Optimizing a PoS Tag Set for Dependency Parsing

Anonymous ACL submission

Abstract

Abstraction ensues.

1 Introduction

PoS tagging is an important preprocessing step for many NLP tasks, such as dependency parsing (Nivre et al., 2007; Hajič et al., 2009), named entity recognition (Tjong Kim Sang and De Meulder, 2003) and sentiment analysis (Wilson et al., 2009). Whereas much effort has gone into the development of PoS-taggers, to the effect that this task is often considered more or less a solved task, considerably less effort has been devoted to the empirical evaluation of the PoS tag sets themselves. Error analysis of PoS-taggers indicate that, whereas tagger improvement though means of learning algorithm or feature engineering seems to have reached something of a plateau, linguistic assessment of the distinctions made by the PoS tag set may be a venue worth investigating further (Manning, 2011). Clearly, the utility of a PoS tag set is tightly coupled with the the downstream task for which it is performed. Even so, PoS tag sets are usually employed in a "one size fits all" fashion, regardless of the requirements posed by the task which makes use of this information.

It is well known that syntactic parsing often benefits from quite fine-grained morphological distinctions (Zhang and Nivre, 2011; Seeker and Kuhn, 2013). Morphology interacts with syntax though phenomena such as agreement and case marking and incorporating information on morphological properties of words can therefore often improved parsing performance. However, in a realistic setting where the aim is to automatically parse raw text, the generation of morphological information will often require a separate step of morphological analysis that can be quite costly.

In this paper, we optimize a Norwegian PoS tag set for the task of data-driven dependency parsing. We show that the introduction of morphological distinctions not present in the original tag set, whilst compromising tagger accuracy, actually leads to significantly improved parsing accuracy. In a set of experiments, various morphological distinctions are introduced and evaluated both intrinsically, i.e. in terms of PoS-tagging accuracy, and extrinsically, in terms of parsing accuracy. Our results show that the introduction of morphological distinctions not present in the original tag set, whilst compromising tagger accuracy, actually leads to significantly improved parsing accuracy. This optimization allows us to bypass the additional step of morphological analysis, framing the whole preprocessing problem as a simple tagging task.

The article is structured as follows. We start out by reviewing previous work on tag set evaluation in Section 2 and Section 3 provides an overview of the treebank used for experimentation. Section 4 describes the experimental setup used in our work and Section 5 goes on to provide the results from our optimization experiments. Finally, Section 7 summarizes our main findings and discusses some avenues for future work.

2 Previous/Related work

There has been little previous work dedicated specifically to extrinsic evaluation of the effects of PoS tag sets on downstream applications. There has, however, been some work that evaluates the effects of PoS tag set granularity on PoS tagging. This includes investigation of the effects of PoS tag sets on tagging of Swedish (Megyesi, 2001; Megyesi, 2002) and English (MacKinlay, 2005).

(Megyesi, 2001; Megyesi, 2002) trained and evaluated a range of PoS taggers on the Stockholm-Umeå Corpus (SUC) (Gustafson-Capková and Hartmann, 2006), annotated with a tag set based on a Swedish version of PAROLE tags totaling 139 tags. Furthermore, they investigated the effects of tag set size on tagging by mapping the original tag set into smaller subsets designed for parsing. They argue that a tag set with complete morphological tags may not be necessary for all NLP applications, for instance syntactic parsing. They found that the smallest tag set comprising 26 tags yields the lowest tagger error rate. However, for some of the taggers, augmenting the tag set with more linguistically informative tags may actually lead to a drop in error rate. They argue that this shows that the size of the tag set as well as the type of information in the tags are crucial factors for tagger performance. However/unfortunately, they do not report results of parsing with the various PoS tag sets.

Similarly, MacKinlay (2005) investigated the effects of PoS tag sets on tagger performance in English, specifically the Wall Street Journal portion of Penn Treebank (Marcus et al., 1993). They mapped the original tag set of Penn Treebank to more fine-grained tag sets using linguistic insight to investigate whether additional linguistic information included in finer-grained tags could assist the tagger. Experimenting with both lexically and syntactically conditioned modifications such as distinguishing between count nouns and non-count nouns and between transitive and intransitive verbs, they found that more fine-grained tag sets rarely led to improvements in tagger accuracy; the most successful modification yielded an improvement in tagger accuracy of 0.05 percentage points.

Transition/overgang her ...

3 Data

3.1 The Norwegian Dependency Treebank

We used the newly developed Norwegian Dependency Treebank (NDT) (Solberg et al., 2014), the first publicly available treebank for Norwegian. It was developed at the National Library of Norway in collaboration with the University of Oslo, and contains manual syntactic and morphological annotation. The treebank contains data from both varieties of Norwegian; 311 000 tokens of Bokmål and 303 000 tokens of Nynorsk. The annotated texts are mostly newspaper text, but also include government reports, parliament transcripts and ex-

Tag	Description
`adj`	Adjective
`adv`	Adverb
`det`	Determiner
`inf-merke`	Infinitive marker
`interj`	Interjection
`konj`	Conjunction
`prep`	Preposition
`pron`	Pronoun
`sbu`	Subordinate conjunction
`subst`	Noun
`ukjent`	Unknown (foreign word)
`verb`	Verb

Table 1: Overview of the PoS tag set of NDT.

cerpts from blogs. The annotation process of the treebank was supported by the Oslo-Bergen Tagger (Hagen et al., 2000) and then manually corrected by annotators.

Morphological Annotation The morphological annotation and PoS tag set of NDT follows the Oslo-Bergen Tagger (Hagen et al., 2000; Solberg, 2013), which in turn is largely based on the work of Faarlund et al. (1997). The tag set consists of 12 morphosyntactic PoS tags, with 7 additional tags for punctuation and symbols. The tag set is thus rather coarse-grained, with broad categories such as `subst` (noun) and `verb` (verb). The PoS tags are complemented by a large set of morphological features, providing information about morphological properties such as definiteness, number and tense. These features are used in our tag set modifications, where the coarse PoS tag of relevant tokens are concatenated with one or more of these features to include more linguistic information in the tags.

Syntactic Annotation The syntactic annotation choices in NDT are largely based on the Norwegian Reference Grammar (Faarlund et al., 1997). The annotation choices are outlined in Table 2, taken from/courtesy of Solberg et al. (2014), providing overview of the analyses of syntactic constructions that often distinguish dependency treebanks, such as coordination and the treatment of auxiliary and main verbs. The set of dependency relations comprises 29 dependency relations, including ADV (adverbial), SUBJ (subject) and KOORD (coordination).

Head	Dependent
Preposition	Prepositional complement
Finite verb	Complementizer
First conjunct	Subsequent conjuncts
Finite auxiliary	Lexical/main verb
Noun	Determiner

Table 2: Annotation choices in NDT.

4 Experimental Setup

In preparation to conducting our experiments with linguistically motivated tag set modifications, a concrete setup for the experiments needed to be established, which is presented in the following.

Data Set Split As there was no standardized data set split of NDT due to its very recent development, we needed to establish a data set split (training/development/test). Our data set split of the treebank follows the standard 80-10-10 (training/development/test) split and will be distributed with the treebank and proposed as the new standard(?). In creating the data set split, care has been taken to preserve contiguous texts in the various data sets while keeping the split balanced in terms of genre (and source). Our proposed data set split was used in the Norwegian contribution to the Universal Dependencies project (Øvrelid and Hohle, 2016). The split will be made available at a companion website.

Tagger For our experiments with tag set modifications, we wanted a PoS tagger both reasonably fast and accurate. There is often a trade-off between the two factors/considerations, as the most accurate taggers tend to suffer in terms of speed due to their complexity. However, a tagger that achieves both close to state-of-the-art accuracy as well as very high speed is TnT (Brants, 2000). The fact that TnT was used for evaluating the universal tag set (Petrov et al., 2012), served as another good indication of TnT being appropriate for our task. The sum of these factors led to TnT being the tagger of choice for our experiments.

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. Furthermore, recent dependency parser comparisons (Choi et al., 2015) showed that Mate performed very well on parsing of English, beating a range of contemporary state-of-the-art parsers.

Tag Set Mapping In order to carrying out the tag set modifications, we created a mapping that maps the relevant existing tags to new, more fine-grained tags including more relevant morphological features for the applicable tokens. A given tag set modification involves specifying a tag and one or more features to be concatenated with said tag for all applicable tokens, i.e., tokens that are assigned said tag and contains the set of features in its morphological features.

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. Consequently, the parser was both trained and tested on automatically assigned tags in our experiments.

We removed any morphological features in order to simulate a realistic setting. Moreover, it is crucial that we remove these features when working with automatically assigned tags, as the features is 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 aspect is that we want to isolate the effect of PoS tags, necessitating the exclusion of morphological features.

5 Tag Set Optimization/Experiments

With more fine-grained linguistically motivated distinctions, we increase the linguistic information represented in the tags, which may assist the tagger in disambiguating ambiguous and unknown words, which in turn may aid the parser in recognizing and generalizing syntactic patterns. However, the addition of more linguistic information to the tags and thus a more fine-grained tag set will most likely lead to a drop in tagger accuracy, due to the increase in complexity. The best tagging does not necessarily lead to the best parse, and it is therefore interesting to investigate how the tag set modifications may affect the interplay between tagging and parsing.

5.1 Tag Set Experiments

We modified the tags for nouns, verbs, adjectives, determiners and pronouns in NDT by appending selected sets of morphological features to each tag in order to increase the linguistic information expressed by the tags. For each tag, we first experimented with each of the features in isolation before employing various combinations of them. We based our choices of combinations on how promising the features are and what we deem worth investigating in terms of linguistic utility, in order to see how the features might interact.

Nouns In Norwegian, nouns are assigned gender (feminine, masculine, neuter), definiteness (indefinite or definite) and number (singular or plural). There is agreement in gender, definiteness and number between nouns and their modifiers (adjectives and determiners).

Verbs Verbs are inflected for tense (infinitive, present, preterite, past perfect) in Norwegian and can additionally take on mood (imperative, indicative) and voice (active or passive).

Adjectives Adjectives agree with the noun they modify in terms of gender, number and definiteness in Norwegian.

Determiners Like adjectives, determiners in Norwegian agree with the noun they modify in terms of gender, number and definiteness.

Pronouns Pronouns in Norwegian include personal, reciprocal, reflexive and interrogative. They can exhibit gender, number and person. Personal pronouns have case (accusative or nominative).

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.

Jan Terje Faarlund, Svein Lie, and Kjell Ivar Vannebo. 1997. *Norsk referansegrammatikk*. Universitetsforlaget, Oslo, Norway.

Sofia Gustafson-Capková and Britt Hartmann. 2006. Manual of the Stockholm Umeå Corpus version 2.0.

Kristin Hagen, Janne Bondi Johannessen, and Anders Nøklestad. 2000. A Constraint-Based Tagger for Norwegian. In *Proceedings of the 17th Scandinavian Conference of Linguistics*, pages 31–48, Odense, Denmark.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpanek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 Shared Task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the 6th Conference on Natural Language Learning*.

Andrew MacKinlay. 2005. The Effects of Part-of-Speech Tagsets on Tagger Performance. Bachelor's thesis, University of Melbourne, Melbourne, Australia.

Christopher Manning. 2011. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 171–189.

Category	Feature(s)	MFT	Accuracy	LAS	UAS
Baseline	—	94.14%	97.47%	87.01%	90.19%
Noun	Type, case & definiteness	89.61%	97.05%	88.81%	91.73%
Verb	Finiteness	93.72%	97.35%	87.30%	90.43%
Adjective	Degree	94.13%	97.41%	87.29%	90.44%
Determiner	Definiteness	94.13%	97.49%	87.30%	90.42%
Pronoun	Type & case	94.12%	97.51%	87.30%	90.41%

Table 3: Results of tagging and parsing with the most successful tag set modification for each category.

Tag	Description
`adj|komp`	Comparative adjective
`adj|pos`	Positive adjective
`adj|sup`	Superlative adjective
`det|be`	Definite determiner
`det|ub`	Indefinite determiner
`pron|pers`	Personal pronoun
`pron|pers|akk`	Personal pronoun, accusative
`pron|pers|nom`	Personal pronoun, nominative
`pron|refl`	Reflexive pronoun
`pron|res`	Reciprocal pronoun
`pron|sp`	Interrogative pronoun
`subst|appell`	Common noun
`subst|appell|be`	Common noun, definite
`subst|appell|be|gen`	Common noun, definite, genitive
`subst|appell|ub`	Common noun, indefinite
`subst|appell|ub|gen`	Common noun, indefinite, genitive
`subst|prop`	Proper noun
`subst|prop|gen`	Proper noun, genitive
`verb|fin`	Finite verb
`verb|infin`	Nonfinite verb

Table 4: Overview of the optimized tag set.

Tag set	MFT	Accuracy	LAS	UAS
Original	**94.14%**	**97.47%**	87.01%	90.19%
Full	85.12%	93.46%	87.13%	90.32%
Optimized	89.20%	96.85%	**88.87%**	**91.78%**

Table 5: Results of tagging and parsing with the optimized tag set, compared to the initial tag sets.

Mitchell Marcus, Beatrice Santorino, and Mary Ann Marcinkiewicz. 1993. Building A Large Annotated Corpus of English: The Penn Treebank. Technical report, University of Philadelphia, Philadelphia, PA, USA.

Beáta Megyesi. 2001. Comparing Data-Driven Learning Algorithms for PoS Tagging of Swedish. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 151–158, Pittsburgh, PA, USA.

Beáta Megyesi. 2002. *Data-Driven Syntactic Analysis: Methods and Applications for Swedish*. Ph.D. thesis, Royal Institute of Technology, Stockholm, Sweden.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the 6th Conference on Natural Language Learning*, pages 915 – 932.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 2089–2096, Istanbul, Turkey.

Wolfgang Seeker and Jonas Kuhn. 2013. Morphological and syntactic case in statistical dependency parsing. *Computational Linguistics*, 39(1):23–55.

Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, 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.

Per Erik Solberg. 2013. Building Gold-Standard Treebanks for Norwegian. In *Proceedings of the 19th Nordic Conference of Computational Linguistics*, pages 459–464, Oslo, Norway.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, pages 142–147, Stroudsburg, PA, USA.

Theresa Wilson, Janyce Wiebe, and Paul Hoffman. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399 – 433.

Yue Zhang and Joakim Nivre. 2011. Transition-Based Dependency Parsing with Rich Non-Local Features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Lingustics: Human Language Technologies*, pages 188–193, Portland, OR, USA.

Lilja Øvrelid and Petter Hohle. 2016. Universal Dependencies for Norwegian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, Portorož, Slovenia.