participant's system, then average these language accuracies within each language genus, and finally report the average of these genus accuracies to rank participants as well as all these accuracies for each language genus (Section 5) to see whether systems behave differently on different language families. We judge statistical significance using a non-parametric two-tailed paired permutation test with 5k samples each.

3.2 Baselines

We provide two baselines. The first is a simple lower-bound baseline based on observing feature frequencies in WALS (Baseline_frequency in Figure 2). For each unobserved feature in the test set, we predict the most frequent feature value from the training set.

The second uses the k-nearest neighbours (k-NN) algorithm with a simple feature set to predict each unobserved feature, with $k = 1$ (Baseline_knn-imputation in Figure 2). Each language is represented by a language vector ($\vec{l} \in \mathbb{R}^{64}$) trained as a part of a multilingual character-based language model (Östling and Tiedemann, 2017). During inference, for a language l and unobserved feature y, we find the nearest neighbour to $\vec{l}$ for which y has been observed, similar to Bjerva and Augenstein (2018a,b).

[1] Distances calculated with WALS language locations.

3.3 Submissions

We received eight submissions from five teams across the *constrained* and *unconstrained* subtasks, as described below.

ÚFAL (Vastl et al. (2020), Charles University) submitted a *constrained* system which ensembled two approaches: first, estimating the correlation of feature values within languages enables missing feature prediction, and second, using a neural network to predict whether feature values match a specific language after training one network with all provided WALS feature values and pre-computed language embeddings. By ensembling both using confidence scores, they were able to improve on each individual approach and produce the best accuracy of all constrained and unconstrained submissions.

CrossLingference (Jäger (2020), University of Tübingen) submitted an *unconstrained* system using inferred phylogenetic trees. These were built with Continuous Time Markov Processes using Swadesh lists from the Automated Similarity Judgment Project (ASJP), with k-nearest neighbour estimations based on geographic information as back-off for test set languages not in both Glottolog and ASJP. Ancestral state reconstruction allows the inference of features for ancestral states from the provided surface features (WALS), and similarly, for this year's shared task, unknown feature values for non-ancestral languages can be inferred individually by rerooting the tree to a related language.

NUIG (Choudhary (2020), NUI Galway) submitted a *constrained* system with independent classifiers to predict each WALS feature. The outputs of independent classifiers are then fed into a shared encoder with feed-forward and self-attention layers in order to make use of feature correlations. Their model does not use other known features for WALS feature prediction at inference time, relying only on the 5-dimensional inputs of longitude, latitude, genus, family, and country-code.

NEMO (Gutkin and Sproat (2020), Google London and Tokyo) submitted *constrained* systems which first computed probabilities of represented feature values across each language's genetic (genus and family), and areal (features from languages within a 2,500 kilometre radius, computed from provided latitude and longitude with the Haversine formula), and *implicational universals* or rather, priors for certain features given commonly associated feature-value pairs in the data.

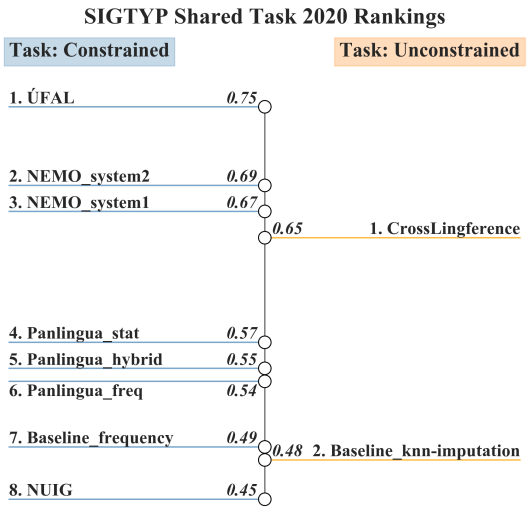

Figure 2: Macro-averaged rankings of all submissions

They compared several classifiers' performance using these sparse features, ultimately submitting systems using ridge regression. The two submitted systems differ in whether these features were computed for the test set or only train and dev.

Panlingua (Kumar et al., 2020), a team effort across KMI, Panlingua, and IIT KGP, submitted *constrained* systems from three approaches: two rule-based systems (one statistical, and one frequency-based baseline) and one hybrid system. Their baseline is similar to the organizers' frequency-base baseline, except that it produces the most frequent value for a feature within a genus if available, backing off to language family, and then the overall most-frequent value. The hybrid system uses 180 different SVM classifiers for the 180 features which were present in the training set. The statistical system provides a two-step back off procedure if neither a feature has been seen for either a languages' genus or family in training: first, finding the most frequent values in nearby languages using Haversine distance, and if these are too distant, turning to nearby language families. This system performed best on the held-out data.

4 Results

Figure 2 shows the overall results and rankings for all shared task submissions. The rankings use macro-averaged accuracies as this equally weights the controlled genera (the exception is the comparison to micro-averaged accuracies in Figure 3). This year's shared task was separated into two subtasks: *constrained* systems which used only the WALS

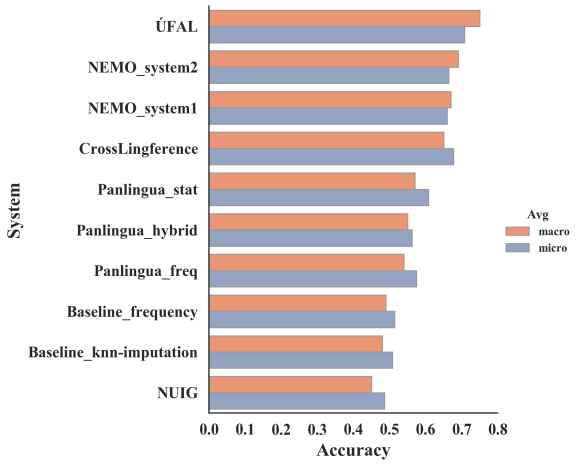

Figure 3: Comparison of macro-averaged and micro-averaged accuracies across submissions

features and data provided, and *unconstrained* systems, open to use of any data or pre-trained models. Accordingly, we have two winning systems: **ÚFAL** for *constrained*, and **CrossLingference** for *unconstrained*, with **ÚFAL** producing the best results overall across both subtasks.

Results for each unobserved genus, shown in comparison to results across genera observed in training, may be found in Table 2.

WALS feature value formatting is not standardized and, unfortunately, the test data was released containing additional tabs within the feature values for some features, which adversely affected teams who may have used tab-separation for data preprocessing. Many teams accounted for this and submitted feature values for all 2417 features across the 149 languages in the test data, but for two teams this led to missing features in their submissions: **CrossLingference** was missing 7 features across 7 languages, affecting their results by 1%; **Panlingua** was missing 61 features across 15 languages in their rule-based submission and 57 across 11 for their additional two submissions, affecting their results by 2%. When evaluating without the affected features, rankings were not changed, nor were there significant differences between submitted systems.

4.1 Subtask 1: Constrained Setting

The nine systems in the constrained setting used a diverse set of model features and architectures. When computing pairwise significances with a paired permutation test, we find that these systems cluster into three groups, within each the systems are not significantly different from each other: {**1**}, {**2-3**}, and {**4-8**}. Teams submitting multiple systems were able to improve their accuracy within their own submissions, but we did not find that their individual submissions were statistically significantly different. Similar differences in overall accuracy do not necessarily indicate statistically significant margins: for example, the 1st and 2nd systems have the same margin (0.05) as the 4th and 7th, but the latter are not significantly different while the former are.

Finer-grained analysis of results across controlled genera, and comparing results across different levels of representation in the training data, can be found in Section 5.

4.2 Subtask 2: Unconstrained Setting

CrossLingference submitted the only unconstrained system, which used additional data in the form of Swadesh lists to infer phylogenetic trees. This system outperforms the unconstrained knn-imputation baseline on all evaluated conditions. When we contextualize this submission by comparing it to those in the constrained setting, we find that it joins the second cluster with the two submissions from **NEMO**; interestingly, when features are micro-averaged rather than macro-averaged, these teams reorder, with **CrossLingference** outperforming the two **NEMO** systems, seen in Figure 3. This is somewhat counter-intuitive, given the way each system uses phylogenetic information. While **CrossLingference** explicitly models phylogenetic information through its model structure, **NEMO** takes a frequentist approach where the counts and probabilities of each feature within a language's genus, family, and geographic area are pre-computed and passed as sparse features to feature classifiers. One might expect the latter to perform better on a micro-average where overall data frequencies would be more heavily weighted than each genus, but this was not the case here. We explore this further in Section 5.

5 Analysis

Our test data was constructed to enable comparison across controlled phylogenetic and geographic relationships, and randomly sampled features from languages covered in training as is common in previous work.

Submission	Tucanoan (8)	Madang (9)	Mahakiranti (13)	Nilotic (15)	Mayan (17)	N. Pama-Nyungan (24)	Other genera (63)
ÚFAL	**0.73**	**0.78**	**0.74**	0.71	**0.80**	**0.76**	**0.76**
NEMO_system2	0.71	0.72	0.72	**0.76**	0.76	0.67	0.69
NEMO_system1	0.70	0.72	0.68	0.75	0.71	0.68	0.67
Panlingua_stat	0.70	0.64	0.55	0.55	0.33	0.62	0.58
Panlingua_hybrid	0.65	0.64	0.57	0.51	0.34	0.61	0.53
Panlingua_freq	0.59	0.64	0.53	0.55	0.31	0.59	0.55
Baseline_frequency	0.51	0.53	0.37	0.49	0.41	0.58	0.53
NUIG	0.51	0.56	0.35	0.45	0.32	0.45	0.48
CrossLingference	0.71	0.73	0.67	0.68	0.57	0.60	0.65
Baseline_knn-imputation	0.48	0.57	0.46	0.48	0.32	0.52	0.51

Table 2: Macro-averaged results across each unobserved genus, as compared to genera with languages observed in training with randomly sampled splits, shown with number of languages in each genus.

5.1 Overall Results

Table 2 compares submission accuracy on features from diverse WALS macroareas unobserved in training data, and other observed genera. We see that overall rankings hold when evaluated on observed languages. However, this is not the case for several of our unobserved genera. With respect to the shared task baselines, we find that the frequency baseline, which naively picks the most well-represented values for each feature, is most representative for the larger 'other genera' category, which represents the majority of the training data but does not account for the diversity of typological features and values across many languages. Nonetheless, for most of the unobserved genera, the frequency baseline performed better than the knn-imputation baseline, which was significantly better for Mahakiranti only, primarily due to correct prediction of "OV" ordering across multiple features.

Interestingly, while the first 3 systems perform better on macro-averaged accuracy than micro-averaged (Figure 3), this is not true for all other systems, suggesting that they rely more on getting frequent and "easy" features right, relying on frequency in training data. Note that the six unobserved genera come from separate macroareas across six different continents, and have a more even distribution of feature values than the 'other genera.'

5.2 Differences across Genera

Looking at specific genera, we see that Mayan caused the greatest split between submitted systems, with the first two clusters performing very well, and the frequency baseline performed better than the majority of systems. On the other end of the spectrum, certain genera (Tucanoan and Madang) with well-represented features were relative equalizers, with the least variance in results across the submitted systems.

Within those teams which submitted multiple systems, there were only certain cases in which these performed significantly differently from each other. **Panlingua** submitted three different systems; two rule-based (one statistical and one frequency-based), and one hybrid model. For most genera, these performed very similarly, with consistently better results from the statistical rule-based system than the others, though there were no statistically significant differences shown by paired permutation tests. However, this was not the case for Tucanoan, where the statistical (and to a less degree, hybrid) model significantly outperformed the other. These systems had equal performance on four of the Tucanoan languages {Cubeo,Secoya,Siona,Koreguaje}, but quite divergent on the remaining four languages {Desano,Retuarã,Tucano,Tuyuca}. This second set required predicting values for several features concerning the order of Subject, Object, Verb, which the statistical model was able to correctly predict through better back-off choices, but swayed by the more frequent SVO languages in training, their frequency baseline and SVM-based classifiers were not.

5.3 Differences among Features

Table 3 shows the features with the highest and lowest accuracies across all submissions. We find that the features with highest accuracies also have the most consistent performance across all systems, and typically have the most frequent values for

	Feature	# Langs	Avg. Accuracy	Std. Deviation
Highest	front_rounded_vowels	4	0.65	0.08
	inclusive/exclusive_forms_in_pama-nyungan	2	0.65	0.09
	distributive_numerals	3	0.64	0.08
	optional_double_negation_in_svo_languages	1	0.64	0.08
	voicing_in_plosives_and_fricatives	4	0.63	0.08
Lowest	verb-initial_with_clause-final_negative	1	0.44	0.21
	multiple_negative_constructions_in_svo_languages	3	0.41	0.27
	suppletion_in_imperatives_and_hortatives	2	0.40	0.13
	languages_with_two_dominant_orders_of_subject,_object,_and_verb	2	0.39	0.20
	the_position_of_negative_morphemes_in_verb-initial_languages	9	0.38	0.28

Table 3: Features with the highest and lowest overall accuracies across *all* submissions, with number of languages containing the feature in the test data (183 total languages)

those features. The most difficult features, on the other hand, have the least frequently occurring values in the training data, and have higher variance – interestingly, the top four systems were nonetheless able to achieve greater than 65% accuracy on these features, while the remaining systems' accuracies were $\sim 20\%$.

5.4 Impact of Blanking Ratio

Since the languages in our test sets all remove and retain different numbers of features, we can see whether the blanking ratio, that is, the ratio of features that participants have to impute to the features listed for that language, correlates with the performance of the system in question on that language. Calculating Pearson's correlation coefficient for every system individually, we realize that they range from -0.23 (for a NEMO system) to 0.31 (for a Panlingua system), almost all of which statistically significantly different from 0.

Why do these correlations differ so much from system to system? To answer that question, we plot these correlation coefficients as a function of the system's overall performance in Figure 4. It turns out that a system's overall performance and how sensitive it is to the blanking ratio are highly correlated: the stronger systems are much more negatively affected by the removal of more features (their correlations are negative), weaker systems are not only not harmed, but seem to find the languages where only a few features are blanked harder still (having positive correlations).

6 Related Work

Previous work can be divided into research on predicting typological features automatically, cross-lingual transfer learning which utilises typology

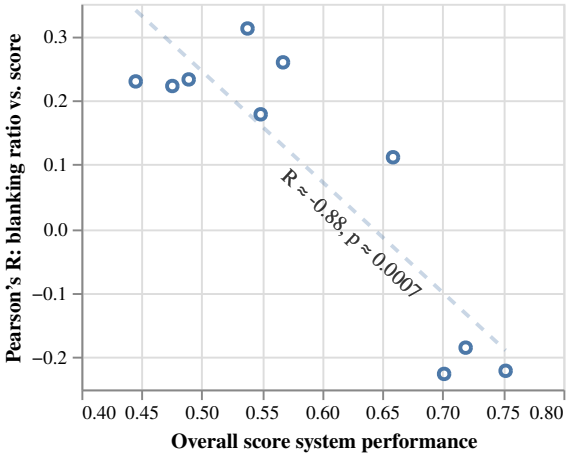

Figure 4: Correlation coefficients (between blanking ratio and language performance) for each system as a function of that system's performance. The correlation of $\mathrm{R} = -0.88$ is significant at $p < .001$.

to inform sharing, probing of representations for what typological knowledge they encode, and finally, work on how best to represent a language in terms of its typological features.

6.1 Predicting Typological Features

Typological knowledge bases are both sparse and skewed in terms of language–feature annotations. They are sparse in the sense that most languages only have annotations for a handful of features and skewed in the sense that a few features have much wider coverage than others. Luckily, such features often correlate with one another, which allows for prediction of those features from others. For instance, languages where the verb precedes the object tend to have prepositions, e.g. Norwegian, whereas languages where the object precedes the verb word tend to have postpositions, e.g. Japanese.

A survey of approaches to prediction of features

is provided in Ponti et al. (2019a, § 4.3). Some common approaches include prediction based on language representations learned as a by-product of model training (Östling and Tiedemann, 2017; Malaviya et al., 2017; Bjerva and Augenstein, 2018a; Bjerva et al., 2019c) and matrix factorisation (Murawaki, 2017; Bjerva et al., 2019a).

6.2 Typologically Informed Sharing

Cross-lingual sharing informed by typology has been investigated for, among others, parsing (Naseem et al., 2012; Täckström et al., 2013; Zhang and Barzilay, 2015; de Lhoneux et al., 2018), language modeling (Tsvetkov et al., 2016; Ponti et al., 2019b), machine translation (Daiber et al., 2016; Ponti et al., 2018), and morphological inflection (Chaudhary et al., 2019). Many of these approaches use language embeddings with sparse features encoding WALS feature values. Oncevay et al. (2020) find that combining information from typological databases with embeddings learned during training of an NMT model can be beneficial for multilingual NMT.

6.3 Typological Probing

Several recent papers study typological feature prediction as a probing task for evaluating cross-lingual sentence encoders (Choenni and Shutova, 2020; Bjerva and Augenstein, 2018a; Nooralahzadeh et al., 2020; Zhao et al., 2020). Typically, hidden representations are probed for whether or not they might encode a typological feature by, e.g., using them in a separate classifier (Malaviya et al., 2017; Bjerva and Augenstein, 2018a; Nooralahzadeh et al., 2020). Östling and Tiedemann (2017) learn language representations during multilingual language modelling and find that the resulting representations can reproduce relatively credible phylogenetic trees.

Bjerva and Augenstein (2018a) learn language representations under NLP tasks such as POS tagging and grapheme-to-phoneme conversion, and find that typological features related to the task at hand are sometimes encoded. Nooralahzadeh et al. (2020) use a typological probing task in experiments for zero- and few-shot NLI and QA, finding that languages which share typological properties benefit from sharing. Zhao et al. (2020) attempt to induce language-agnostic representations, e.g. by reducing the typological gaps between languages, and find that this is beneficial for NLI and MT. Gerz et al. (2018) show that there is a correlation between typological features related to morphology and model performance in language modelling, and Cotterell et al. (2018) further show that inflectional morphology affects performance in both n-gram and LSTM-based language models.

7 Conclusions

This paper documents the first SIGTYP shared task on prediction of typological features in WALS. The 8 system submissions from 5 teams showed that a variety of different methods can be applied to the task. Interestingly, the best system only achieved a macro-averaged accuracy of 75%, indicating that the task is far from solved. This further shows that the evaluation set-up in which we controlled for both phylogenetic relationships and geographic proximity is a challenging one. We expect that further exploration of unconstrained systems to have the most potential for predicting features in such cases, where little or nothing is known about a language.

Acknowledgments

This research has received funding from the Swedish Research Council under grant agreement No 2019-04129, as well as the German Research Foundation (DFG project number 408121292).

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively Multilingual Word Embeddings. *CoRR*, abs/1602.01925.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Trans. Assoc. Comput. Linguistics*, 7:597–610.

Johannes Bjerva and Isabelle Augenstein. 2018a. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916, New Orleans, Louisiana. Association for Computational Linguistics.

Johannes Bjerva and Isabelle Augenstein. 2018b. Tracking Typological Traits of Uralic Languages in Distributed Language Representations. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 76–86, Helsinki, Finland. Association for Computational Linguistics.

Johannes Bjerva, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019a. A probabilistic generative model of linguistic typology. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1529–1540, Minneapolis, Minnesota. Association for Computational Linguistics.

Johannes Bjerva, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019b. Uncovering probabilistic implications in typological knowledge bases. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3924–3930, Florence, Italy. Association for Computational Linguistics.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019c. What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

Aditi Chaudhary, Elizabeth Salesky, Gayatri Bhat, David R. Mortensen, Jaime Carbonell, and Yulia Tsvetkov. 2019. CMU-01 at the SIGMORPHON 2019 shared task on crosslinguality and context in morphology. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 57–70, Florence, Italy. Association for Computational Linguistics.

Rochelle Choenni and Ekaterina Shutova. 2020. What does it mean to be language-agnostic? probing multilingual sentence encoders for typological properties. *CoRR*, abs/2009.12862.

Chinmay Choudhary. 2020. NUIG: Multitasking Self-attention based approach to SigTyp 2020 Shared Task. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*. Association for Computational Linguistics.

Bernard Comrie. 1988. Linguistic typology. *Annual Review of Anthropology*, 17:145–159.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *ACL*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *NeurIPS*, pages 7057–7067.

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.

William Croft. 2002. *Typology and Universals*. Cambridge University Press.

Joachim Daiber, Miloš Stanojević, and Khalil Sima'an. 2016. Universal reordering via linguistic typology. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3167–3176, Osaka, Japan. The COLING 2016 Organizing Committee.

Hal Daumé III and Lyle Campbell. 2007. A Bayesian Model for Discovering Typological Implications. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 65–72, Prague, Czech Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, Brussels, Belgium. Association for Computational Linguistics.

Alexander Gutkin and Richard Sproat. 2020. NEMO: Frequentist Inference Approach to Constrained Linguistic Typology Feature Prediction in SIGTYP 2020 Shared Task. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*. Association for Computational Linguistics.

Gerhard Jäger. 2020. Imputing typological values via phylogenetic inference. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*. Association for Computational Linguistics.

Ritesh Kumar, Deepak Alok, Akanksha Bansal, Bornini Lahiri, and Atul Kr. Ojha. 2020. KMI-Panlingua-IITKGP at SIGTYP2020: Exploring rules and hybrid systems for automatic prediction of typological features. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*. Association for Computational Linguistics.

Miryam de Lhoneux, Johannes Bjerva, Isabelle Augenstein, and Anders Søgaard. 2018. Parameter sharing between dependency parsers for related languages. In *Proceedings of the 2018 Conference on*

Empirical Methods in Natural Language Processing, pages 4992–4997, Brussels, Belgium. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14. Association for Computational Linguistics.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics.

Yugo Murawaki. 2017. Diachrony-aware induction of binary latent representations from typological features. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 451–461, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–637, Jeju Island, Korea. Association for Computational Linguistics.

Johanna Nichols, Alena Witzlack-Makarevich, and Balthasar Bickel. 2013. The autotyp genealogy and geography database: 2013 release. *Zurich: University of Zurich*.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-Shot Cross-Lingual Transfer with Meta Learning. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Arturo Oncevay, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. In *Proceedings of EMNLP*. Association for Computational Linguistics. ArXiv preprint arXiv:2004.14923.

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain. Association for Computational Linguistics.

Edoardo Maria Ponti, Helen O'horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019a. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Edoardo Maria Ponti, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018. Isomorphic transfer of syntactic structures in cross-lingual NLP. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1531–1542, Melbourne, Australia. Association for Computational Linguistics.

Edoardo Maria Ponti, Ivan Vulić, Ryan Cotterell, Roi Reichart, and Anna Korhonen. 2019b. Towards zero-shot language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2900–2910, Hong Kong, China. Association for Computational Linguistics.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1061–1071, Atlanta, Georgia. Association for Computational Linguistics.

Yulia Tsvetkov, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W Black, Lori Levin, and Chris Dyer. 2016. Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1357–1366, San Diego, California. Association for Computational Linguistics.

Martin Vastl, Daniel Zeman, and Rudolf Rosa. 2020. Predicting Typological Features in WALS using Language Embeddings and Conditional Probabilities: ÚFAL Submission to the SIGTYP 2020 Shared Task. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*. Association for Computational Linguistics.

Viveka Velupillai. 2012. *An introduction to linguistic typology*. John Benjamins Publishing Company Amsterdam, Philadelphia.

Takashi Wada, Tomoharu Iwata, and Yuji Matsumoto. 2019. Unsupervised Multilingual Word Embedding with Limited Resources using Neural Language Models. In *ACL (1)*, pages 3113–3124. Association for Computational Linguistics.

Yuan Zhang and Regina Barzilay. 2015. Hierarchical low-rank tensors for multilingual transfer parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1857–1867, Lisbon, Portugal. Association for Computational Linguistics.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2020. Inducing Language-Agnostic Multilingual Representations. *arXiv preprint arXiv:2008.09112*.