Improving the Accuracy of Polish POS Tagging by Using Voting Ensembles

Łukasz Kobyliński

Institute of Computer Science, Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland
lkobylinski@ipipan.waw.pl

Abstract

Recently, several new part-of-speech (POS) taggers for Polish have been presented. This is highly desired, as the quality of morphosyntactic annotation of textual resources (especially reference text corpora) has a direct impact on the accuracy of many other language-related tasks in linguistic engineering. For example, in the case of corpus annotation, most automated methods of producing higher levels of linguistic annotation expect an already POS-analyzed text on their input. In spite of the improvement of Polish tagging quality, the accuracy of even the best-performing taggers is still well below 100% and the mistakes made in POS tagging propagate to higher layers of annotation. One possible approach to further improving the tagging accuracy is to take advantage of the fact that there are now quite a few taggers available and they are based on different principles of operation. In this paper we investigate this approach experimentally and show improved results of POS tagging accuracy, achieved by combining the output of several state-of-the-art methods.

Keywords: POS tagging, combining taggers, voting ensembles, Polish taggers.

1. Introduction

Part-of-speech tagging is a central task of Natural Language Processing (NLP). New and improved methods of automatically assigning POS tags to tokens in text are constantly being researched, as the accuracy of morphosyntactic annotation is important on its own and influences the results of other NLP tasks, such as parsing or word sense disambiguation. For English, the task may be considered nearly solved, as taggers achieve an accuracy of over 97%. In the case of highly inflectional languages, such as Polish, there is still a large margin of tagger-made mistakes, as the authors of even the best taggers report accuracy not higher than 91% (Waszczuk, 2012).

The problem of producing an accurate morphosyntactic layer of annotation is of a crucial importance in the case of text corpora. Such corpora are either annotated manually by qualified linguists, or automatically, using taggers. For large corpora, such as the National Corpus of Polish (Przepiórkowski et al., 2012), which contains more than 1 billion tokens, relying on manual tagging of the whole corpus is infeasible, because of time and cost constraints. Both manual and automated methods are thus often used, by annotating only a selected, representative part of a corpus by hand and using it as a gold-standard annotation to train automated taggers. Taggers are then used to generate annotations for the remaining part of the corpus.

Independently of the research concerning new taggers, one of the important trends in machine learning is the observation that often a combination of multiple classifiers may produce more accurate results, than any of the individual classifiers on its own. It is because the errors made by different classifiers are usually not exactly the same and there exists at least a theoretical possibility of creating a new, better-performing method by selecting the right classifier from an ensemble for each classified test case.

2. Previous Work

There have been many attempts to create an ensemble of taggers for English and other languages, for which multiple tagging methods exist. In the case of English, Brill and Wu (1998) reported a considerable reduction (6.9%) of the number of errors produced in tagging by using a voted combination of three different taggers. Van Halteren et al. (2001) achieved a much higher, 19.1% reduction of tagging error rate in comparison to the best individual tagger. The authors have adopted a simple voting strategy, pairwise voting and stacking of four different taggers: a trigram tagger, a transformation based learning system, a tagger built around a memory-based learning method and a maximum entropy model.

In the case of Polish, an evaluation of an ensemble of taggers has been presented by Śniatowski and Piasecki (2012). The performance of the system has been estimated using a now outdated tagset and a smaller corpus consisting of ca. 880 000 tokens. The authors have also used a method of evaluation, which is now considered to give unfair advantage to some taggers, as it measures the POS tagging disambiguation accuracy and not the accuracy of tagging plain text, which is usually the real world scenario. In another evaluation Radziszewski and Śniatowski (2011a) have provided results of experiments of combining three taggers using the currently used tagset and a larger corpus, but still employing the approach measuring the disambiguation accuracy, as opposed to the accuracy of tagging plain text.

3. Tagger Evaluation Procedure

As indicated in the previous section, there is an important distinction to be made between different understandings of what a tagger evaluation procedure is. Radziszewski and Acedański (2012) pointed out that in many recent tagger comparisons the accuracy of morphosyntactic disambiguation is commonly reported as the tagger performance metric. This is however biased, as the accuracy of tagging unknown words is not evaluated properly in such a scenario, as the correct tag is always present as one of the options in the tagger input. Furthermore, in a real-world scenario the user is interested in tagging plain

text and does not have access to previously morphologically analyzed resources.

Following that line of thought, we have decided to perform the comparisons using plain text as input and report tagger accuracy using the *accuracy lower bound* (Acc_{lower}) metric, proposed by Radziszewski and Acedański (2012). The metric penalizes all segmentation changes in regard to the gold standard and treats such tokens as misclassified. Furthermore, we report separate metric values for both known and unknown words to assess the performance of guesser modules built into the taggers. These are indicated as Acc^K_{lower} for known and Acc^U_{lower} for unknown words.

All the experiments have been performed on the manually annotated part of the National Corpus of Polish (Przepiórkowski et al., 2012), version 1.1, which consists of ca. 1 million tokens. We will refer to this dataset as NCP1M. The ten-fold cross-validation procedure has been followed, by re-evaluating the methods ten times, each time selecting one of ten parts of the corpus for testing and the remaining parts for training the taggers. The provided results are averages calculated over ten training and testing sequences. Each of the taggers and each tagger ensemble has been trained and tested on the same set of cross-validation folds, so the results are directly comparable.

Each of the training folds has been reanalyzed, according to the procedure described in (Radziszewski, 2013), using the Maca toolkit (Radziszewski and Śniatowski, 2011b). The idea of a morphological reanalysis of the gold-standard data is to allow the trained tagger to see similar input that is expected in the tagging phase. The training data is firstly turned into plain text and analyzed using the same mechanism that will be used by the tagger during actual tagging process. The output of the analyzer is then synchronized with the original gold-standard data, by using the original tokenization. Tokens with changed segmentation are taken from the gold-standard intact. In the case of tokens, for which the segmentation did not change in the process of morphological analysis, the produced interpretations are compared with the original. A token is marked as an unknown word, if the correct interpretation has not been produced by the analyzer. This is to mimic a real-world scenario of tagging plain text with the chosen morphological analyzer.

In our experiments, Maca has been run with the `morfeusz-nkjp-official` configuration, which uses Morfeusz SGJP analyzer (Woliński, 2006) and no guesser module.

4. State-of-the-Art Polish Taggers

To enable a fair comparison of tagging results, we have firstly evaluated each of the individual taggers on the same data, which has been then used for further experiments concerning tagger ensembles.

Pantera (Acedański, 2010) is an adaptation of the Brill's algorithm to morphologically rich languages, such as Polish. Pantera includes several techniques of improving the tagging of inflectional languages, such as multi-pass tagging and transformation templates. In the experiments, we have used the learning threshold value of 6, as recommended by the author.

WMBT (Radziszewski and Śniatowski, 2011a) is a memory based tagger, which disambiguates the set of possible tags in multiple tiers. The number of tiers is equal to the number of attributes in the tagset, including the grammatical class. Tokens in each of the individual tiers are classified using a k-Nearest Neighbors classifier. We have used the supplied `nkjp-guess.ini` configuration file when using the tagger.

WCRFT (Radziszewski, 2013) is a tiered tagger, based on Conditional Random Fields (CRF), a mathematical model similar to Hidden Markov Models. A separate CRF model is used to disambiguate distinct grammatical attributes. The `nkjp_s2.ini` configuration has been used during evaluation.

Concraft (Waszczuk, 2012) is another approach to adaptation of CRFs to the problem of POS tagging. In Concraft, the CRF layers are mutually dependent and the results of disambiguation from one of the layers may propagate to another. The first of the two layers used by the tagger includes tags related to POS, case and person, while the second contains all other grammatical categories.

n	Tagger	Acc_{lower}	Acc^K_{lower}	Acc^U_{lower}
1	Pantera	88.95%	91.22%	15.19%
2	WMBT	90.33%	91.26%	60.25%
3	WCRFT	90.76%	91.92%	53.18%
4	Concraft	91.07%	92.06%	58.81%

Table 1: The accuracy of individual state-of-the art POS taggers for Polish (evaluated on the NCP1M corpus, ten-fold cross-validation).

The accuracy of tagging for each of the methods, as evaluated using the scheme described in the previous section, has been presented in Table 1. They are generally on par with previously published results. There is a statistically significant difference between the accuracies of the taggers, with a significance level of 0.05.

5. Combination of Taggers

5.1. Tagger Complementarity

Following the approach proposed by Brill and Wu (1998), we have evaluated the relative differences between the sets of errors made by the individual taggers. The tagger complementarity $Comp(A, B)$ measures how different the mistakes made by the two taggers A and B are. The calculated value is the percentage of time when tagger A is wrong that tagger B is correct:

$$Comp(A, B) = (1 - \frac{e_{AB}}{e_A}) * 100,$$

where e_{AB} is the number of common errors, both in A and B, while e_A is the number of errors made by tagger A. The results for the used taggers, as evaluated on the NCP1M corpus, are presented in Table 2.

As the results in Table 2 suggest, there is a large overlap in the sets of errors made by the taggers (all values are below 50%, while for completely independent taggers the value would be 100%). There is however still hope of

A \ B	Pantera	WMBT	WCRFT	Concraft
Pantera	0.00%	42.33%	42.16%	45.22%
WMBT	34.09%	0.00%	35.30%	39.52%
WCRFT	30.78%	32.25%	0.00%	33.97%
Concraft	32.21%	34.52%	31.72%	0.00%

Table 2: Tagger complementarity.

achieving a lower rate of mistakes, especially in the case of ensembles containing the Concraft tagger, for which the complementarity values are the highest.

5.2. Theoretical Bounds

A theoretical upper bound of the expected accuracy of an ensemble of classifiers may be calculated as the number of times all taggers make a mistake, while tagging the test dataset. Even if only one of the classifiers provides the correct answer, there is a possibility of developing a tagger selection method, which is able to accurately distinguish between correct and wrong tagger decisions. The accuracy of such an "Oracle" for each of the possible tagger ensembles has been presented in Table 3 and Figure 1. It is worth noting that such a theoretical upper bound rises to 95.82% in the case of an ensemble constructed on the basis of all the evaluated taggers.

The accuracy of the best individual tagger, which is 91.07% for Concraft, is the natural lower bound, below which creating an ensemble is pointless. With regard to tagger selection strategy, we may think of another lower bound, such as a random selection strategy, over which a more elaborate approach should have a significant advantage.

5.3. Evaluation of Tagger Combinations

We have evaluated the accuracy of each possible tagger ensemble, consisting of 2, 3 and 4 classifiers. Each of the methods has been used in exactly the same setup and using the same parameters, as during the individual evaluation. Taggers have been previously trained on the same set of 10 training folds, which have been prepared according to the procedure described in Section 3. Tagging has been performed on test folds, which have been created by turning the original gold-standard data into plain text and performing morphological analysis using the Maca framework and Morfeusz SGJP analyzer.

The first evaluated approach to ensemble decision selection was a simple voting strategy. In this case the decision of each of the taggers is equally weighted and the tag produced by the highest number of taggers is selected as the most probable. In the case of ties, the first tagger in the ensemble is selected as the winner.

In the second scenario, we have weighted the individual taggers with weights equal to their individual accuracy, as tested in the preliminary experiments, presented in Table 1. This is essentially making the best individual tagger win in the case of tie. The results of both experiments are presented in Table 3 and Figure 1.

All tagger combinations consisting of at least three methods proved to produce better quality tagging results than the best single tagger. In some cases the differences in accuracy between three–tagger combinations are very slight, but the inclusion of the best performing tagger (Concraft) clearly improves overall ensemble accuracy and the difference is statistically significant. It is interesting to note that the best performing combination, consisting of all available taggers, improves not only the overall accuracy over the Concraft tagger, but also the quality of tagging unknown words over the WMBT tagger, individually performing best in that field.

The difference in results between weighted and simple voting strategies turned out not to be statistically interesting, as in most cases the disagreements between taggers have been resolved by a simple majority vote. There have been only a few "true disagreements", in which the number of possible options has been greater than two. The difficulty of creating an ensemble of taggers lies in the ability to select correctly the (often outvoted) single tagger that in a particular context produces the correct POS tag.

6. Conclusions

We have shown that even a simple combination of Polish taggers by decision voting may lead to an improvement of the tagging accuracy and, as the result, the quality of the produced morphosyntactic annotation of natural language text. We have made a re-evaluation of state-of-the-art Polish taggers, using the most current and largest training data and the recommended evaluation procedure. In the direct comparison of all the possible tagger combination configurations and the individual taggers, it is clear that using more than one tagger produces statistically significant improvement in tagging accuracy.

The theoretical upper bound of the performance of such an ensemble (95.82%) brings us closer to the results reported for English, which are still above 97%. The investigation of other, more sophisticated strategies of classifier selection, most probably based on machine learning techniques, remains for further work. Another direction of possible research is the inclusion of even greater number of taggers, regardless of their individual accuracy, to incorporate a greater variation in the set of answers produced by the ensemble.

Acknowledgements

This work has been funded by the National Science Centre project number DEC-2011/01/N/ST6/01107.

The author would like to thank Adam Radziszewski for sharing valuable information concerning the evaluation of taggers.

References

Acedański, Szymon, 2010. A morphosyntactic Brill tagger for inflectional languages. In *Advances in Natural Language Processing*.

Brill, Eric and Jun Wu, 1998. Classifier combination for improved lexical disambiguation. In *Proceedings of the 17th international conference on Computational linguistics - Volume 1*, COLING '98. Stroudsburg, PA, USA: Association for Computational Linguistics.

Strategy:	Random	Simple Voting			Weighted Voting			Oracle
Taggers	Acc_{lower}	Acc_{lower}	Acc^{K}_{lower}	Acc^{U}_{lower}	Acc_{lower}	Acc^{K}_{lower}	Acc^{U}_{lower}	Acc_{lower}
1+2+3	90.01%	91.46%	92.57%	55.19%	91.53%	92.60%	56.55%	94.98%
1+2+4	90.14%	91.67%	92.74%	56.99%	91.81%	92.81%	59.45%	95.25%
1+3+4	90.30%	91.69%	92.85%	54.02%	91.88%	92.89%	59.08%	95.09%
2+3+4	90.71%	91.86%	92.78%	62.07%	91.90%	92.82%	62.31%	95.15%
1+2+3+4	90.30%	91.95%	92.87%	62.18%	**92.01%**	**92.91%**	**62.81%**	95.82%

Table 3: The accuracy of tagger combinations. Taggers are identified by numbers, as given in Table 1.

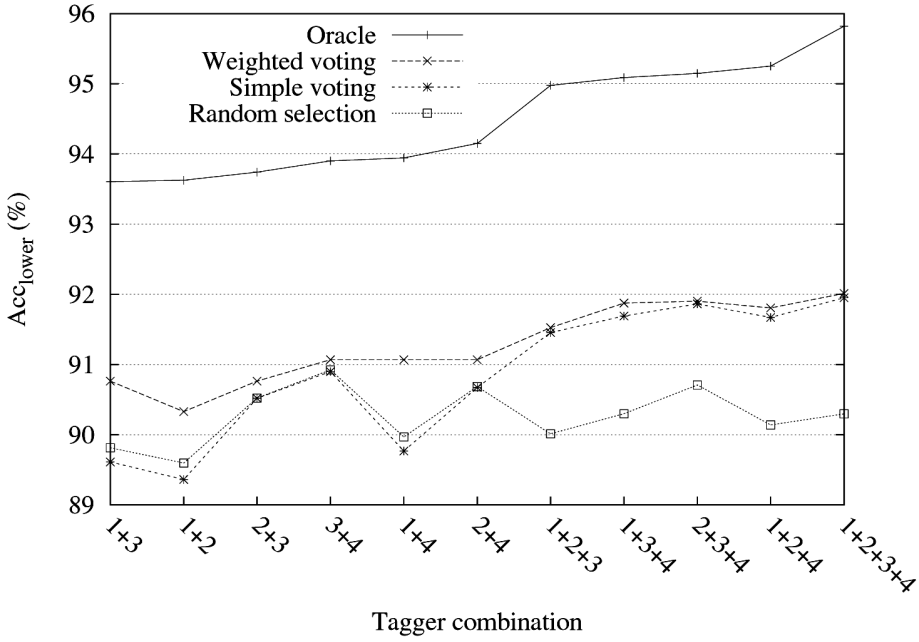

Figure 1: The accuracy of tagger combinations. Oracle: theoretical upper bound. Random selection: a lower bound for evaluating tagger combinations. Taggers are identified by numbers, as given in Table 1.

Śniatowski, Tomasz and Maciej Piasecki, 2012. Combining Polish morphosyntactic taggers. In Pascal Bouvry, Mieczysław A. Kłopotek, Franck Leprévost, Małgorzata Marciniak, Agnieszka Mykowiecka, and Henryk Rybiński (eds.), *Security and Intelligent Information Systems*, volume 7053 of *LNCS*. Springer-Verlag.

Przepiórkowski, Adam, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk (eds.), 2012. *Narodowy Korpus Języka Polskiego*. Warsaw: Wydawnictwo Naukowe PWN.

Radziszewski, Adam, 2013. A tiered CRF tagger for Polish. In R. Bembenik, Ł. Skonieczny, H. Rybiński, M. Kryszkiewicz, and M. Niezgódka (eds.), *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*. Springer Verlag.

Radziszewski, Adam and Szymon Acedański, 2012. Taggers gonna tag: an argument against evaluating disambiguation capacities of morphosyntactic taggers. In *Proceedings of TSD 2012*, LNCS. Springer-Verlag.

Radziszewski, Adam and Tomasz Śniatowski, 2011a. A Memory-Based Tagger for Polish. In *Proceedings of the LTC 2011*.

Radziszewski, Adam and Tomasz Śniatowski, 2011b. Maca — a configurable tool to integrate Polish morphological data. In *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*.

van Halteren, Hans, Walter Daelemans, and Jakub Zavrel, 2001. Improving accuracy in word class tagging through the combination of machine learning systems. *Comput. Linguist.*, 27(2):199–229.

Waszczuk, Jakub, 2012. Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*. Mumbai, India.

Woliński, Marcin, 2006. Morfeusz — a practical tool for the morphological analysis of Polish. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski (eds.), *Intelligent Information Processing and Web Mining*, Advances in Soft Computing. Berlin: Springer-Verlag.