Overview

- Goal: Improve Dozat and Manning’s (2017) LSTM parser for CoNLL 2017
  - Higher-precision POS tagger to boost label accuracy
  - Character-level embedding model for morphologically rich languages

Basic parser

- Bidirectional LSTM over word/tag embeddings
- Four separate fully connected ReLU layers
  - Arc-head, arc-dep
  - Rel-head, rel-dep
- Biaffine layer on all arc-dep/arc-head pairs to get the \((n \times n)\) matrix of edge scores

Character model

- Unidirectional LSTM over character embeddings
  - Concatenate together two “summary” vectors:
    - Linear attention over top hidden states (Cao and Rei, 2016)
    - Final cell state (Ballesteros et al., 2015)
  - Linearly transform to the desired size

Conclusion

- High effective POS tagger critical to system’s parsing success
- Character-level model extended effectiveness of Dozat and Manning (2017) to morphologically rich languages
- Primarily tuned on medium/large datasets, more tuning on small datasets may achieve better accuracy
- Graph-based parser demonstrated advantages over transition-based approach on more highly nonprojective treebanks

Tagger study

- Parsers with our tagger outperformed identical ones with UDPipe v1.1’s (Straka et al., 2015) tagger \((p < .05)\)
- Parser improvement correlated with tagger improvement \((p < .05)\)

Character model study

- Adding character model improves performance \((p < .05)\)
- Improvement correlated with morphological complexity \((p < .05)\)