How to Train Dependency Parsers with Inexact Search for Joint Sentence Boundary Detection and Parsing of Entire Documents
Anders Björkelund  Agnieszka Faleńska  Wolfgang Seeker  Jonas Kuhn
IMS, University of Stuttgart, Germany
{anders, falenska, seeker, jonas}@ims.uni-stuttgart.de

Summary

- **Task**: Joint dependency parsing and sentence boundary detection (SBD)
  - SBD is trivial for copy-edited text, but challenging for non-standard orthography (e.g., speech, web content)
  - Four SBD propagates to the parser and deteriorates parsing performance
  - Hypothesis: Syntax can be helpful for finding sentence boundaries
    - That is, a joint system could improve SBD (and possibly parsing)

- **System**: Transition-based parser with sentence boundary transition
  - Beam search for approximate search
  - Operates on documents rather than sentences. Often orders of magnitude more tokens – potential complexity issue
  - Standard training methods for inexact search (early update and max-violation) yield bad models when training on documents

- **Conclusion**: DLaSO outperforms early update and max violation when training on documents
  - Syntax helps to disambiguate sentence boundaries

Training

- Greedy – plain greedy perceptron, uses all training data
- Structured perceptron with beam search
  - Early update – not necessarily using all training data
  - Max-violation – not necessarily using all training data
- DLASO – uses all training data

Why Early and Max-violation Don’t Work

- Early and max-violation do not use all training data when training instances are full documents

Increasing beam size does not help

- Minimal improvements for max-violation
  - Still worse than DLASO

Task

- Predict sentence boundaries and syntactic structure jointly

Data

- *WSJ*: Wall Street Journal, copy-edited (standard)
- *Switchboard*: Spoken transcripts (lowercased, no punct)
- *WSJ*: *WSJ* similar to Switchboard (lowercased, no punct)
- *Joint*: High gains from syntax

Sentiment Boundary Baselines

- OPENNLP – requires punctuation
- CORNLP – requires punctuation
- MARMT – sequence tagger, does not require punctuation
- NOSYNTAX – (joint) parser, but with trivial parse trees

Final Results

- **Sentence boundaries**
  - WSJ: 98.21 76.65 52.82
  - Switchboard: 85.66 78.93 83.37

- **Parsing**
  - WSJ: 99.11 74.98 52.83
  - Switchboard: 89.81 78.93 83.37

References