The greatest PoS tags, just the best really

Anonymous ACL submission

Abstract

Abstraction ensues.

1 Introduction

2 Previous/Related work

Megyesi, Swedish
MacKinlay, English
OBT, Norwegian

3 Data

3.1 The Norwegian Dependency Treebank

The newly developed Norwegian Dependency Treebank (NDT) (Solberg et al., 2014) is the first publicly available treebank for Norwegian. The treebank was used for training and evaluation in our experiments. The tag set consists of 12 morphosyntactic PoS tags, with 7 additional tags for punctuation and symbols.

PoS Tag Set

Dependency Relations

4 Experimental Setup

Tagger TnT (Brants, 2000)

Parser In choosing a syntactic parser for our experiments, we considered previous work on dependency parsing of Norwegian, specifically that of (Solberg et al., 2014). They found the Mate parser (Bohnet, 2010) to be the most successful parser for the parsing of NDT. Recent dependency parser comparisons (Choi et al., 2015) also showed that Mate performed very well on parsing of the English portion of the OntoNotes 5 corpus, beating a range of contemporary state-of-the-art parsers.

Tag Set Mapping In order to alter the tag set of NDT, we created a mapping for carrying out the tag set modifications. We created a mapping for carrying out the tag set modifications that maps the relevant existing tags to new, more fine-grained tags including more relevant morphological features for the applicable tokens.

Baseline It is common practice to compare the performance of PoS taggers to a pre-computed baseline for an initial point of comparison. For PoS tagging, a commonly used baseline is the Most Frequent Tag (MFT) baseline, which we use in our experiments. This involves labeling each word with the tag it was assigned most frequently in the training. All unknown words, i.e., words not seen in the training data, are assigned the tag most frequently assigned to words seen only once in the training. Unknown and infrequent words have in common that they rarely occur, and we might therefore expect them to have similar properties.

Tags & Features As we seek to quantify the effects of PoS tagging in a realistic setting, we want to run the parser on automatically assigned PoS tags. For the training of the parser, however, we have two options: using either gold standard or automatically assigned tags. In order to settle on a configuration, we conducted experiments with gold standard and automatically assigned tags to see how they differ with respect to performance. The results of our experiments reveal that the combination of training and testing on automatic tags is superior to training on gold standard tags and testing on automatic tags, surprisingly. Consequently, the parser was both trained and tested on automatically assigned tags in our experiments.

Note that it is absolutely crucial that the morphological features in the treebank are removed when using automatic tags, as they are still gold standard. For instance, if a verb token is erroneously tagged as a noun, we could potentially have a noun token with verbal features such as tense, which markedly obfuscates the training and parsing. Another important factor is that we want to isolate the effect of PoS tags, necessitating the exclusion of morphological features.

5 Tag Set Optimization

6 Optimized Pipeline

7 Summary/Conclusion and Future Work

References

Bernd Bohnet. 2010. Very High Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97, Beijing, China.

Thorsten Brants. 2000. TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference*, Seattle, WA, USA.

Jinho D. Choi, Joel Tetreault, and Amanda Stent. 2015. It Depends: Dependency Parser Comparison Using A Web-Based Evaluation Tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 387–396, Beijing, China.

Per Erik Solberg, Arne Skjrholt, Lilja vrelid, Kristin Hagen, and Janne Bondi Johannessen. 2014. The Norwegian Dependency Treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 789–795, Reykjavik, Iceland.