Tagger Voting Improves Morphosyntactic Tagging Accuracy on Croatian Texts

Željko Agić[1], Marko Tadić[2], Zdravko Dovedan[1]
[1]*Department of Information Sciences*
[2]*Department of Linguistics*
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb
{zeljko.agic, marko.tadic, zdravko.dovedan}@ffzg.hr

Abstract. *We present results of an experiment dealing with combining outputs of five part-of-speech taggers via tagger voting in order to improve the overall accuracy of morphosyntactic tagging of Croatian texts using a subset of the Multext-East v3 tagset. The increase in accuracy over the best-performing single tagger is shown to exist, but not to be statistically significant. We discuss the performance of the five single taggers, the overlaps between tagger pairs, the reduced tagset and the voting scheme, along with scores for five meaningful tagger combinations in the voting scheme and future work plans.*

Keywords. Tagger voting, morphosyntactic tagging, Croatian language

1. Introduction

In the paradigm of stochastic part-of-speech or morphosyntactic tagging, the performance of taggers on texts of a certain language is shown (cf. e.g. [2]) to be proportional with the size of the morphosyntactically annotated corpora on which their language models are trained and inversely proportional with the size of the tagset that encodes morphological features in corpora of a given language. A more detailed elaboration of this topic from a perspective of tagging Croatian texts is given in the introductory section of [4]. Basically, given the available corpora and appropriate tagset for a certain language, various stochastic taggers will reach certain peak tagging accuracies on unseen texts, from which it will be hard to progress in terms of further increase in accuracy. For current state-of-the-art taggers, this peak accuracy will be placed somewhere between (cf. e.g. [22]) 96.46 and 97.33 percent correctly assigned tags when tagging English, with a ca 10-15 percent decrease when tagging other, morphologically more complex languages, i.e. languages for which the available corpora are annotated using a significantly larger tagset, such as Croatian or Czech (cf. [1] and [15]). Keeping in mind the issues of large tagsets and small available corpora of languages as Croatian, it becomes clear why even a basic natural language processing task, such as part-of-speech tagging, still poses a challenge.

Three basic, somewhat disjoint approaches to raising the performance level of part-of-speech taggers can be found in the literature today (again, cf. [4] for a detailed discussion):

- hybridizing the stochastic taggers by introducing them with rule-based and language-specific modules or resources,
- manipulating the language models of these taggers, e.g. by dynamically shrinking and expanding the tagset at tagger runtime and
- combining outputs of different taggers and tagging paradigms via tag voting schemes and meta-taggers.

An example of the first approach, i.e. tagger hybridization, is an experiment [1] in tagging Croatian texts by introducing an available Croatian lexical resource – namely the Croatian morphological lexicon [17] [18] – to a second order hidden Markov model tagger as a module for handling unknown word forms at tagger runtime. Introducing lexical resources resulted in raising the overall tagger accuracy and is also present in this paper as the CroTag tagger.

Tiered tagging or tier-tagging [7] [20] [21] is a notable approach for the method of manipulating with the language model of (not exclusively) stochastic taggers. It implements a lossless algorithm that maps the existing tagset used in annotation of the training corpus into a smaller hidden tagset, which is then used to do the actual tagging. After the tagging is completed, the annotations from the hidden tagset are expanded into actual tags and presented to the user as the output of the tagging procedure. Another approach that falls within this group is the one introducing lossy tagset reductions with regards to applications of tagging in other, more complex natural language processing systems. Namely, in this approach, the full set of morphosyntactic tags is elaborately reduced in a meaningful way by expert linguists, excluding the morphological features that are not required in the resulting larger scale system.

The third approach, dealing with improving the accuracy of morphosyntactic tagging by combining taggers [5] [11] [14], is the focus of the experiment presented here. For purposes of this paper, we perceive two different approaches to combining taggers and targeting higher performance levels: (1) tagger voting and (2) tagging by classifying or meta-tagging. Both approaches stem from the same underlying idea: annotating the text by a number of different taggers and merging the provided annotations by a certain merge strategy. It is the specific strategy of merging the output tags provided by the taggers into a single output tag that differentiates the two approaches. In the first approach, a specifically designed, most often rule-based post-processing module is used to choose a single output tag. One of the most straightforward approaches is tagger voting, in which a certain, preferably odd number of taggers is run on the input and the output is voted on, choosing the output tag on basis of the votes. In this paradigm, taggers are considered as voters and tags as candidate outputs – numbers of votes are assigned to each of the output tags and the tag having the highest number of votes (i.e. the one which the majority of taggers outputted the most) is chosen as the final output. Other approaches implement more sophisticated voting or disambiguating strategies (cf. [5]) that rely on observing actual outputs in terms of conducting qualitative analyses and creating disambiguation modules according to their results. Finally, meta-tagging by classifying is an approach where the choice of specific voting or disambiguation strategy is left to machine learning algorithms rather than manual analysis. Namely, e.g. in [11] and [14] for tagging Slovene and Swedish, machine classifiers are trained on held-out data to automatically choose between taggers on basis of sentence (tag, word form) context. These pre-trained models are then used to disambiguate the output of multiple taggers.

In this contribution, we present results of utilizing the before-mentioned straightforward approach to tagger voting to tagging Croatian texts. Five morphosyntactic taggers with three distinct underlying paradigms are chosen and used to tag Croatian texts. Their single outputs are evaluated, combined and then disambiguated using the simple voting scheme. The following sections of the paper present the experiment plan and obtained results, followed by a discussion and insight on future work prospects.

2. Experiment setup

By default, an experiment with part-of-speech tagging requires a manually annotated and ten-folded gold standard corpus and a tagger or, in this case, taggers. As in previous experiments with morphosyntactic tagging of Croatian (cf. [1] to [4]), the CW100 newspaper corpus was also used in this experiment. Detailed description of the corpus can be found in [1] to [4], while table 1 provides only a short overview. The corpus is split into ten different parts, equal in number of sentences contained. Nine parts are used for creating the language model for the tagger and the tenth is always used for validating that model. All counts and results are tenfold cross-validated. In this specific experiment, we used a tagset reduction of the full Multext-East v3 [8] morphosyntactic tagset in order to reduce the training time overhead for certain taggers' model-building procedures. The reduction itself is linguistically founded and is exposed in detail in [4]. As shown in table 1, the reduction reduced the number of tags from 879.60 different tags on average in the CW100 training sets to 48.00 tags in the training sets used in this experiment and from 473.20 to 41.80 tags in the testing sets.

Table 1. Overview of corpus subsets (average)

Set	Tokens	Unique	Tags	Reduced
Train	106676.10	23426.40	879.60	48.00
Test	11852.90	4638.60	473.20	41.80

As the purpose of this specific experiment was to show whether or not voting improves tagging accuracy of Croatian texts, considering how tagset size influences only the learning rates and peak accuracy of the taggers, we argue that the observations made with the reduced tagset are equally valid with regards to the full Multext-East v3 tagset for Croatian, but keeping in mind the loss of information encoded in the features dropped from the tagset by the reduction.

Five morphosyntactic taggers were used in the experiment. Three of them were second order hidden Markov model taggers – CroTag [1], HunPos [10], TnT [6] – while SVMTool [9] is based on support vector machines (SVMs) and TreeTagger [12] [13] uses decision trees. The taggers were chosen simply by overlapping these criteria: (1) speed of training and annotation, (2) previously documented performance on Croatian and English texts and (3) notable differences in underlying paradigms. We intuitively considered the third point to be of particular importance, as we expected that the taggers implementing different tagging paradigms, e.g. HMM vs. SVM, would all reach satisfactory tagging accuracies, while disagreeing about tags more frequently than taggers implementing the same tagging paradigm would. We expected that the voting scheme would benefit from such a combination of good accuracy scores and high disagreement about tags for specific words. It should be noted here that the tagset reduction compromise was made with respect to the speed of the training procedure of the SVM tagger, while some taggers were rejected entirely (see the acknowledgements section), either because of these three requirements or because of demands on input and output format of the corpus.

The experiment itself was conducted in the following way. First each of the taggers was evaluated on the tenfold-sets. Afterwards, their outputs on each of the sets were merged into group outputs that were then also evaluated as an indicator of performance for the voting tagger that consists out of all five taggers. Finally, four other voting taggers were defined by combining the single taggers in groups of three: (1) CroTag, HunPos and TnT, (2) CroTag, SVMTool and TreeTagger, (3) CroTag, HunPos and SVMTool, (4) CroTag, HunPos and TreeTagger. These tagger assemblies were also evaluated in the same testing scenario. The results are provided in the following section.

3. Results and discussion

The presentation of the obtained results starts with an insight on accuracy of single taggers. Beside this information being a logical starting point for discussion on results, it is also interesting to observe it on its own as this is also, at least to our knowledge, the first evaluation of taggers based on support vector machines (SVMTool) and decision trees (TreeTagger) in the task of tagging Croatian texts. However, it should be noted that detailed investigation into properties of these tagging paradigms with respect to specifics of Croatian was not conducted in this experiment, as our focus was on combining these taggers via the simple voting scheme. Scores are given in table 2, presenting overall average tagging accuracy on the testing sets, followed by scores on tagging known and unknown word forms, i.e. word forms that were or were not encountered while training the taggers. All the averaged scores are followed by corresponding 95-percent confidence intervals.

Table 2. Accuracy of single taggers

	Overall	Known	Unknown
CroTag	**90.35**±0.52	94.14±0.35	**71.09**±1.43
HunPos	90.06±0.52	94.03±0.36	69.92±1.19
SVMTool	89.79±0.54	**94.19**±0.35	67.66±1.51
TnT	90.30±0.53	94.16±0.33	70.68±1.39
TreeTagger	88.31±0.43	93.47±0.27	62.12±0.95

Expectedly, the single highest scoring tagger in the experiment was CroTag, as it was run in hybrid mode, i.e. utilizing the information from an inflectional lexicon of Croatian for easier handling of unknown word forms. This is clearly shown in the column containing scores on unknown tokens, where CroTag outperforms all the other taggers, while SVMTool and TnT in turn outperform CroTag in tagging known word forms. All three highest performing taggers in this task were hidden Markov model taggers and it should be noted, as indicated by the confidence intervals and verified using the two-tailed t-test, the difference between them was not statistically significant. In addition, the difference between the three hidden Markov model taggers and the SVMTool tagger was also not statistically significant. However, the difference in scores between these four taggers and the TreeTagger is shown to be statistically significant by the two-tailed t-test. It should once again be noted that none of the taggers, CroTag excluded, were additionally fine-tuned for this task.

Table 3. Agreement between taggers

	HunPos	SVMT	TnT	TreeTagger
CroTag	95.24	92.84	97.22	91.67
HunPos	/	92.50	97.10	90.90
SVMT	/	/	92.95	90.26
TnT	/	/	/	91.79

In choosing tagger assemblies, i.e. defining voting taggers for the experiment, as presented in the previous section, we followed two sets of results: overall accuracy from table 2 and tagger agreement given in table 3. The tagger agreement or overlap results have expectedly proven the intuition. Overlaps are large between the three HMM taggers and 3-6 percent smaller when comparing the HMM group to SVMTool and TreeTagger. The overlap of the latter two taggers was also smaller than the one between the HMMs. Inferring from the comparable overall accuracies of specific taggers presented in table 2 and sizes of their overlaps from table 3, intuition would also suggest which voting scheme would yield an improvement in overall tagging accuracy. Table 4 provides actual data to verify that intuition. Once again, the scores are given with 95-percent confidence intervals.

Table 4. Accuracy of tagger voting combinations

	Overall	Known	Unknown
All taggers	90.75±0.51	94.28±0.33	**72.83**±1.40
CT+HP+TnT	90.37±0.52	94.17±0.34	71.09±1.34
CT+SVM+TT	90.04±0.44	94.11±0.25	69.36±1.32
CT+HP+SVM	**90.80**±0.56	**94.38**±0.37	72.61±1.49
CT+HP+TT	90.01±0.50	94.01±0.31	69.75±1.38

Three voting taggers with the highest accuracy scores – the voting tagger using all the outputs (All taggers), combination of CroTag, HunPos and SVMTool and the combination of HMM taggers – outperformed all of the single taggers. Somewhat surprisingly, the highest performing assembly was not the assembly of all taggers but the tagger combining outputs from CroTag, HunPos and SVMTool. This is most likely due to the negative influence that the single lowest scoring tagger in the experiment, i.e. TreeTagger, had on the all-tagger assembly on the one side and the positive influence of higher disagreement between CroTag, HunPos and SVMTool (while still keeping their respective accuracies in the region of 90 percent) on the other side. It should once again be noted that all voting taggers were compared to CroTag by using the two-tailed t-test, which indicated that the observed difference between them was not in fact statistically significant. A brief discussion on this account is provided in the following section of the paper. A visualization of the scores is given in figure 1. The first part of the figure visualizes the overall tagging performance and performance on known and unknown words for all the single taggers, while the second part provides overall accuracies and 95-percent confidence intervals for all the voting taggers compared to CroTag.

4. Conclusions and future work

In this paper, we have presented results of an experiment dealing with combining five different part-of-speech taggers in the task of tagging Croatian texts via simple tagger voting. Tagger

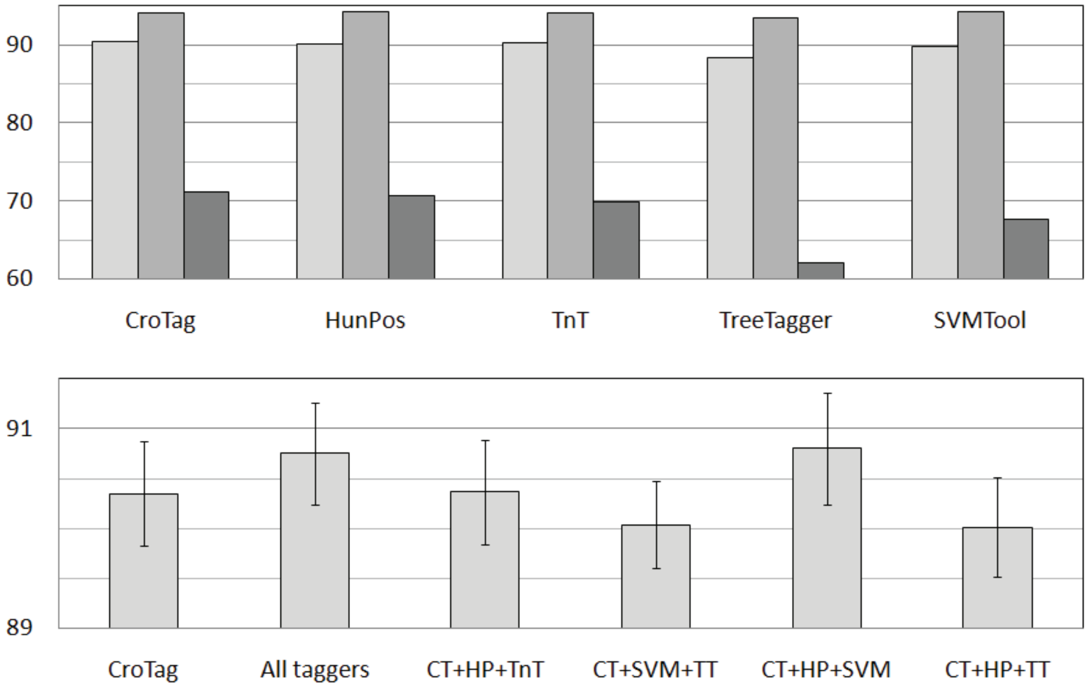

Figure 1. Accuracy of single and voting taggers

voting was shown to outperform single taggers in a majority of test scenarios. However, being that the observed differences between the best single taggers and the voting taggers was shown not to be statistically significant, these results should be considered with caution in terms of planning future research.

Namely, further work dealing with improving morphosyntactic tagging accuracy on Croatian texts will probably be planned along the lines that we already sketched in the introductory section. The tagger voting experiment presented here should be expanded to include other voting schemes, more carefully designed with respect to error analysis that was conducted for tagging Croatian texts with CroTag [3]. It would also be interesting to investigate the effect of tagset design and size to tagger voting. Experiments with using classifiers to disambiguate between outputs of different taggers, along the lines of [11] and [14] should also be conducted for Croatian. An experiment with tiered tagging [7] [20] [21] of Croatian texts is currently pending. Finally, once again reflecting on the results obtained by this experiment, it would be interesting to investigate whether more elaborate approaches to tagging by using single taggers would yield a more substantial improvement in overall tagging accuracy. Namely, we could test transformation based learning taggers or SVM taggers such as SVMTool using elaborate feature selection schemes reflecting the properties of Croatian language and compare them to the best voting taggers from this and future experiments with tagger voting. This line of research might provide a more valuable insight on what the best approach to morphosyntactic tagging of Croatian might be in terms of overall tagging accuracy. If followed by information on technical data for the specific taggers, such as the demands on memory and processing time, it would provide potential users with valuable information when choosing the best paradigm of tagging Croatian in terms of specific requirements of larger natural language processing systems.

5. Acknowledgements

The authors would like to thank Dragutin Miletić for attempting to include two additional morphosyntactic taggers – the Stanford Tagger and CRF++ suite – in the experiment. Due to limited time and other constraints, these were unfortunately not included.

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants No. 130-1300646-1776 and 130-1300646-0645.

6. References

[1] Agić Ž, Tadić M, Dovedan Z. (2008). Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis. Informatica. 32:4, pp. 445-451.

[2] Agić Ž, Tadić M, Dovedan Z. (2008). Investigating Language Independence in HMM PoS/MSD-Tagging. Proceedings of the 30th International Conference on Information Technology Interfaces. Zagreb, SRCE University Computer Centre, University of Zagreb, pp. 657-662.

[3] Agić Ž, Tadić M, Dovedan Z. (2009). Error Analysis in Croatian Morphosyntactic Tagging. Proceedings of the 31st International Conference on Information Technology Interfaces. Zagreb, SRCE University Computer Centre, University of Zagreb, 2009. pp. 521-526.

[4] Agić Ž, Tadić M, Dovedan Z. (2009). Tagset Reductions in Morphosyntactic Tagging of Croatian Texts. The Future of Information Sciences: Digital Resources and Knowledge Sharing. University of Zagreb, pp. 289-298.

[5] Attardi G, Fuschetto A, Tamberi F, Simi M, Vecchi E M. (2009). Experiments in tagger combination: arbitrating, guessing, correcting, suggesting. Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence.

[6] Brants T. (2000). TnT - a statistical part-of-speech tagger. Proceedings of ANLP 2000.

[7] Ceauşu A. (2006). Maximum Entropy Tiered Tagging. Proceedings of the 11th ESSLLI Student Session, June 20th 2006, Malaga, Spain, pp. 173-179.

[8] Erjavec T. (2004). Multext-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. Proceedings of the Fourth International Conference on Language Resources and Evaluation. ELRA, Lisbon-Paris 2004, pp. 1535-1538.

[9] Giménez, J, Márquez, L. (2004). SVMTool: A general POS tagger generator based on Support Vector Machines. Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon, Portugal, 2004.

[10] Halácsy P, Kornai A, Oravecz C. (2007). HunPos - an open source trigram tagger. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 209-212.

[11] Rupnik J, Grčar M, Erjavec T. (2008). Improving morphosyntactic tagging of Slovene by tagger combination. Proceedings of the Slovenian KDD conference – SiKDD 2008. Ljubljana, Slovenia, 2008.

[12] Schmid H. (1994). Probabilistic part-of-speech tagging using decision trees. Proceedings of International Conference on New Methods in Language Processing.

[13] Schmid H. (1995). Improvements In Part-of-Speech Tagging With an Application To German. Proceedings of the ACL SIGDAT-Workshop, pp. 47-50.

[14] Sjöbergh J. (2003). Combining POS-taggers for improved accuracy on Swedish text. NoDaLiDa 2003, 14th Nordic Conference on Computational Linguistics. Reykjavik, 2003.

[15] Spoustová D, Hajič J, Votrubec J, Krbec P, Květoň P. (2007). The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. Proceedings of the Workshop on Balto-Slavonic Natural Language Processing. Prague, Czech Republic, Association for Computational Linguistics, 2007.

[16] Tadić M. (2002). Building the Croatian National Corpus. Proceedings of LREC 2002. ELRA, Pariz-Las Palmas 2002, Vol. II, pp. 441-446.

[17] Tadić M, Fulgosi S. (2003). Building the Croatian Morphological Lexicon. Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages. Budapest, ACL, 2003. pp. 41-46.

[18] Tadić M. (2005). The Croatian Lemmatization Server. Southern Journal of Linguistics. 29:1/2, pp. 206-217.

[19] Tadić M. (2006). Developing the Croatian National Corpus and Beyond. Contributions to the Science of Text and Language. Word Length Studies and Related Issues. Kluwer, Dordrecht 2006, pp. 295-300.

[20] Tufiş D. (1999). Tiered Tagging and Combined Classifiers. Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 1692, pp. 28-33.

[21] Tufiş D, Dragomirescu L. (2004). Tiered Tagging Revisited. Proceedings of the 4th LREC Conference. Lisbon, Portugal, 2004, pp. 39-42.

[22] ACL Wiki: POS Tagging (State of the art). http://aclweb.org/aclwiki/index.php?title=POS_Tagging_(State_of_the_art) (2010-03-01).