Taggers gonna tag: an argument against evaluating disambiguation capacities of morphosyntactic taggers*

Adam Radziszewski[1] and Szymon Acedański[2]

[1] Institute of Informatics, Wrocław University of Technology
[2] Institute of Computer Science, Polish Academy of Sciences

Abstract Usually tagging of inflectional languages is performed in two stages: morphological analysis and morphosyntactic disambiguation. A number of papers have been published where the evaluation is limited to the second part, without asking the question of what a tagger is supposed to do. In this article we highlight this important question and discuss possible answers. We also argue that a fair evaluation requires assessment of the whole system, which is very rarely the case in the literature. Finally we show results of the full evaluation of three Polish morphosyntactic taggers. The discrepancy between our results and those published earlier is striking, showing that these issues do make a practical difference.

Key words: morphosyntactic tagging, morphological analysis, tokenisation, evaluation

1 Introduction

Part-of-speech (POS) tagging is a well-researched Natural Language Processing (NLP) task. Taggers assign POS tags to words and word-like units (*tokens*) in text. In languages with rich inflection the tags usually include significantly more information that just parts-of-speech, e.g. nouns may be specified for values of number, gender and case, adverbs may be specified for degree. In such a setting, the tags are often called morphosyntactic tags and the task is referred to as morphosyntactic tagging.

It has been noted that the tagging accuracy has significant impact on performance of other NLP tasks, such as parsing (Hajič et al. 2001). A reliable tagger evaluation procedure is therefore vital for unbiased selection of the best tagger for a particular application. What is more, using an inaccurate evaluation procedure may bring about long-standing consequences: as long as the generally used procedure neglects pratical aspects related to actual tagger mispredictions, those shortcomings are unlikely to be overcome. In this paper we show that the latter is often the case. We also offer a simple alternative that allows avoiding unjustified bias. Our discussions are grounded on Polish background, although

* Work financed by Innovative Economy Programme, POIG.01.01.02-14-013/09.

the observations are also applicable at least to other inflectional languages. We also perform a new evaluation of two state-of-the art Polish taggers along the lines of the proposed methodology and present our results, which are strikingly different from those previously published.

2 Common practice

A common practice in tagging of inflectional languages is to decompose the process into two stages:

1. morphological analysis (dictionary look-up), resulting in sets of possible tags assigned to each token;
2. morphosyntactic disambiguation, that is ruling out of contextually inappropriate tags assigned during the previous stage.

This practice has become so common[3] that it has already started to influence the perception of what the task of a morphosyntactic tagger is. This consideration has important implications for tagger evaluation: depending on which stage of an evaluated tagger we take as a black box, we will assess different parts of the whole system and get different results. The problem is that the very question of what a tagger is supposed to do is almost never asked, while different answers are implicitly assumed, rendering comparisons of published results impossible. There are at least three possible answers, all of them at least occasionally assumed:

1. The tagger tags running (plain) text, that is a sequence of characters; assumed in Karwańska and Przepiórkowski (2010).
2. The tagger tags a sequence of unlabelled tokens, possibly given sentence boundaries; assumed in Schmid and Laws (2008), Daelemans et al. (2010), Acedański and Przepiórkowski (2010).
3. The tagger's task is only that of morphosyntactic disambiguation; assumed in Piasecki (2007), Acedański (2010), Śniatowski and Piasecki (2011), Radziszewski and Śniatowski (2011), Hajič and Vidová-Hladká (1998).

The second approach assumes that the tokenisation performed by the tagger is perfect (tokenisation is taken from the reference corpus). The third approach is most controversial, since it neglects both tokenisation errors and deficiencies of morphological analysis. In spite of that, this is actually the most popular approach, at least to evaluation of Polish taggers.

3 Proposed methodology

We argue that taggers are best evaluated on plain text, while the two other approaches (as outlined in the previous section) are biased and should be avoided:

[3] E.g. Hajič and Vidová-Hladká (1998) state that "given the nature of inflectional languages (...) it is necessary to employ morphological analysis before the tagging proper".

1. In a typical scenario, the user has access to plain text and is interested in obtaining reasonable tokenisation and accurate labelling of those tokens. No reference morphological analysis is normally available.
2. One of the symptoms is that separate figures for tagging known and unknown words are not reported in the Polish literature, which is otherwise a common practice. This is because in such a setting there are virtually no unknown words — the reference tag is always there to be chosen. This may lead to an absurd situation, when proper tagging of out-of-vocabulary words is the easiest task for the black box[4].
3. Such an approach makes it impossible to assess the impact of different morphological analysers on the overall tagging performance. This is a serious consideration: multiple analysers are available, while the two-stage implementation of tagging facilitates integration with different analysers. At the same time, the choice of morphological analyser has an obvious impact on the final tagging accurracy. Hence, publishing results of only the disambiguation part silently ignores the influence of the choice and quality of the analyser.
4. Such evaluation procedures show only differences in disambiguation strategy, while the tagger's guessing capabilities are not assessed. This discourages the development of better strategies for handling unknown words.

It is also worth noting that a fair evaluation has already been performed for two Polish taggers (Karwańska and Przepiórkowski 2010), although not mentioned[5]. Note that the following publications involving evaluation of Polish taggers continue to use the 'old' disambiguation-based approach: Acedański (2010), Śniatowski and Piasecki (2011), Radziszewski and Śniatowski (2011). A similar situation happened for tagging Czech: Hajič (2000) states in a footnote on p. 3 that they had "been simply ignoring the unknown word problem altogether in the past". We hope that this paper will make these issues explicit and encourage fair tagger evaluation in the future.

3.1 Recommendations

We recommend the following methodology, which we also used for experiments reported in the next section:

1. Ten-fold cross-validation is employed, using a manually annotated corpus (gold standard). We split on the basis of paragraphs to account for taggers which may use context information that crosses sentence boundaries.
2. The main statistic calculated we call *accuracy lower bound*. It is the measure we advocate for general-purpose tagger evaluation, where all discrepancies in

[4] Such a situation indeed occurs in Acedański (2010), Śniatowski and Piasecki (2011), Radziszewski and Śniatowski (2011) as the corpus employed for evaluation assigns exactly two possible tags for all the unknown words: the proper tag and a special *out-of-vocabulary* tag (Przepiórkowski and Murzynowski 2010) — a sufficient winning strategy is to never choose the *out-of-vocabulary* tag.

[5] It was confirmed by Danuta Karwańska (personal communication, 6 October 2011).

segmentation are penalised. We expect the segmentation to match the gold standard exactly to promote authors to create consistent segmenters instead of tweaking evaluation methods to match particular taggers. We calculate this statistic as a percentage of tokens in the gold standard which have a lexically matching segment and it is correctly tagged.

3. We also calculate an additional statistic designed to show the possible influence of segmentation errors on tagging quality. *Accuracy upper bound* is a hypothetical upper limit of tagger performance, treating all the tokens subjected to segmentation changes as correctly tagged. It is a percentage of gold standard segments, which either have a lexically corresponding segment with the correct tag, or have no lexically corresponding segment.

4 Experiments for Polish taggers

It is worth noting that in Polish, tags are composed of a part-of-speech label and a number of labels for grammatical categories. For example, `subst:sg:nom:f` is a feminine, singular noun (substantive) in the nominative case. Some attributes are defined as optional and during evaluation we expand them to multiple tags, following recommendation in Karwańska and Przepiórkowski (2010). As the gold standard we used the published 1 million token manually annotated subcorpus of the National Corpus of Polish, version 1.0 (Przepiórkowski et al. 2010).

4.1 PANTERA

PANTERA (Acedański 2010) is a morphosyntactic tagger based on Brill's Algorithm (Brill 1992) adapted for morphologically rich languages, targeted at Polish. The tagger automatically generates limited-context rules which are then applied in order to the text, pre-tagged using an unigram tagger. PANTERA is also a 2-tier tagger, which first disambiguates the part of speech, case and person, and then the rest of grammatical categories. In our experiments we train PANTERA with the threshold rule quality of 6, using morphological analyser Morfeusz[6] (Woliński 2006) with the TaKIPI guesser module (Piasecki 2007) enabled. The authors report 92.68% accuracy on the NCP.

4.2 WMBT

WMBT (Radziszewski and Śniatowski 2011) is a simple memory-based tagger that operates on as many tiers as there are attributes in the tagset. WMBT itself is actually a disambiguation engine — it should be run after performing tokenisation and morphological analysis. The authors recommend using Maca (Radziszewski and Śniatowski 2011) software for that purpose. Thus, the tests presented here were performed this way, using a Maca configuration recommended for this scenario, namely `morfeusz-nkjp-official-guesser`. The authors report 92.98% accuracy on the NCP.

[6] For all the experiments described in this paper we use a 64-bit version of Morfeusz SGJP, ver. 0.82 (code timestamp 22/02/2010, linguistic data from 15/04/2011).

4.3 MBT

To make the comparison more insightful, we decided to also include a tagger which is not targeted specifically at Polish. MBT (Daelemans et al. 2010) is a generic memory-based tagger that may be used for various languages. MBT is trained with a corpus that must contain a sequence of tokens and their corresponding tags. Optionally, external features may also be included. A particularly interesting feature of MBT is that it creates two models: one for known words and one for unknown words (the latter one is obtained by analysing forms that were infrequent in the training data).

A drawback of MBT when applied to inflectional languages is that it treats feature values atomically, hence it cannot reason using the attribute values inferred from tags. This can be altered by introducing additional features directly to the input before running MBT. We decided to exploit this possibility by testing three variants:

1. A *naive* variant (*Features 0*): each input token is described by its wordform.
2. Simplified feature set (*Features 1*): each token is described by its wordform, but also possible part-of-speech labels and values of three gramamtical categories: number, gender and case taken from a window $(-3, \ldots, +2)$ surrounding the token.
3. Rich feature set (*Features 2*) is exactly the same feature set as used internally by WMBT, that is *Features 2* extended with some tests for morphosyntactic agreement (Radziszewski and Śniatowski 2011), generated using the WCCL toolkit (Radziszewski et al. 2011).

4.4 Evaluation results

We performed two experiments:

1. Evaluation of disambiguation capabilities only (Acc_{dis}). The morphological data and tokenisation were taken directly from the reference corpus.
2. Evaluation according to our proposal: the testing material was turned to plain text and the taggers were to produce valid output. We report the values of accuracy lower and upper bound (Acc_{lower}, Acc_{upper}), but also, accuracy lower bound measured separately for words known and unknown to the morphological analyser (Acc^{K}_{lower} and Acc^{U}_{lower}, respectively).

The results are presented in Table 1. The difference between accuracy lower and upper bounds is not very substantial. In the following discussion we focus on the lower bounds then.

The figures for disambiguation accuracy of PANTERA and WMBT are in line with previous publications. But what is striking here is the gap between these figures and the accuracy of the full tagging process (that is, accuracy lower bound). The gap apparently stems from the fact that the disambiguation accuracy neglects the errors made during morphological analysis and tokenisation. Interestingly, the errors neglected make up almost a half of the total tagging

Tagger	Acc_{dis}	Acc_{lower}	Acc_{upper}	Acc^{K}_{lower}	Acc^{U}_{lower}
PANTERA	92.95%	88.79%	89.09%	91.08%	14.70%
WMBT	93.00%	87.50%	87.82%	89.78%	13.57%
MBT: Features 0	79.31%	79.11%	79.44%	80.30%	40.49%
MBT: Features 1	88.03%	84.14%	84.46%	85.79%	30.74%
MBT: Features 2	87.12%	83.39%	83.72%	85.00%	31.36%

Table 1. Accuracy measures obtained using the proposed methodology.

error. What is more, PANTERA turns out to outperform WMBT, while the opposite seems to hold if observing disambiguation accuracy alone.

As one could expect, both measures do converge for MBT. This is because MBT is by design not given access to external morphological analyser, hence it is also unable to take advantage of the reference morphological analysis available in the reference corpus. This is another argument against evaluation based on the reference morphological analysis: the taggers that assume two-stage operation are able to peek at the reference annotation and win undeserved points.

While the overall tagging accuracy of MBT is not impressive, a couple of interesting observations could be made. First, the introduction of additional features brought substantial improvement. This confirms our presumption that for best results in tagging inflectional languages, values of grammatical categories should be represented as separate features. Even then MBT is still bound to output whole tags during one run; this is a likely explanation for its performance being much lower than that of WMBT, a tiered tagger. On the other hand, the figures recorded for tagging unknown words are much higher in the case of MBT than those achieved by the two state-of-the-art taggers made specifically for Polish: the best result for MBT is 40.5%, while the figures reported for the Polish taggers are lower than 15%. This is probably due to a separate module for tagging unknown words in MBT. The lower accuracy of other taggers could also be attributed to the prevalence of evaluation based on disambiguation capabilities in Polish NLP community: a problem unnoticed is likely to remain unsolved.

5 Conclusion

In this paper we pointed out and discussed a number of choices which need to be made consciously to ensure comparable evaluation of taggers. We also re-evaluated the taggers described in published papers, now in a comparable way. This brought significantly different results. Therefore we proposed a general-purpose tagger evaluation methodology. By providing a simple enough method, we promote publishing comparable and useful tagging performance values.

Bibliography

Acedański, S. (2010). A morphosyntactic Brill tagger for inflectional languages. In Loftsson, H., Rögnvaldsson, E., and Helgadóttir, S., editors, *Advances in Natural Language Processing: Proceedings of the 7th International Conference on Natural Language Processing, IceTAL 2010, Reykjavík, Iceland*, volume 6233 of *Lecture Notes in Artificial Intelligence*, pages 3–14, Heidelberg. Springer-Verlag.

Acedański, S. and Przepiórkowski, A. (2010). Towards the adequate evaluation of morphosyntactic taggers. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010); Poster Session*, pages 1–8, Beijing.

Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 152–155, Morristown, NJ, USA. Association for Computational Linguistics.

Daelemans, W., Zavrel, J., Van den Bosch, A., and van der Sloot, K. (2010). MBT: Memory-Based Tagger, version 3.2. Technical Report 10-04, ILK.

Goźdź-Roszkowski, S., editor (2010). *The proceedings of Practical Applications in Language and Computers PALC 2009*, Frankfurt am Main. Peter Lang.

Hajič, J. (2000). Morphological tagging: Data vs. dictionaries. In *Proceedings of the 6th Applied Natural Language Processing and the 1st NAACL Conference*, pages 94–101.

Hajič, J., Krbec, P., Květoň, P., Oliva, K., and Petkevič, V. (2001). Serial combination of rules and statistics: A case study in Czech tagging. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 268–275. Association for Computational Linguistics.

Hajič, J. and Vidová-Hladká, B. (1998). Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. In *Proceedings of the COLING - ACL Conference*, pages 483–490. ACL.

Karwańska, D. and Przepiórkowski, A. (2010). On the evaluation of two Polish taggers. In Goźdź-Roszkowski (2010).

Piasecki, M. (2007). Polish tagger TaKIPI: Rule based construction and optimisation. *Task Quarterly*, 11(1–2):151–167.

Przepiórkowski, A., Górski, R. L., Łaziński, M., and Pęzik, P. (2010). Recent developments in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta. ELRA.

Przepiórkowski, A. and Murzynowski, G. (2010). Manual annotation of the National Corpus of Polish with Anotatornia. In Goźdź-Roszkowski (2010).

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

Radziszewski, A. and Śniatowski, T. (2011). A memory-based tagger for Polish. In *Proceedings of the 5th Language & Technology Conference, Poznań*.

Radziszewski, A., Wardyński, A., and Śniatowski, T. (2011). WCCL: A morpho-syntactic feature toolkit. In *Proceedings of the Balto-Slavonic Natural Language Processing Workshop*. Springer.

Schmid, H. and Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of COLING 2008*, volume 1, pages 777–784. Association for Computational Linguistics.

Śniatowski, T. and Piasecki, M. (2011). Combining Polish Morphosyntactic Taggers. In *Proceedings of the 2011 International Joint Conference on Security and Intelligent Information Systems*. Springer Berlin / Heidelberg.

Woliński, M. (2006). Morfeusz — a Practical Tool for the Morphological Analysis of Polish. In *Intelligent Information Processing and Web Mining*, pages 511–520.