Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: The Case of Slovene

Nikola Ljubešić,[*][†] **Tomaž Erjavec**[†]
[*] Dept. of Information and Communication Sciences, University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb, Croatia
nikola.ljubesic@ffzg.hr
[†] Dept. of Knowledge Technologies,
Jožef Stefan Institute,
Jamova cesta 3, SI-1000 Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Abstract

In this paper we present a tagger developed for inflectionally rich languages for which both a training corpus and a lexicon are available. We do not constrain the tagger by the lexicon entries, allowing both for lexicon incompleteness and noisiness. By using the lexicon indirectly through features we allow for known and unknown words to be tagged in the same manner. We test our tagger on Slovene data, obtaining a 25% error reduction of the best previous results both on known and unknown words. Given that Slovene is, in comparison to some other Slavic languages, a well-resourced language, we perform experiments on the impact of token (corpus) vs. type (lexicon) supervision, obtaining useful insights in how to balance the effort of extending resources to yield better tagging results.

Keywords: Part-of-Speech tagging, evaluation, Slavic languages

1. Introduction

Part-of-speech or, better, morphosyntactic tagging is still an interesting topic of research, esp. for highly inflected languages, such as Czech (Straková et al., 2014), Polish (Radziszewski, 2013) or Slovene (Grčar et al., 2012). Such languages with their large tagsets of morphosyntactic descriptions (MSDs) and often limited training data still offer significant room for improvement in tagging accuracy. A related research question is how to best split the effort needed to compile larger training corpora against extending the tagger background lexicon, a problem already investigated for French (Denis and Sagot, 2012). In this paper we address both questions.

We develop a new tagger, esp. optimised for tagging unknown (and partially unknown) words, useful for cases where the background lexicon is small or inflectionally incomplete, and test it on Slovene data.

We build on the approach by Grčar et al. (2012) as it is conceptually simple and is, in addition to Denis and Sagot (2012), one of the few proposals that treats tagging of seen and unseen tokens as an identical problem, utilising the knowledge from a morphosyntactic lexicon indirectly in form of classification features. It therefore does not consider the lexicon supplied MSDs as the only possible MSDs for the word, i.e. it uses the lexicon-as-features rather than the lexicon-as-constraint approach.

In contrast to Grčar et al. (2012) we replace an instance classifier (SVM) with a sequential one (CRF) and test additional features, which significantly improves their results, with an error reduction of $\sim$25% on both known and unknown words.

Finally, we repeat the experiment performed for French (Denis and Sagot, 2012) on the impact of the amount of token (corpus) and type (lexicon) supervision for Slovene.

2. Related Work

For Czech, Straková et al. (2014) describe the open-source tagger MorphoDiTa, based on the averaged perceptron, yielding 95.75% accuracy on MSD tagging and 97.8% on lemmatisation when training and tuning on the large 2 million word Prague Dependency Treebank PDT 2.5 (Bejček et al., 2012). The features used are defined in Spoustová et al. (2009). However, unknown word guessing is performed by a special program outside MorphoDiTa, which is not part of the tagger distribution.

For Polish Radziszewski (2013) describes a tiered CRF tagger, solving separately different levels of grammatical description, filtering thereby the hypothesis space obtained from a morphological lexicon. As features they use word forms, set of possible MSDs, set of possible numbers, genders and cases, gender, number and case agreement with following word and with previous and following word. Finally they use character-level features encoding the use of uppercase letters etc. They report an accuracy of 90.67% when 10-folding on a 1.2 million token corpus. Given the tiered nature of their CRF, this system is computationaly very expensive. The set-up is also rather complex and seems difficult to re-implement, while the code itself is not available as open source.

Kobyliński (2014) proposed an ensemble of existing classifiers, obtaining slight improvements for Polish with an accuracy of 92.05%. Waszczuk (2012) introduces a constrained version of CRF which classifies each token to one of the MSDs given by the morphological lexicon. They perform an initial ordering of possible MSDs of unknown words, keeping k=10 top candidates, after which they perform morphosyntactic disambiguation on all MSD candidates. One specific feature this work uses on unknown words is the packed shape of the word where uppercase characters are replaced with 'u', lowercased with 'l', digits with 'd' and others with 'x' and all repeating characters are removed. This work does not extract specific grammatical categories from MSDs. During disambiguation the complex tags are separated into two layers. The first one holds the PoS, case and person, while the second layer encodes the remaining information. The authors evaluate the tagger on the same corpus as Radziszewski (2013) and obtain an accuracy of 91.44%.

For Slovene Grčar et al. (2012) train a maximum-entropy classifier. The morphological lexicon is encoded as a suffix trie, where on each node all the seen MSDs are encoded. The suffix trie is used for producing two types of features: all the possible MSDs given the longest suffix of a token to be found in the trie, and each MSD by itself. The features used are words in the context of a window size 7, the result of the classification in the left context (the authors use an instance-level classifier, so no decoding of the optimal path is performed) and the hypotheses from the suffix trie for the right-side context. They additionally encode suffixes up to length 4 for the window of same size. Finally they use character-level features such as whether the token is lowercased, is punctuation, number etc. They report an accuracy of 92.49% when 10-folding on the ssj500k (Krek et al., 2013) corpus.

For Croatian and Serbian, languages closely related to Slovene, Agić et al. (2013) train a HunPos (Halácsy et al., 2007) model on a 90k-token Croatian manually annotated corpus (Agić and Ljubešić, 2014). They report ∼84% accuracy in annotating both Croatian and Serbian.

In our work we mostly follow Grčar et al. (2012) as they propose the most straightforward approach, considering tagging of seen and unseen tokens as an identical problem, feeding the knowledge from a morphosyntactic lexicon indirectly in form of features. However, we extend their approach by using a sequential tagger and inspecting many additional features.

3. The Dataset

For Slovene a number or resources are available as open datasets in the CLARIN.SI[1] repository. For our experiments we used ssj500k 1.3 (Krek et al., 2013), a 500k word corpus manually annotated with context-disambiguated MSDs (and lemmas) and the Sloleks morphological lexicon 1.2 (Dobrovoljc et al., 2015) which contains about 100,000 lemmas with their full inflectional paradigms.

In addition to improving the state-of-the-art for tagging of Slovene, our motivation also comes from the fact that this language has easy-to-obtain, high-quality and reasonably sized resources necessary for the task, thereby showing the optimal way on how to proceed for other morphologically complex languages, especially similar South Slavic ones such as Croatian, Serbian and Macedonian, with much less available resources.

The corpus was sequentially split into 10 folds, using the first 9 folds for the development of the tagger, in particular for determining the optimal set of tagger features, in a 9-fold cross-validation setting. We use the 10th fold for evaluating the final set of features and for performing the experiment on the impact of corpus vs. lexicon supervision on the tagging task. During the split of the corpus we did not shuffle sentences, but documents, thereby obtaining a more realistic split of the data.

The Sloleks morphological lexicon is used in both experiments in the form of classification features. Prior to feature extraction we encode the lexicon as a suffix tree (McCreight, 1976) in which each node contains all MSDs that were observed occurring with the specific suffix. The features we regularly encode by using this suffix tree are separate MSDs (or their specific morphological information) that are found in the suffix tree under the longest-to-be-found suffix of a specific surface form. Therefore, as a simple example, if we had the (token, MSD) pairs (označevanja, Ncnsg) and (kampanja, Ncfsn) encoded in the suffix tree, and if we observed the surface form ”banja”, we would enrich it with the ”Ncnsg” and ”Ncfsn” information, while for the surface form ”razumevanja” we would enrich it with ”Ncnsg” only. In the remainder of the paper we call these features MSD hypotheses.

4. Tagger Features and Evaluation

In this section we describe the feature selection process for our tagger. For extracting the features we use our own Python code while we train our models using CRFsuite (Okazaki, 2007). We close this section with an evaluation of the final tagger.

During the feature selection process we discriminate between two sets of features. The first, the core feature set, consists of features that were traditionally proven to work well for PoS tagging (Ratnaparkhi, 1996; Toutanova et al., 2003). On the second, the experimental feature set, we ran a large number of experiments in the quest for a (near-to) optimal feature set for the given language (family).

4.1. The Core Feature Set

The core feature set consists of the following features:

- lowercased tokens at positions -2, -1, 0, +1, +2
- focus token suffixes of length 1..4
- focus token packed representation giving information about the case of the word and whether it occurs at the beginning of the sentence, e.g. `ull-START` (starts with upper-case followed by at least two lower case character at the start of the sentence)
- focus token suffix trie MSD hypotheses

4.2. The Experimental Feature Set

The experimental feature set consists of the following features:

- lowercased tokens from a wider context
- suffixes of length greater than 4
- suffixes of tokens on positions -2, -1, +1, +2
- using suffix features only for tokens not covered by the lexicon

[1] http://www.clarin.si/

- distinguishing between suffix trie MSD hypotheses containing the complete token vs. containing its suffix only

- weighting the MSD hypotheses by the number of lexicon entries satisfying them

- ambiguity classes in the form of MSD hypothesis sets obtained from the suffix trie

- MSD hypotheses of tokens in positions -2, -1, +1, +2

- coarse MSD hypotheses of tokens in positions -2, -1, +1, +2; one coarse feture encodes the PoS and, if applicable, gender and number; another encodes the case

We ran experiments using our development set in a 9-fold cross-validation setting, using MSD accuracy as our evaluation metric. First, single experimental features were added to the core set. In the second batch different combinations of experimental features that were proven to be informative were added to the core feature set. After testing a large number of combinations (although not performing an exhaustive search over the defined space), we determined that the following experimental features improve the results:

- lowercased tokens on positions -3 and +3

- MSD hypotheses of tokens in positions -2, -1, +1 and +2

The remaining features showed no positive impact and were therefore discarded either as non-informative or already implicitly covered by other features.

Furthermore we realised that we obtain comparable results if we include MSD hypotheses only if the full token was observed in the lexicon, leaving MSD hypotheses of suffixes aside. This is probably to be explained by the size of the lexicon as only ∼2% of the words in the whole 500k-token corpus are not covered by the lexicon and the fact that including suffix features already deals with tokens not covered by the lexicon.

However, given that in the next section we perform experiments in which we drastically decrease the size of the lexicon used, we decided to keep for the the remainder of the experiments the more complex feature of MSD hypotheses given the longest suffix.

4.3. Final Evaluation

Table 1 gives the results of the evaluation of the optimised set of features in the final test set (the 10th fold) by measuring, on all tokens, the MSD accuracy, the PoS accuracy (the first letter of the MSD) and the extended PoS (the first two letters of the MSD) and the MSD accuracy on two types of unknown tokens:

1. lowercased token not seen in the training corpus;

2. lowercased token not seen in the training corpus nor in the lexicon.

Evaluation	Accuracy
MSD all tokens	94.27%
PoS all tokens	98.94%
Extended PoS all tokens	98.46%
MSD Type1 unknowns	84.39%
MSD Type2 unknowns	64.37%

Table 1: Evaluation results on the test set (10th fold)

All the experiments were run with CRFsuite (Okazaki, 2007) using 10 iterations of the passive aggressive learning algorithm.[2] The results are significantly better than Grčar et al. (2012) who used the same data and achieved 92.49% for the MSD (-1.78%, error reduction of 23.7%), 98.55% for PoS accuracy (-0.39%, error reduction of 26.9%) and 54.03% on MSD accuracy for unknown words (-10.34%, error reduction of 22.5%). It should be noted that Sloleks is quite a large lexicon, so unknown words tend to be foreign proper names, for which it is quite difficult, even for humans, to assign the correct MSD[3].

5. Token vs. Type Supervision

While developing resources necessary for high-quality morphosyntactic tagging of morphologically rich languages, regularly the question emerges about the ratio of corpus (token) and lexicon (type) supervision. It is our intention to cast some light on that dependence with the experiments described in this section, simplifying thereby, as we will show, the very important decisions that have to be made during that process.

We repeat the experiments for French by Denis and Sagot (2012) on Slovene while using the tagger presented in the previous section.

We train our systems on the development data (9 out of 10 folds) and evaluate them on our test set (10th fold). During the experiments we control the amount of token supervision (the size of the training corpus), and type supervision (the size of the lexicon encoded in the suffix trie). We increase the size of the corpus by following the order in our training set. The size of the lexicon is increased by adding lemmas in order of their descending corpus frequency (information present in Sloleks), thereby following a realistic scenario where the most frequent tokens / lemmas are added to the lexicon first.

We train all together 110 systems, ranging the size of the corpus from 10% to 100% and the size of the lexicon from 0% to 100%, all in 10% increments.

Figure 1 shows the results of these experiments as the estimated contours of specific tagger accuracies in the two-dimensional space of the amount of token and type supervision. We use the `gnuplots` basis spline algorithm (`bspline`) for smoothing.

Beside the accuracy contours, we plot the time contours, i.e. the estimate of the time (in hours) necessary to produce

[2] This learning setting was proven to give results comparable to the ones obtained by L-BFGS learning until convergence.

[3] In our test set only 1.5% of tokens were not seen neither in the corpus nor the lexicon.

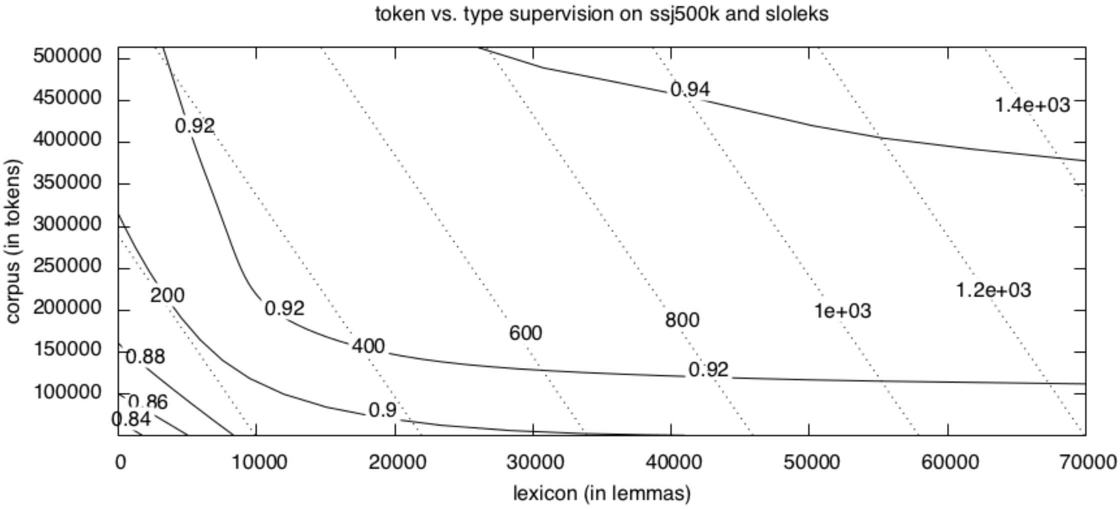

Figure 1: Token vs. type supervision presented with accuracy contours (full lines) and time contours in hours (dotted lines)

a specific amount of corpus and lexicon data. We plot the time contours with shaded lines.

The time contours are estimated by our previous measurements that the time necessary to correct a tag in the corpus is ∼2.5 seconds, and that adding a new lexeme to the lexicon while using the approach of predicting the lemma and paradigm of an OOV, as described in Ljubešić et al. (2015), takes on average ∼60 seconds.

The plot shows the results of the lexicon size up to 70,000 lemmas as no improvement on any corpus size can be observed after that point.

The first observation we can make is that for Slovene the ratio between token and type supervision for the task of morphosyntactic tagging is far from optimal as similar results to the best ones could have been obtained with a lexicon one third of size (roughly 30k lemmas) and slightly more corpus supervision, cutting thereby the resource production costs to less than a half.

Secondly, if we are interested in the optimal path through the contour plot up to that point, we have to look for locations where, for a given time investment, best accuracy is obtained. At the initial stages (up to accuracy of 0.88) corpus supervision is of much greater importance than lexicon supervision, but that phenomenon steadily levels off, which can be observed on the 0.84, 0.86 and 0.88 accuracy contours as they move steadily towards parallelity with the 200 hours time contour. Around the 0.9 accuracy contour the lexicon starts gaining in importance which can be observed on the closest entry point to 0.9 accuracy being just below 200,000 tokens of corpus supervision and just 5,000 lemmas for lexicon supervision while the same point for the 0.92 contour is around 200,000 corpus tokens and 12,000 lexicon entries. The overall phenomenon can probably be explained by the fact that on smaller sizes of corpora and lexicons the intersection of those two is higher, therefore the lexicon not giving any added value. With the increase in the size of these datasets, because of the Zipfian distribution in corpora which does not apply to lexicons, the lexicons start giving more information which is not covered by the corpora.

If we wanted to fit a 0-intercept linear function to our data for the optimal ratio of corpus and lexicon supervision, it would be somewhere at the 0.06 lexicon entries per token, i.e. 16.7 tokens per lexicon entry. This means that the overall rule of thumb for Slavic languages regarding the optimal relation between token and type supervision should be that for each lexeme added to the inflectional lexicon around 15 to 20 tokens in the corpus should be manually annotated. It is very important to note that this relation holds only for manually annotated corpora over the size of 200.000 tokens. Below that size investing in annotated corpora only makes most sense from the standpoint of obtaining maximum accuracy for the time invested in developing resources.

6. Conclusion

We have introduced a new CRF-based tagger and evaluated its performance on Slovene, reducing the error of the previously best performing Slovene tagger by 25%.

The code of the tagger, consisting of a lexicon compilation tool, a feature extractor, training and tagging scripts, as well as its models for Slovene and other South Slavic languages is available from `https://github.com/uzh/reldi/tree/master/tools/tagger`.

In contrast to most other taggers of highly inflected languages, tagging of unknown words is directly incorporated into the tagger architecture, rather than being handled by an additional process. Additionally, unlike most taggers developed, our tagger does not consider the ambiguity classes of words in the lexicon as being complete and correct, as this is often not the case for highly inflected languages with limited resources. Rather, the lexicon provides only features which are then considered by the classifier.

The other contribution of the paper is the analysis of corpus vs. lexicon supervision for developing highly accurate morphosyntactic taggers. We show that for Slovene, and probably all other Slavic languages, up to the corpus training size of around 200 thousand tokens investing in a lexicon does not pay off. On larger corpus sizes the optimal ratio between the lexicon size and the corpus size is around 15 tokens per lexicon entry.

7. Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran), the Slovenian Research Agency within the national basic research project "Resources, Tools and Methods for the Research of Nonstandard Internet Slovene" (J6-6842, 2014-2017) and the Swiss National Science Foundation grant no. IZ74Z0_160501 (ReLDI).

8. Bibliographical References

Agić, Ž. and Ljubešić, N. (2014). The SETimes.HR linguistically annotated corpus of Croatian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Agić, v., Ljubešić, N., and Merkler, D. (2013). Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. In *Proceedings of the Fourth Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 48–57, Sofia, Bulgaria, August. Association for Computational Linguistics.

Bejček, E., Panevová, J., Popelka, J., Straňák, P., Ševčíková, M., Štěpánek, J., and Žabokrtský, Z. (2012). Prague dependency treebank 2.5 – a revisited version of pdt 2.0. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, pages 231–246, Mumbai, India. Coling 2012 Organizing Committee.

Denis, P. and Sagot, B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Language Resources and Evaluation*, 46(4):721–736, December.

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., and Romih, M. (2015). *Morphological lexicon Sloleks 1.2*. Slovenian language resource repository CLARIN.SI. `http://hdl.handle.net/11356/1039`.

Grčar, M., Krek, S., and Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In *Zbornik Osme konference Jezikovne tehnologije*, Ljubljana, Slovenia.

Halácsy, P., Kornai, A., and Oravecz, C. (2007). HunPos: an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 209–212, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kobyliński, Ł. (2014). PoliTa: a Multitagger for Polish. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Krek, S., Erjavec, T., Dobrovoljc, K., Može, S., Ledinek, N., and Holz, N. (2013). *Training corpus ssj500k 1.3*. Slovenian language resource repository CLARIN.SI. `http://hdl.handle.net/11356/1029`.

Ljubešić, N., Esplà-Gomis, M., Klubička, F., and Preradović, N. M. (2015). Predicting Inflectional Paradigms and Lemmata of Unknown Words for Semi-automatic Expansion of Morphological Lexicons. In *Proceedings of Recent Advances in Natural Language Processing*, pages 379–387.

McCreight, E. M. (1976). A space-economical suffix tree construction algorithm. *Journal of the Association for Computing Machinery*.

Okazaki, N. (2007). CRFsuite: a fast implementation of Conditional Random Fields (CRFs). `http://www.chokkan.org/software/crfsuite/`.

Radziszewski, A. (2013). A Tiered CRF Tagger for Polish. In *Intelligent Tools for Building a Scientific Information Platform*, volume 467 of *Studies in Computational Intelligence*, pages 215–230. Springer Berlin Heidelberg.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In *Proceedings of EMNLP*. ACL.

Spoustová, D., Hajič, J., Raab, J., and Spousta, M. (2009). Semi-supervised Training for the Averaged Perceptron POS Tagger. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 763–771, Stroudsburg, PA, USA. Association for Computational Linguistics.

Straková, J., Straka, M., and Hajic, J. (2014). Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*, pages 13–18.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Waszczuk, J. (2012). Harnessing the CRF Complexity with Domain-Specific Constraints. The Case of Morphosyntactic Tagging of a Highly Inflected Language. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings*, pages 2789–2804.