Syntactic Typology from Plain Text Using Language Embeddings

Taiqi He
University of California, Davis
tqhe@ucdavis.edu

Kenji Sagae
University of California, Davis
sagae@ucdavis.edu

1 Introduction

We examine the question of whether information about linguistic typology can be derived automatically solely from text corpora, without access to any kind of annotation or parallel data. We describe an ongoing effort to use text from various languages to develop an unsupervised approach to characterize languages through *language embeddings*, which encode information about the structure of languages as vectors. We then explore whether these language vector representations encode typological information, which would traditionally require human expertise.

In recent work, Wang and Eisner (2017) showed that it is possible to predict word order for various languages using models based only on part-of-speech tag sequences, showing that syntactic typology can be modeled to an extent from sequences, without the need for full structural annotations. In contrast, we do not leverage any kind of linguistic annotation, relying instead on multilingual word embeddings, which cab be derived from plain text (Lample et al., 2017).

2 Method

Our approach is based on the idea behind a denoising autoencoder (Vincent et al., 2008) applied to many languages simultaneously. Given a text corpus with unrelated sentences in each language, we use an encoder-decoder model that learns to reorder, or denoise, sentences in each language.

We first map the words from the various languages into a common representation, leveraging the multilingual 300-dimensional word embeddings from Facebook project MUSE (Lample et al., 2017). We replaced each word in each sentence, regardless of language, with the nearest English word in the multilingual word embedding space. Although it would also be possible to use the multilingual word embeddings directly, as an approximation that allowed for more convenient experimentation, we used English words as a pivot.

We then have a corpus with sentences consisting of English words in the original orders from the different languages. The words in each sentence of this corpus are reordered randomly, creating the input, or source, sequences. The target sequences are the corresponding original sentences. The model then must learn to reorder words in each language, from a random order, to the original order. Additionally, we provide the model with information about what language the sentence is from by appending to each word on both the source and target sides a feature that corresponds to the language identity. Table 1 shows how our target sentences are represented, with English words, original word orders, and language features, along with the original sentence for comparison. A 50-dimensional embedding of this language identity feature is learned along with the reordering task. The intuition is that the model will learn that reordering the same words is done differently depending on whether the language is English, French, Turkish, Vietnamese, etc., but that certain languages are more similar to, or more different from, each other. Once the model is trained, the language feature embedding that helps the model learn how to reorder words for a specific language is the language embedding.

After training, we retrieved the language embeddings from the decoder and encoder of the model. Our initial analysis indicates that the language embeddings on the encoder and the decoder learn similar language relationships, and we discuss below only the results from the decoder embeddings, which appear to be of higher quality.

Our BiLSTM encoder and LSTM decoder both have two layers of 500 units. We used 29 lan-

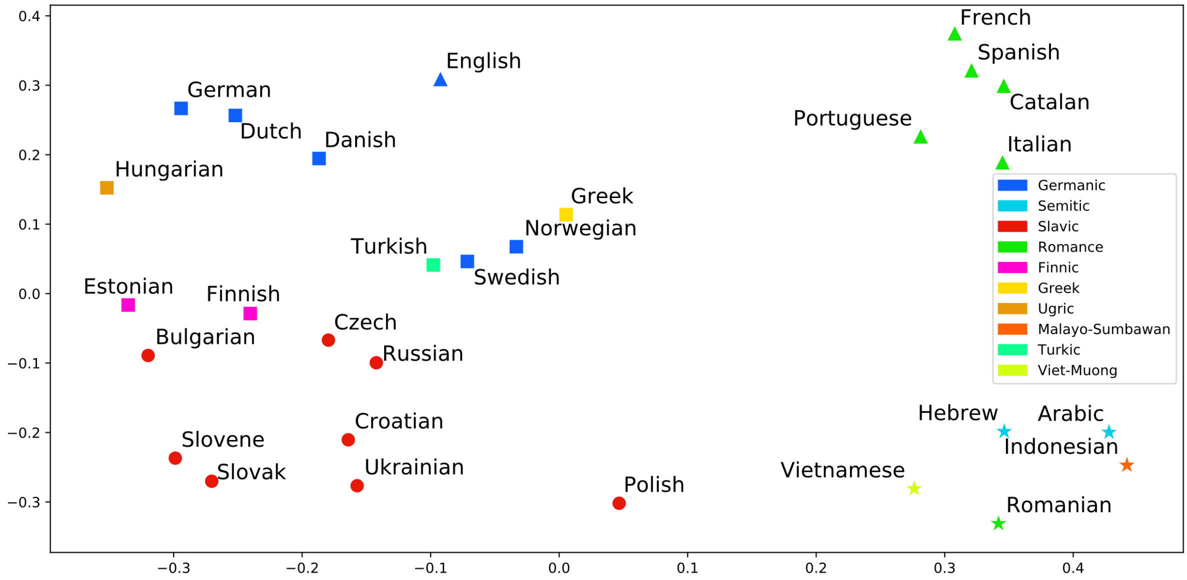

Figure 1: PCA projection of the language embeddings. Shapes represent automatically derived clusters. The Rand score between actual (color) and predicted (shape) language categorizations is .55.

Original	**Transformed**
Er hat den roten Hund nicht gesehen	he|de has|de the|de red|de dog|de not|de seen|de
No vio al perro rojo	not|es saw|es the|es dog|es red|es
Il n'a pas vu le chien rouge	he|fr not|fr has|fr seen|fr the|fr dog|fr red|fr

Table 1: The sentence *He didn't see the red dog* transformed from German, Spanish, and French to English. Word orders were preserved and a label denoting the origin language was attached to each word.

guages, with sentence counts varing from 200 thousand to 1.9 million, although for most languages there were approximately 1 million sentences. Although the MUSE embeddings used in our experiments were created using bilingual dictionaries, violating our goal of deriving the language representations from text only, we also plan to examine the use of multilingual embeddings obtained from text alone (Chen and Cardie, 2018).

3 Examining Language Embeddings

Figure 1 shows the two-dimensional PCA projection of the normalized language embeddings. We can clearly see clustering of Slavic languages on the lower left, Romance on the upper right, and Germanic on the upper left. Our dataset also had two Finnic languages, which appear right above the Slavic languages, and two Semitic languages, which appear on the lower right. The other languages are from families underrepresented in our dataset, and appear either mixed with the Germanic languages (in the case of Hungarian, Turkish and Greek), or far on the lower right corner (Vietnamese, Indonesian). Romanian, a Romance language, appears miscategorized by our language embeddings, also on the lower right corner.

In addition to actual language relationships, represented by color, we also present the result of spectral clustering with four categories through different shapes. This indicates that, broadly, the language embeddings did capture similarities within language families and dissimilarity across language families. Finally, we trained linear models to predict WALS (Dryer and Haspelmath, 2013) features for each language based on the language's 50-dimensional embedding. Even with only 28 training samples (the models were evaluated by leaving one language out), the models predict features related to verbal categories, word order, nominal categories, morphology and lexicon above the level of a majority baseline, with average accuracy of 0.76, while phonology, nominal syntax and other were identical to a majority baseline, with average accuracy 0.72. Despite the surprising failure to capture nominal syntax, it does appear that the language embeddings capture some aspects of syntactic typology.

References

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 261–270, Brussels, Belgium. Association for Computational Linguistics.

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

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised Machine Translation Using Monolingual Corpora Only. *arXiv:1711.00043 [cs]*. ArXiv: 1711.00043.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM.

Dingquan Wang and Jason Eisner. 2017. Fine-Grained Prediction of Syntactic Typology: Discovering Latent Structure with *Supervised* Learning. *Transactions of the Association for Computational Linguistics*, 5:147–161.