Polyglot Parsing for One Thousand and One Languages (And Then Some)

Ali Basirat[*] **Miryam de Lhoneux**[*] **Artur Kulmizev**[*]
Murathan Kurfalı[†] **Joakim Nivre**[*] **Robert Östling**[†]

[*]Department of Linguistics and Philology
Uppsala University

[†]Department of Linguistics
Stockholm University

1 Introduction

Cross-lingual model transfer (Zeman and Resnik, 2008; McDonald et al., 2011) is a commonly used technique for parsing low-resource languages, which relies on the existence of pivot features, such as universal part-of-speech tags or cross-lingual word embeddings. In order for the technique to be really successful, it must also be possible to identify one or more suitable source languages, a task for which language similarity metrics have been exploited (Rosa and Zabokrtsky, 2015). When training parsers on multiple languages, whether for the purpose of model transfer or not, recent studies have also shown that it is beneficial to encode information about language similarity in the form of embeddings, which can be initialized using typological information (Ammar et al., 2016; Smith et al., 2018).

In this project, we try to combine these techniques on an unprecedented scale by building a parser for 1266 low-resource languages, using the following resources:

- Treebanks for 27 languages from Universal Dependencies (Nivre et al., 2016).
- Pre-trained word embeddings for a mostly overlapping set of 27 languages from Facebook (Bojanowski et al., 2016) aligned into a multilingual space (Smith et al., 2017).
- A parallel corpus of Bible translations in the high-resource languages and 1266 additional languages (Mayer and Cysouw, 2014).

The basic idea is to first create cross-linguistically consistent word and language embeddings for all languages based on the embeddings and annotated resources available for the high-resource languages, relying on massively multilingual word alignment to project information to the low-resource languages. Given the word and language embeddings, we can then train a multilingual parser on a suitable subset of treebanks for high-resource languages and use it to parse any of the 1293 languages. In this paper, we present an early progress report, describing the methods used to derive cross-linguistically consistent word and language embedddings, as well as some preliminary parsing experiments.

2 Methods

2.1 Multilingual Word Embeddings

Our goal is to create a multilingual word embedding space that covers all 1293 languages. We achieve this by multi-source projection from the aligned word embeddings of Smith et al. (2017) which are trained on Wikipedia data in 27 languages. First, we perform pairwise word alignment (168 × 1480 texts, since many languages have multiple translations) of the Bible corpus using the bitext alignment tool of Östling and Tiedemann (2016) and use the union of the word alignments produced in each alignment direction. Then, we let the embedding of each low-resource language token be the mean of all tokens in the high-resource languages it is aligned to. Only the 25% of tokens that form the most coherent cluster are used for projection, to compensate for noise in the word alignments.

2.2 Language Embeddings

To suit the parsing task at hand, we use two models aimed at capturing syntactic information about languages.

Language Modeling (LM) This model is based on a word-based language model, using a single LSTM for all languages of the multilingual word embeddings. Sentences from different languages are mixed during training, and the prediction at each timestep is conditioned on the (100-dimensional) embedding of the given language as well as the embedding of the previous word in the sequence. Since it is not straightforward to use standard softmax loss with a multilingual vocabulary, we use the cosine distance between the predicted and actual embeddings of the following word. As we are only interested in learning language embeddings, this turns out to to be sufficient. Word embeddings are fixed during training.

Projected Dependencies (SVD) This model is based on word order features extracted from projected dependency trees. Using pairwise word alignments as per above, dependency link statistics are projected from Bible translations parsed with the tool of Qi et al. (2018) trained on the UD treebanks. We then use maximum spanning tree decoding for each low-resource language, and discard low-confidence dependencies where less than 25% of aligned source texts agree on the dependency relation. Finally, we create a matrix of head-initial/head-final ratio for each (dependency label, head POS, dependent POS) tuple covering all languages, and reduce its dimensionality to 100 using Singular Value Decomposition.

2.3 Multilingual Parsing

We use and extend UUParser[1] (de Lhoneux et al., 2017; Smith et al., 2018), an evolution of the transition-based parser of Kiperwasser and Goldberg (2016). In this parser, BiLSTMs are employed to learn useful representations of tokens in context, while a multi-layer perceptron is used to predict transitions and arc labels, taking as input the BiLSTM vectors of a few tokens at a time. When the parser is applied to data from multiple languages, the representation fed to the BiLSTM for each input token consists of (1) a pre-trained word embedding, (2) the output of a character-level BiLSTM, and (3) a language embedding. In this project, we use the multilingual word embeddings from Section 2.1 as (1) and the language embeddings from Section 2.2 as (3).

3 Preliminary Experiments

Training			Test	
Afrikaans	Finnish	Russian	Arabic	OCSlavonic
Bulgarian	Hungarian	Slovenian	Belarusian	Serbian
Catalan	Indonesian	Spanish	Coptic	Telugu
Danish	Italian	Swedish	Gothic	Urdu
English	Polish	Turkish	Hindi	Uyghur
Estonian	Portuguese	Ukrainian	Marathi	Vietnamese

Table 1: Training and test languages

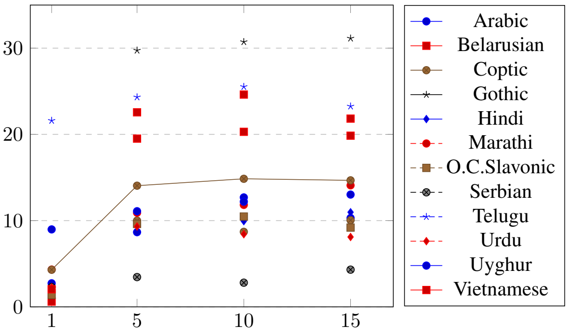

Figure 1: LAS for test languages with training subsets containing, for each test language, the 1, 5, 10 or 15 most similar languages. Average = solid line.

In our preliminary experiments, we have used two disjoint subsets of the total set of 1293 languages, listed in Table 1. The set of *training* languages include all 18 languages that have both a UD treebank (with a training and a development set) and *pre-trained* word embeddings. The set of *test* languages include all 12 languages that have both a UD treebank (with a development set) and *projected* word embeddings. The idea is that, as long as the test language treebanks are used only for evaluation, not for training, the results can be cautiously generalized to other unseen languages. This division results in a training set of size 18, and a test set of size 12 as shown in Table 1.

Figure 1 shows the labeled attachment scores obtained when combining the 1, 5, 10 or 15 most similar training languages for each test languages, as determined by the cosine similarity of the LM language embeddings, and training the parser for 100 epochs on 100 sentences from each training language. We see that the scores generally increase when more training languages are added, although the results are so far very modest with only a few languages reaching a score of 20% or better: Belarusian, Gothic, Telugu and Vietnamese. By comparison, LAS for the training languages range from 27.6 for Turkish to 79.9 for Portuguese with an average of 65.0, which means that the best test language score (Gothic) exceeds the worst training language score (Turkish). One of the challenges for future research is to explain the large variance across languages and relate it to factors such as the quality of word and language embeddings, the similarity of training and test languages, and properties of the parsing architecture.

[1] https://github.com/UppsalaNLP/uuparser

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many languages, one parser. 4:431–444.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional lstm feature representations. 4:313–327.

Miryam de Lhoneux, Yan Shao, Ali Basirat, Eliyahu Kiperwasser, Sara Stymne, Yoav Goldberg, and Joakim Nivre. 2017. From raw text to Universal Dependencies – Look, no tags! In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 207–217.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. pages 62–72.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Dan Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Rudolf Rosa and Zdenek Zabokrtsky. 2015. KLcpos3 – a language similarity measure for delexicalized parser transfer. pages 243–249.

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018. 82 treebanks, 34 models: Universal dependency parsing with multi-treebank models. In *Proceedings of the 2018 CoNLL Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR*, abs/1702.03859.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 35–42.