The MIT Press

Journals

Books   Journals   Digital   Resources   About   Contact

Journal Resources

Editorial Info
Abstracting and Indexing
Release Schedule
Advertising Info

Author Resources

Submission Guidelines
Publication Agreement
Author Reprints

Parsing Noun Phrases in the Penn Treebank

David Vadas and James R. Curran

◎ **Download Options** ＞

Abstract Authors

Noun phrases (nps) are a crucial part of natural language, and can have a very complex structure. However, this np structure is largely ignored by the statistical parsing field, as the most widely used corpus is not annotated with it. This lack of gold-standard data has restricted previous efforts to parse nps, making it impossible to perform the supervised experiments that have achieved high performance in so many Natural Language Processing (nlp) tasks.

We comprehensively solve this problem by manually annotating np structure for the entire Wall Street Journal section of the Penn Treebank.

The inter-annotator agreement scores that we attain dispel the belief that the task is too difficult, and demonstrate that consistent np annotation is possible. Our gold-standard np data is now available for use in all parsers.

We experiment with this new data, applying the Collins (2003) parsing model, and find that its recovery of np structure is significantly worse than its overall performance. The parser's F-score is up to 5.69% lower than a baseline that uses deterministic rules. Through much experimentation, we determine that this result is primarily caused by a lack of lexical information.

To solve this problem we construct a wide-coverage, large-scale np Bracketing system. With our Penn Treebank data set, which is orders of magnitude larger than those used previously, we build a supervised model that achieves excellent results. Our model performs at 93.8% F-score on the simple task that most previous work has undertaken, and extends to bracket longer, more complex nps that are rarely dealt with in the literature. We attain 89.14% F-score on this much more difficult task. Finally, we implement a post-processing module that brackets nps identified by the Bikel (2004) parser. Our np Bracketing model includes a wide variety of features that provide the lexical information that was missing during the parser experiments, and as a result, we outperform the parser's F-score by 9.04%.

These experiments demonstrate the utility of the corpus, and show that many nlp applications can now make use of np structure.

Forthcoming

Most Read   See More

⚲ Lexicon-Based Methods for Sentiment Analysis (14019 times)
Maite Taboada et al.
Computational Linguistics
Volume: 37, Issue: 2, pp. 267-307

⚲ Computational Linguistics and Deep Learning (10513 times)
Christopher D. Manning
Computational Linguistics
Volume: 41, Issue: 4, pp. 701-707

⚲ Near-Synonymy and Lexical Choice (3658 times)
Philip Edmonds et al.
Computational Linguistics
Volume: 28, Issue: 2, pp. 105-144

(Note that the Most Read numbers are based on the number of full text downloads over the last 12 months.)

Most Cited

See More

⚲ Lexicon-Based Methods for Sentiment Analysis (436 times)
Maite Taboada et al.
Computational Linguistics
Volume: 37, Issue: 2, pp. 267-307

⚲ A Systematic Comparison of Various Statistical Alignment Models (174 times)
Franz Josef Och et al.
Computational Linguistics
Volume: 29, Issue: 1, pp. 19-51

⚲ Opinion Word Expansion and Target Extraction through Double Propagation (147 times)
Guang Qiu et al.
Computational Linguistics
Volume: 37, Issue: 1, pp. 9-27

(Note that the Most Cited numbers are based on Crossref's Cited-by service and reflect citation information for the past 24 months.)
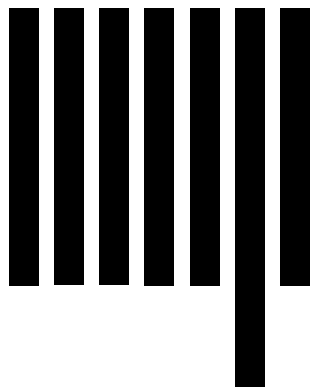
◎ Download Options ⟩

Favorite ♡

Sign up for Alerts 🔔

Download Citation ↓

RSS TOC 📶

RSS Citation 📶

Submit your article

Support OA at MITP 🔒

Journals

Terms & Conditions

Privacy Statement

Contact Us

Conditions Statement Us

Books

US

UK

Connect

© 2018 The MIT Press

One Rogers Street Cambridge MA 02142-1209

Suite 2, 1 Duchess Street London, W1W 6AN, UK

Technology Partner: Atypon Systems, Inc.
CrossRef Member
COUNTER Member
The MIT Press colophon is registered in the U.S. Patent and Trademark Office.
Site Help