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

UPPSALA UNIVERSITET

Ali Basirat, Miryam de Lhoneux, Artur Kulmizev
Murathan Kurfalı, Joakim Nivre, Robert Östling

Stockholm University

## Introduction

- **Goal**
  - A parser for 1000+ languages
  - Train a multilingual parser for high-resource languages
  - Use the parser to parser low-resource languages
- **Pivot features**
  - Multilingual word embeddings
  - Language embeddings
- **Resources**
  - Treebanks for 27 languages from Universal Dependencies
  - Pre-trained cross-lingual word embeddings for the 27 languages
  - A parallel corpus of Bible translations for 1293 languages
- **Evaluation**
  - Train the parser on the high-resource languages
  - Test the parser on a sample of low-resource languages
  - Generalize the results to all low-resource languages

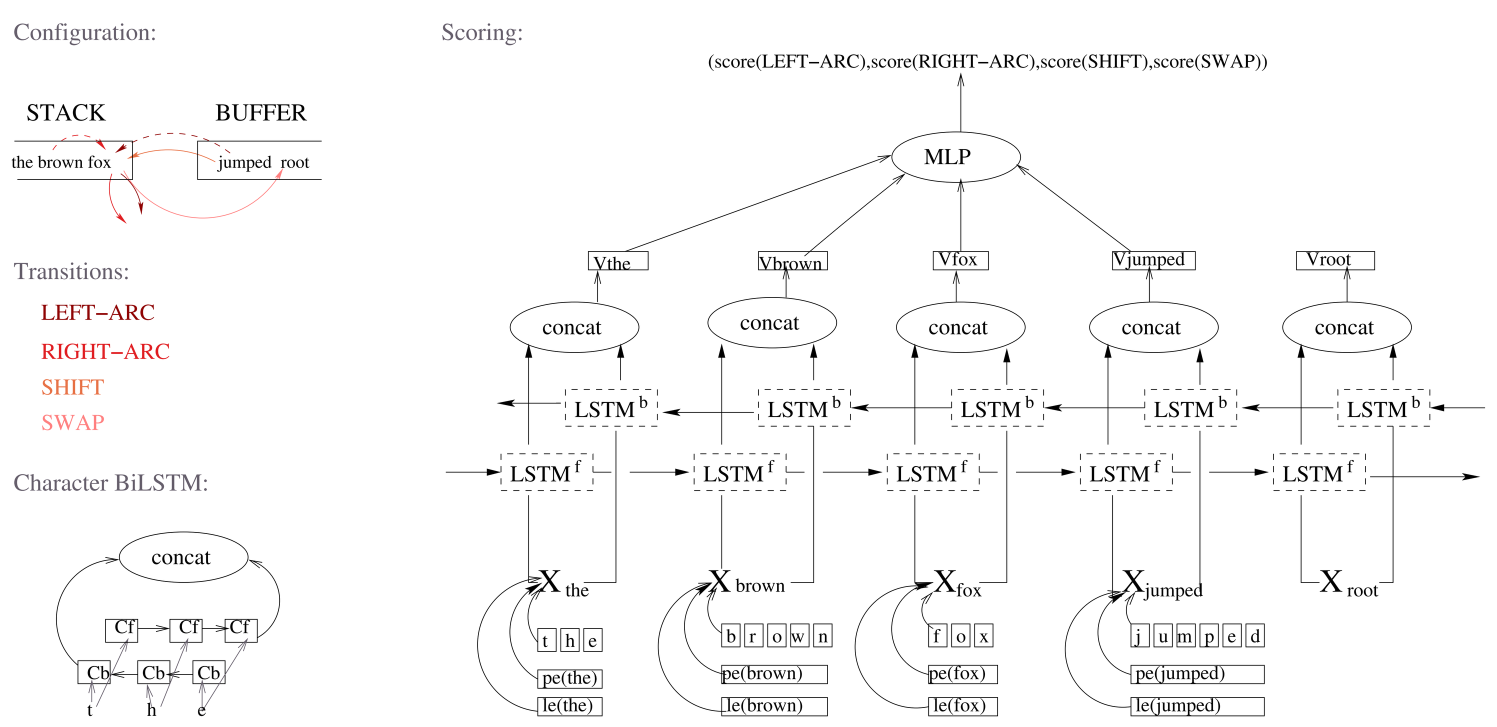
## Methods

- **Multilingual word embeddings**
  - Multilingual embeddings for 27 high resource languages [4]
  - Pairwise word alignments of (168×1480) Bible translations [2]
  - Multi-source projection through word alignments (mean vector of all aligned tokens) for the remaining 1266 low-resource languages
- **Language embeddings**

  Two sets of languages embeddings aimed at capturing the syntactic information about languages:

  (i) *Language embeddings based on language models*
  - A language model (LM) with a single LSTM is trained with fixed word embeddings
  - Prediction is conditioned on 100-dim language embedding and the embedding of the previous word
  - Cosine distance is used as the loss function

  (ii) *Language embeddings based on projected dependencies (SVD)*
  - Bibles of high resource languages are parsed using [3]
  - Dependency link statistics are projected to low resource languages using word alignments
  - Maximum spanning tree decoding for low-resource languages
  - A matrix of head-initial/final ratio for each (dep. label, head/depd. POS) tuple covering all languages is created
  - The dimensionality is reduced to 100 using Singular Value Decomposition (SVD)
- **Multilingual parsing**
  - UUParser [1]: a transition-based dependency parser

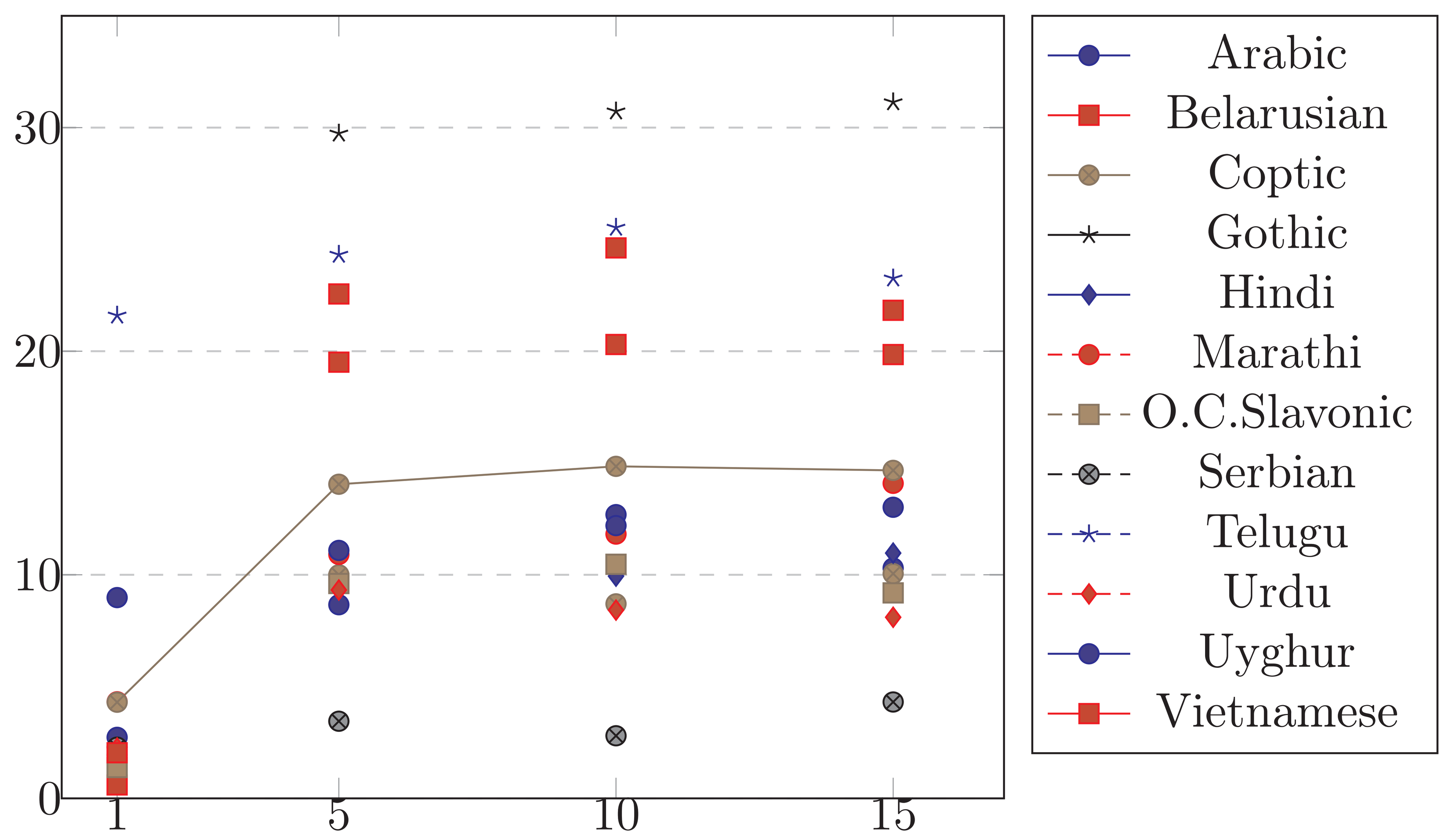



## Preliminary experiments

- **Two disjoint subsets from 1293 languages**
  - 18 training languages have both a UD treebank and pre-trained word embeddings
  - 12 test languages have both a UD treebank and projected word embeddings
  - All languages have LM-based language embeddings

| 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 |

- **Idea: as long as test languages are not used for training, the results can be generalized to all unseen languages**




Y-axis: LAS for test languages with training subsets containing, for each test language, the 1, 5, 10 or 15 most similar languages (X-axis). Average = solid line.

- **Summary of the results**
  - The training languages are selected based on the cosine similarities with the test languages
  - Epochs: 100, number of sentences per training language: 100
  - The parsing scores increase as more training languages are added
  - LAS for the training languages range from 27.6 for Turkish to 79.9 for Portuguese with an average of 65.0
  - The best test language score (Gothic) exceeds the worst training language score (Turkish)
- **Future challenges**
  - Explain the large variance across languages
  - Examine the quality of word and language embeddings
  - Measure the similarity of training and test languages
  - Experiment with different parsing architectures

## References

[1] M. de Lhoneux, Y. Shao, A. Basirat, E. Kiperwasser, S. Stymne, Y. Goldberg, and J. Nivre. 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, 2017.

[2] R. Östling and J. Tiedemann. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146, October 2016.

[3] P. Qi, T. Dozat, Y. Zhang, and C. D. Manning. 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, Oct. 2018. Association for Computational Linguistics.

[4] S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR*, abs/1702.03859, 2017.