What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian

Nikola Ljubešić
Jožef Stefan Instute
Jamova cesta 39
1000 Ljubljana, Slovenia
nikola.ljubesic@ijs.si

Kaja Dobrovoljc
Jožef Stefan Institute
Jamova cesta 39
1000 Ljubljana, Slovenia
kaja.dobrovoljc@ijs.si

Abstract

We present experiments on Slovenian, Croatian and Serbian morphosyntactic annotation and lemmatisation between the former state-of-the-art for these three languages and one of the best performing systems at the CoNLL 2018 shared task, the Stanford NLP neural pipeline. Our experiments show significant improvements in morphosyntactic annotation, especially on categories where either semantic knowledge is needed, available through word embeddings, or where long-range dependencies have to be modelled. On the other hand, on the task of lemmatisation no improvements are obtained with the neural solution, mostly due to the heavy dependence of the task on the lookup in an external lexicon, but also due to obvious room for improvements in the Stanford NLP pipeline's lemmatisation.

1 Introduction

Morphosyntactic annotation and lemmatisation are crucial tasks for languages that are rich in inflectional morphology, such as Slavic languages. These tasks are far from solved, and the recent CoNLL 2017 (Zeman et al., 2017) and CoNLL 2018 (Zeman et al., 2018) shared tasks on multilingual parsing from raw text to Universal Dependencies (Nivre et al., 2016) have given the necessary spotlight to these problems. In addition to the advances due to multi- and cross-lingual settings, the participating systems have also confirmed the predominance of neural network approaches in the field of natural language processing.

In this paper we compare the improvements obtained on these two tasks in three South Slavic languages (Slovenian, Croatian and Serbian) by moving from traditional approaches to the neural ones. The tool that we use as the representative of the traditional approaches is `reldi-tagger` (Ljubešić and Erjavec, 2016; Ljubešić et al., 2016), the previous state-of-the-art for morphosyntactic tagging and lemmatisation of the three focus languages due to (1) carefully engineered features for the CRF-based tagger, (2) integration of an inflectional lexicon both for the morphosyntactic tagging and the lemmatisation task and (3) lemma guessing for unknown word forms via morphosyntactic-tag-specific Naive Bayes classifiers, predicting the transformation of the surface form. The tool that we use as the representative for the neural approaches is `stanfordnlp`, the Stanford NLP pipeline (Qi et al., 2018), a state-of-the-art in neural morphosyntactic and dependency syntax text annotation. The system took part in the CoNLL 2018 shared task (Zeman et al., 2018) as one of the best-performing systems, which would have, with "an unfortunate bug fixed", placed among the top-three for all evaluation metrics, including lemmatisation and morphology prediction. The tool is, additionally, released as open source and has a vivid development community,[1] with a named entity recognition module being in development.

2 Experiment Setup

We perform our comparison of the traditional and the neural tool of choice on the two tasks on data splits defined in the `babushka-bench` benchmarking platform[2] which currently hosts data and results for the three South Slavic languages we use in these experiments, namely Slovenian, Croatian and Serbian. It is organised as a git repository, with scripts for transferring datasets from the CLARIN.SI repository,[3] and splitting them into

[1] https://github.com/stanfordnlp/stanfordnlp

[2] https://github.com/clarinsi/babushka-bench

[3] https://www.clarin.si/repository/xmlui

Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pages 29–34,
Florence, Italy, 2 August 2019.

training, development, and testing portions. While the primary usage of this platform are in-house experiments on the available and emerging technologies, other researchers are more than welcome to further enrich the repository.

The name of the repository has its roots in the erroneous, but popular naming of the Matryoshka doll in South Slavic languages, as the datasets are split into train, dev and test portions in a random fashion, but with a fixed random seed. This enables splitting the same datasets on the annotation layers that were not applied over the whole dataset (as is often the case with costly annotations of syntax, semantic etc.), and simultaneously ensuring that no spillage between train, dev and test between the various layers would occur. There are many cases where such a split comes handy for benchmarking, one example being using the whole datasets for training taggers and just portions of the datasets (i.e. the manually parsed subsets) to train parsers that require tagging as upstream processing.

For evaluating morphosyntactic tagging and lemmatisation in `babushka-bench`, we use a modified CoNLL 2018 shared task evaluation script to enable evaluation without parsing present. This script calculates the F1 metric between the gold and the real annotations, taking into account the possibility of different segmentation, which is not the case in these experiments as we use gold segmentation from the datasets to focus on the tasks of morphosyntactic tagging and lemmatisation. When modelling morphosyntax, we predict morphosyntactic descriptions (MSDs), position-based encodings of part-of-speech and feature-value pairs, as defined in the MULTEXT-East tagset (Erjavec, 2012). The training-data-defined size of the tagset for each of the three languages lies between 600 and 1300 MSDs, depending on the language and the size of the training data. This is the default tagset for the `reldi-tagger` and is also supported by the `stanfordnlp` tool, where language-specific tags (XPOS) are predicted as one of the three outputs by the tagging module (the other two being UD parts-of-speech (UPOS) and features (FEATS)). The datasets we use for our experiments are the three official datasets for training standard language technologies for these languages. These are the ssj500k dataset for Slovenian (Krek et al., 2019), the hr500k dataset for Croatian (Ljubešić et al., 2018) and the SETimes.SR dataset for Serbian (Batanović et al., 2018). While the Slovenian and Croatian datasets are both around 500 thousand tokens in size, the Serbian dataset is significantly smaller with only 87 thousand tokens in size. We additionally make use of the inflectional lexicons of these three languages, Sloleks for Slovenian (Dobrovoljc et al., 2019), hrLex for Croatian (Ljubešić, 2019a) and srLex for Serbian (Ljubešić, 2019b), all containing more than 100 thousand lemmas with around 3 million inflected forms.

While learning neural morphosyntactic taggers, we also experiment with various embeddings, mostly (1) the original CoNLL 2017 word2vec (w2v) embeddings for Slovenian and Croatian (Ginter et al., 2017) (there are none available for Serbian), based on the CommonCrawl data, and (2) the CLARIN.SI embeddings for Slovenian (Ljubešić and Erjavec, 2018), Croatian (Ljubešić, 2018a) and Serbian (Ljubešić, 2018b), either trained with fastText (fT) or with word2vec (w2v)[4] on large, just partially publicly available texts due to copyright restrictions.

Our experiments are split into two main parts: experiments on morphosyntactic tagging in Section 3.1, backed with the comparison of the difference of the most frequent errors in the traditional and neural approaches, and the experiments on lemmatisation in Section 3.2.

3 Results

3.1 Morphosyntax

We first compare the results of the two tools on morphosyntactic annotation, trained on the training portion of the datasets of the three languages, with development data used if necessary.[5] The results of the two taggers on the two languages are presented in Table 1.

The results show significant differences between `reldi-tagger` and `stanfordnlp`, with relative error reduction of 43% for Slovenian, 27% for Croatian and 40% for Serbian. Regarding

[4]Currently only the fastText versions are available for download in the repository.

[5]While `stanfordnlp` uses the development data for updating the learning rate and optimization algorithm, `reldi-tagger` did not make any use of the development data during this training phase. However, during the development of `reldi-tagger`, a series of feature selections and hyperparameter values were investigated on held-out data, so we can consider for that tool to have used development data indirectly, as well.

tool	distributional information	Slovenian	Croatian	Serbian
reldi-tagger	Brown clusters	94.21	91.91	92.03
stanfordnlp	CoNLL w2v embeddings	96.45	93.85	94.78
stanfordnlp	CLARIN.SI w2v embeddings	**96.79**	**94.18**	94.91
stanfordnlp	CLARIN.SI fT embeddings	96.72	94.13	**95.23**

Table 1: F1 results in morphosyntactic annotation with the traditional and neural tool and different distributional information.

	Slovenian			Croatian			Serbian		
	true	pred	freq	true	pred	freq	true	pred	freq
reldi-tagger	Ncmsan	Ncmsn	109	Xf	Npmsn	162	Xf	Npmsn	28
	Ncmsn	Ncmsan	71	Qo	Cc	118	Ncmsan	Ncmsn	22
	Ncnsa	Ncnsn	61	Ncmsan	Ncmsn	117	Npmsan	Npmsn	13
	Ncfpa	Ncfsg	47	Ncmsn	Ncmsan	98	Ncmsn	Ncmsan	12
	Agpnsn	Rgp	41	Ncfpa	Ncfsg	56	Ncmsg	Ncmpg	12
	Ncfpn	Ncfsg	36	Cs	Rgp	55	Ncfpn	Ncfsg	12
	Ncnsn	Ncnsa	35	Ncmpg	Ncmsg	53	Ncmpg	Ncmsg	11
	Agpnsa	Agpnsn	31	Ncmsg	Ncmpg	50	Npmsay	Npmsg	9
	Sa	Sl	27	Agpnsny	Rgp	48	Ncnsn	Ncnsa	8
	Npmsay	Npmsg	27	Ncnsa	Ncnsn	43	Ncfpa	Ncfsg	8
stanfordnlp	Ncmsn	Npmsn	54	Xf	Npmsn	111	Xf	Npmsn	20
	Pp3fpa–y	Pp3mpa–y	31	Qo	Cc	96	Ncmsan	Ncmsn	10
	Ncmsan	Ncmsn	28	Cs	Rgp	75	Ncmpg	Ncmsg	10
	Cc	Rgp	28	Npmsn	Xf	74	Npfsn	Npmsn	8
	Ncmsn	Ncmsan	27	Mro	Mdo	57	Ncmsn	Ncmsan	8
	Xf	Npmsn	20	Ncmsg	Ncmpg	50	Npmsan	Npmsn	7
	Ncnsn	Ncnsa	18	Ncmsan	Ncmsn	42	Ncnsn	Ncnsa	5
	Pp3nsa–y	Pp3msa–y	17	Ncmpg	Ncmsg	38	Ncmsg	Ncmpg	5
	Npfsn	Npmsn	17	Rgp	Cs	37	Npmsn	Npmsan	4
	Mlc-pn	Mlc-pa	17	Cc	Qo	36	Ncnsa	Ncnsn	4

Table 2: Most frequent errors by the traditional and neural tagger on Slovenian, Croatian and Serbian.

the usage of different embedding collections with stanfordnlp, there are no drastic differences, but the CLARIN.SI embeddings show to be better suited than the CoNLL embeddings, which does not come as a surprise as the former are based on more text, which is frequently also of higher quality. The distinction between word2vec (w2v) and fastText (fT) embeddings shows to be minimal, but fastText seems to be more beneficial when smaller amounts of training data are available, as is the case with Serbian.

For the error analysis, as well as downstream experiments on lemmatisation, for which morphosyntactic annotation is a prerequisite, we take the stanfordnlp tool with CLARIN.SI fastText embeddings, as these settings achieve the best results on average.

To identify the differences in morphosyntactic tagging errors between the traditional and neural tagger, we analyse the 10 most frequent confusions per tagger for each of the three languages. Our results presented in Table 2 show that some of the most frequent errors in reldi-tagger are substantially reduced by stanfordnlp, such as the confusion between masculine nouns in singular accusative (Ncmsan) and nominative (Ncmsn), which shows the neural tagger to be more capable in modelling long-range dependencies. Namely, whether a male noun is in the nominative or accusative case depends mostly on whether one of these two cases already occurred somewhere in the clause.

Another regular confusion in morphosyntactic tagging in general, which is also heavily resolved

tool	morphosyntax	Slovenian	Croatian	Serbian
reldi-tagger	gold	99.46	98.17	97.89
reldi-tagger	reldi-tagger	98.35	96.82	96.44
reldi-tagger	stanfordnlp	98.77	97.22	97.26
stanfordnlp	gold	97.75	96.22	95.29
stanfordnlp	stanfordnlp	97.51	95.85	95.18
stanfordnlp+lex	gold	99.30	98.11	97.78
stanfordnlp+lex	stanfordnlp	98.74	97.22	97.13

Table 3: F1 results in lemmatisation with the traditional and neural tool and different upstream processing.

by the neural tagger, is that between adjectives in the neutrum nominative (Agpnsn) and adverbs (Rgp), which, again, requires information from a wider context, i.e., whether there is a noun to which the potential adjective can be attached to.

An error type which requires more of a semantic understanding is the distinction between proper nouns (Npmsn) and foreign residuals (Xf) in Croatian and Serbian. In these two languages, the rule is that proper nouns of foreign origin (*Easy Jet*, *Feng Shui*) are annotated as foreign residuals. This type of error is in good part resolved via word embedding information where this distinction is obviously encoded, while in the 1000 hierarchical Brown clusters this is obviously not the case.

Interestingly, some shared errors are even more frequent in the neural `stanfordnlp` predictions, such as the disambiguation between homonymous conjunctions (Cc, Cs) and adverbs (Rgp) for Croatian and Slovenian (e.g. *već*, *tako*, *zato*), which does come as a surprise as this distinction requires long-range information which should be more available in the neural approach.

3.2 Lemmatisation

Given that morphosyntactic information is usually expected as the input to lemmatisation, we compare the lemmatisation performance of the two tools if (1) gold morphosyntax is given, (2) the morphosyntax predicted by the tool itself is used and (3) the best predicted morphosyntax by `stanfordparser` is used. In addition to that, we also expand `stanfordnlp` with a simple intervention in the lemmatisation procedure, in which the lexicon lookup is not performed over the training data only, but the external inflectional lexicons as well, naming this modified tool `stanfordnlp+lex`.

The results of the lemmatisation experiments are given in Table 3. The results show that `reldi-tagger` outperforms the original `stanfordnlp` by a substantial margin, which does not come as a surprise as `reldi-tagger` uses a large inflectional lexicon. A simple lexicon intervention with `stanfordnlp+lex` closes the gap between the two, with almost no difference in lemmatisation quality for any of the languages.

Regarding different upstream processing, as expected, preprocessing with `stanfordnlp` closes one third of the gap between preprocessing with `reldi-tagger` and having perfect, gold morphosyntactic annotation.

Investigating the differences between the decisions of `reldi-tagger` and `stanfordnlp+lex` shows that these mostly differ in handling named entities, with both tools missing the correct lemma with similar frequency. For `stanfordnlp+lex` in particular, some errors can be attributed to the fact it does not rely on the morphological feature (FEATS) information when looking up the lexicon and producing lemma predictions, causing errors such as generating a feminine proper noun lemma for a correctly tagged masculine proper noun.

4 Conclusion

In this paper we have presented the set up of the long-term evaluation platform for benchmarking current and future NLP tools for the three South Slavic languages, a practice which is still far too rare. We did a comparative evaluation of two state-of-the art tools with different architectures (traditional vs. neural) and confirmed that the neural approach yields significant improvements in tagging, especially because of better long-range dependency modelling and more distributional semantic information available.

For lemmatisation, the results of both approaches are very close, especially because of a heavy dependence on the lookup in a large inflec-

tional lexicon, but with obvious room for improvement in the neural lemmatisation process.

The presented results give important pointers for the development of future state-of-the-art tools for the three languages, but also Slavic languages in general.

Acknowledgments

The work described in this paper was funded by the Slovenian Research Agency within the national basic research projects "Resources, methods and tools for the understanding, identification and classification of various forms of socially unacceptable discourse in the information society" (J7-8280, 2017–2020) and "New grammar of contemporary standard Slovene: sources and methods" (J6-8256, 2017–2020), the national research programme "Language Resources and Technologies for Slovene" (P6-0411), the Slovenian-Flemish bilateral basic research project "Linguistic landscape of hate speech on social media" (N06-0099, 2019–2023), and the Slovenian research infrastructure CLARIN.SI.

References

Vuk Batanović, Nikola Ljubešić, Tanja Samardžić, and Tomaž Erjavec. 2018. *Training corpus SETimes.SR 1.0*. Slovenian language resource repository CLARIN.SI.

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, Miro Romih, špela Arhar Holdt, Jaka Čibej, Luka Krsnik, and Marko Robnik-šikonja. 2019. *Morphological lexicon Sloleks 2.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1230.

Tomaž Erjavec. 2012. Multext-east: morphosyntactic resources for central and eastern european languages. *Language Resources and Evaluation*, 46(1):131–142. https://doi.org/10.1007/s10579-011-9174-8.

Filip Ginter, Jan Hajič, Juhani Luotolahti, Milan Straka, and Daniel Zeman. 2017. *CoNLL 2017 Shared Task - Automatically Annotated Raw Texts and Word Embeddings*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, Taja Kuzman, Jaka Čibej, Špela Arhar Holdt, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, and Anja Zajc. 2019. *Training corpus ssj500k 2.2*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1210.

Nikola Ljubešić. 2018a. *Word embeddings CLARIN.SI-embed.hr 1.0*. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić. 2018b. *Word embeddings CLARIN.SI-embed.sr 1.0*. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić and Tomaž Erjavec. 2018. *Word embeddings CLARIN.SI-embed.sl 1.0*. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić. 2019a. *Inflectional lexicon hrLex 1.3*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1232.

Nikola Ljubešić. 2019b. *Inflectional lexicon srLex 1.3*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1233.

Nikola Ljubešić, Željko Agić, Filip Klubička, Vuk Batanović, and Tomaž Erjavec. 2018. *Training corpus hr500k 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1183.

Nikola Ljubešić and Tomaž Erjavec. 2016. Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: the Case of Slovene. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. 2016. New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics. https://nlp.stanford.edu/pubs/qi2018universal.pdf.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics. http://www.aclweb.org/anthology/K18-2001.

Daniel Zeman et al. 2017. CoNLL 2017 Shared Task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.