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,falensaa,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)
  - Poor 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 [Collins and Roark 2004]
  - Max-violation – not necessarily using all training data [Huang et al. 2012]
  - DLASO – uses all training data [Björkelund and Kuhn 2014]

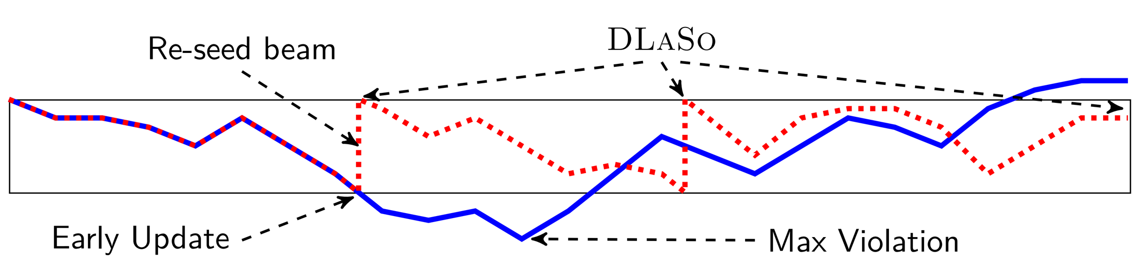
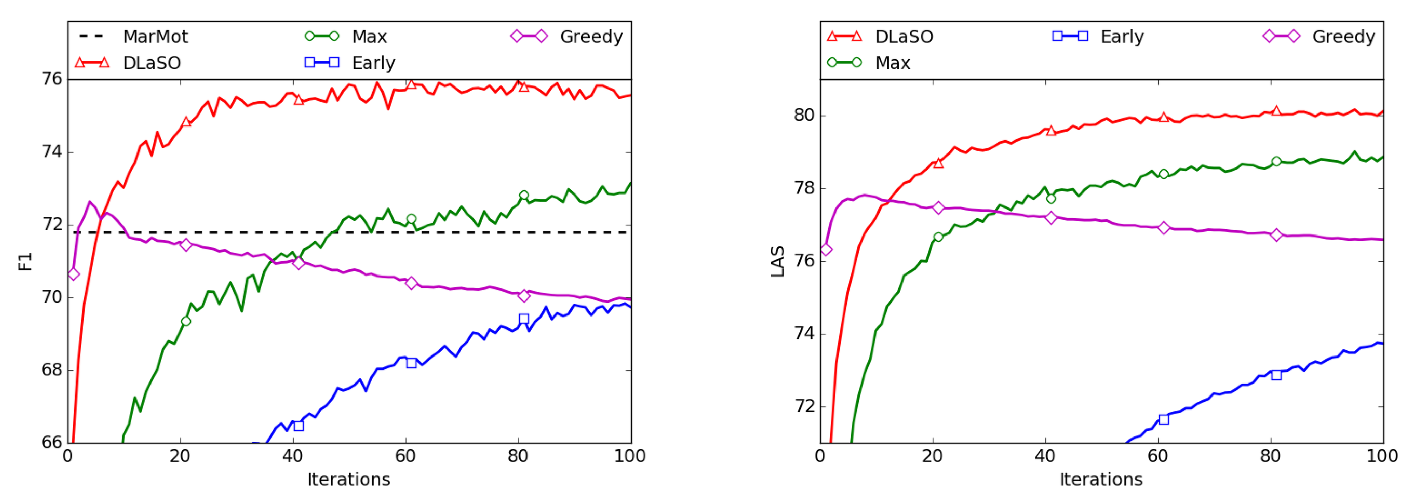


Figure: Difference between training methods for beam search

Performance of update methods

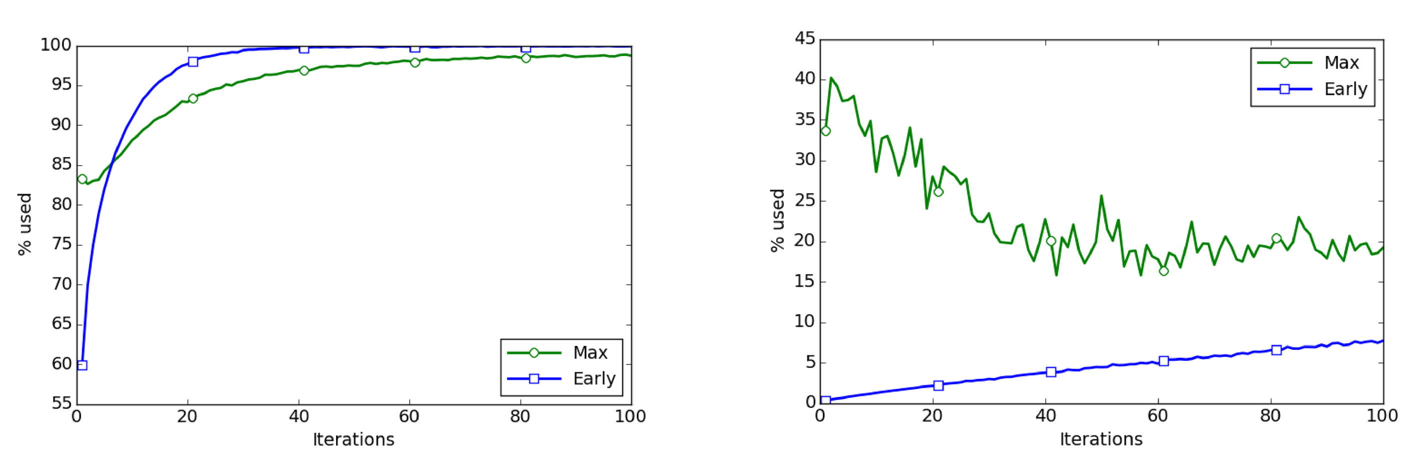


(a) Sentence boundary detection (b) Parsing

Figure: Performance of different update strategies on the Switchboard development set.

Why Early and Max-violation Don't Work

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



(a) Training on sentences (b) Training on documents

Figure: Average length of training sequences used for training for early update and max-violation

Increasing beam size does not help

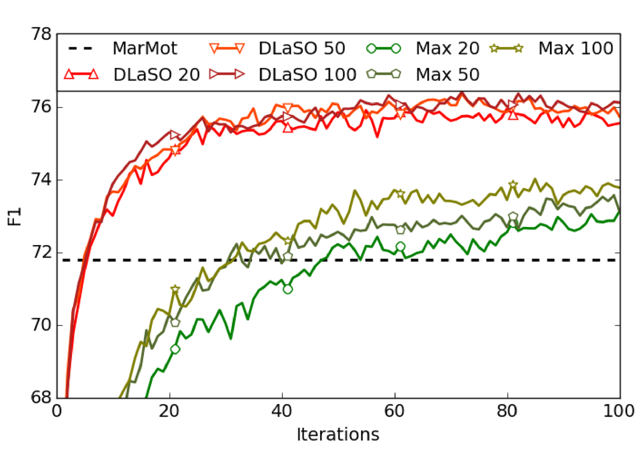


Figure: SBD F_1 when varying beam size

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

Task

Predict sentence boundaries and syntactic structure jointly

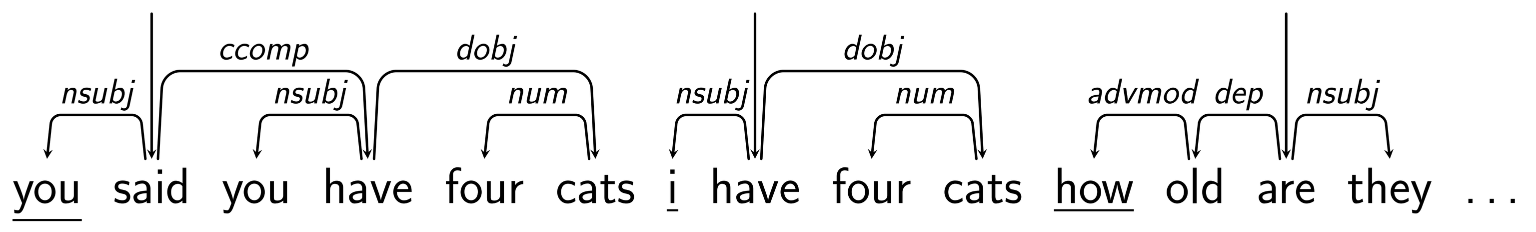


Figure: Sample document from the Switchboard corpus. Sentence starts are underlined

Transition System

- ArcStandard system with SWAP [Nivre 2009]
- Additional transition SB marks new sentences
- State augmented to hold sentence-initial tokens

Transition				Preconditions
LEFTARC	$(\sigma\|s_1\|s_0, \beta, A, S)$	$\Rightarrow$	$(\sigma\|s_0, \beta, A \cup \{s_0 \rightarrow s_1\}, S)$	$s_1 \neq 0$
RIGHTARC	$(\sigma\|s_1\|s_0, \beta, A, S)$	$\Rightarrow$	$(\sigma\|s_1, \beta, A \cup \{s_1 \rightarrow s_0\}, S)$	
SHIFT	$(\sigma, b_0\|\beta, A, S)$	$\Rightarrow$	$(\sigma\|b_0, \beta, A, S)$	$b_0 \neq \mathrm{LAST}(S) \vee \|\sigma\| = 1 \vee \mathrm{SWAPPED}(\beta)$
SWAP	$(\sigma\|s_1\|s_0, \beta, A, S)$	$\Rightarrow$	$(\sigma\|s_0, s_1\|\beta, A, S)$	$s_1 < s_0$
SB	$(\sigma, b_0\|\beta, A, S)$	$\Rightarrow$	$(\sigma, b_0\|\beta, A, S \cup \{b_0\})$	$\mathrm{LAST}(S) < b_0 \wedge \neg\mathrm{SWAPPED}(\beta)$

Figure: Transition system

Experimental Setup

- **Data**
  - WSJ: Wall Street Journal, copy-edited (standard)
  - Switchboard: Spoken transcripts (lowercased, no punct)
  - WSJ*: WSJ similar to Switchboard (lowercased, no punct)
- **Evaluation**
  - Sentences: F-measure on sentence-initial tokens
  - Parsing: Labeled Attachment Score (LAS)
- **Sentence Boundary Baselines**
  - OPENNLP – requires punctuation http://opennlp.apache.org
  - CORENLP – requires punctuation [Manning et al. 2014]
  - MARMOT – sequence tagger, does not require punctuation [Müller et al. 2013]
  - NOSYNTAX – (joint) parser, but with trivial parse trees

Baseline SBD performance

- OPENNLP and CORENLP can't be applied on Switchboard and WSJ* due to lack of punctuation
- All systems roughly equal on WSJ
- MarMot and NoSyntax are reasonable baselines

	WSJ	Switchboard	WSJ*
OPENNLP	98.09	–	–
CORENLP	98.60	–	–
MARMOT	98.21	71.78	52.82
NOSYNTAX	99.11	74.98	52.83

Table: Dev set results (F_1) for baselines

Final Results

Sentence boundaries

- WSJ: All roughly equal
- Switchboard: Low syntactic complexity, no improvement with JOINT
- WSJ*: High gains from syntax (JOINT)

	WSJ	Switchboard	WSJ*
MARMOT	97.64	71.87	53.02
NOSYNTAX	98.21	76.31†	55.15
JOINT	98.21	76.65†	65.34†‡

†: significant improvement over MARMOT
‡: significant improvement over NOSYNTAX

Table: Test set SBD results (F_1)

Parsing

- WSJ: All roughly equal
- Switchboard: Slight improvements over baselines
- WSJ*: Big advantage for JOINT

	WSJ	Switchboard	WSJ*
GOLD	90.22	84.99	88.71
MARMOT	89.81	78.93	83.37
NOSYNTAX	89.95	80.30†	83.61
JOINT	89.71	79.97†	85.66†‡
JOINT-REPARSED	89.93	80.61†‡*	85.38†‡

†: significant improvement over MARMOT
‡: significant improvement over NOSYNTAX
*: significant improvement over JOINT

Table: Test set parsing results (LAS)

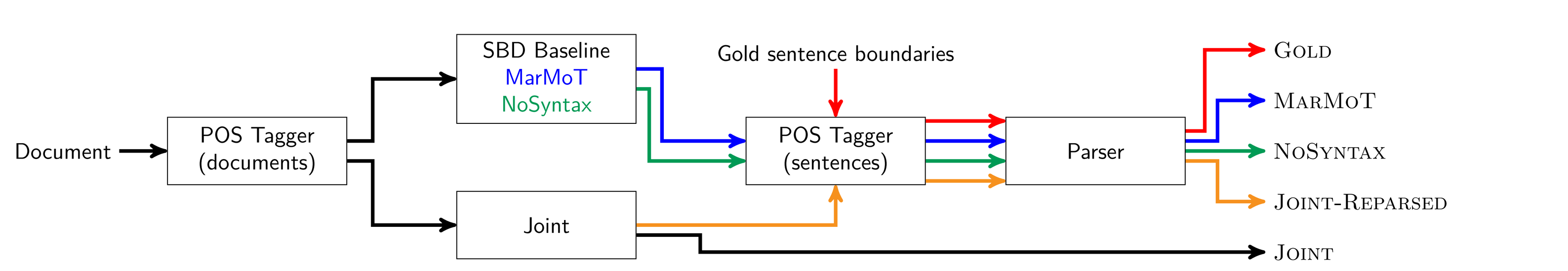


Figure: Overview of parsing pipelines

Acknowledgements

This work was supported by the German Research Foundation (DFG) in project D8 of SFB 732. We also thank spaCy GmbH for making it possible for the third author to finish his contribution to this work.

References

Björkelund, A. and Kuhn, J. (2014). Learning Structured Perceptrons for Coreference Resolution with Latent Antecedents and Non-local Features. In *ACL*.

Collins, M. and Roark, B. (2004). Incremental Parsing with the Perceptron Algorithm. In *ACL*.

Huang, L., Fayong, S., and Guo, Y. (2012). Structured Perceptron with Inexact Search. In *NAACL-HLT*.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *ACL System Demonstrations*.

Müller, T., Schmid, H., and Schütze, H. (2013). Efficient higher-order CRFs for morphological tagging. In *EMNLP*.

Nivre, J. (2009). Non-projective dependency parsing in expected linear time. In *ACL-IJCNLP*.