Comparing CRF and LSTM performance on the task of morphosyntactic tagging of non-standard varieties of South Slavic languages

Nikola Ljubešić
Dept. of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39
1000 Ljubljana, Slovenia
nikola.ljubesic@ijs.si

Abstract

This paper presents two systems taking part in the Morphosyntactic Tagging of Tweets shared task on Slovene, Croatian and Serbian data, organized inside the VarDial Evaluation Campaign. While one system relies on the traditional method for sequence labeling (conditional random fields), the other relies on its neural alternative (bidirectional long short-term memory). We investigate the similarities and differences of these two approaches, showing that both methods yield very good and quite similar results, with the neural model outperforming the traditional one more as the level of non-standardness of the text increases. Through an error analysis we show that the neural system is better at long-range dependencies, while the traditional system excels and slightly outperforms the neural system at the local ones. We present in the paper new state-of-the-art results in morphosyntactic annotation of non-standard text for Slovene, Croatian and Serbian.

1 Introduction

In this paper we present two systems taking part in the MTT (Morphosyntactic Tagging of Tweets) shared task, part of the VarDial Evaluation Campaign (Zampieri et al., 2018). In the task, general-domain and in-domain datasets with tokens manually annotated with morphosyntactic descriptions (MSDs), are given, together with large web-based datasets, for three South Slavic languages: Slovene, Croatian and Serbian. The challenge of the task is to exploit similarity of standard vs. non-standard variants, as well as the overall proximity of the three languages in question.

While the first system, JANES, relies on the traditional method for sequence labeling, namely conditional random fields (CRF), the second system, JSI, relies on the currently hugely popular neural networks, more precisely bidirectional long short-term memories (BiLSTM).

The contributions of this paper are the following: (1) a direct comparison of CRFs and BiLSTMs on a series of datasets, where CRFs are equipped with carefully engineered features, not generic ones, and (2) a new state-of-the-art in tagging non-standard varieties of the three languages in question.

2 System Descriptions

2.1 Datasets Distributed inside the Shared Task

Before we describe our two systems participating in the task, we quickly quantify the available resources through token number in Table 1 as these heavily influence our decisions in the system setup. The `twitter.*` datasets come from the Janes-Tag manually annotated dataset of Slovene computer-mediated communication (Erjavec et al., 2017) and the ReLDI-NormTagNER-* manually annotated datasets of Croatian (Ljubešić et al., 2017b) and Serbian (Ljubešić et al., 2017c) tweets. They are all similar in size, with cca. 40 thousand tokens available for training, 8 thousand for development and 20 thousand for testing.

Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 156–163
Santa Fe, New Mexico, USA, August 20, 2018.

The `standard.train` datasets mostly cover the general domain. While the Slovene and Croatian datasets are similar in size with around 500 thousand tokens, the Serbian dataset is significantly smaller with only 87 thousand tokens.

The `web.auto` datasets are large web-based datasets, slWac for Slovene (Erjavec et al., 2015), hrWaC for Croatian and srWaC for Serbian (Ljubešić and Klubička, 2014). These are automatically annotated with state-of-the-art taggers of standard language for Slovene (Ljubešić and Erjavec, 2016) and Croatian and Serbian (Ljubešić et al., 2016).

	twitter.train	twitter.dev	twitter.test	standard.train	web.auto
Slovene	37,756	7,056	19,296	586,248	895,875,492
Croatian	45,609	8,886	21,412	506,460	1,397,757,548
Serbian	45,708	9,581	23,327	86,765	554,627,647

Table 1: Size of datasets distributed through the MTT shared task. Sizes are in number of tokens.

2.2 The JANES System

The JANES system (the name of the system comes from the Slovene national project JANES inside which the system was developed[1]) is based on conditional random fields (CRFs) (Lafferty et al., 2001), exploiting the following handcrafted features:

- lowercased focus token (token for which features are being extracted)
- lowercased tokens in a window of $\{-3, -2, -1, 1, 2, 3\}$ form the focus token
- focus token suffixes of length $\{1, 2, 3, 4\}$
- features encoding whether the focus token starts with `http` (link), `#` (hashtag) or `@` (mention)
- Brown cluster binary paths for the focus token, with the path length of $\{2, 4, 6, 8\}$

These features were proven to yield optimal results in our previous work on tagging non-standard Slovene (Ljubešić et al., 2017a).

The Brown clusters, the output of a method for context-dependent hierarchical word clustering (Brown et al., 1992), were calculated from the web data that were made available through the shared task, namely the slWaC web corpus of Slovene (Erjavec et al., 2015) and the hrWaC and srWaC corpora of Croatian and Serbian (Ljubešić and Klubička, 2014). We have used default parameters for calculating Brown clusters, except for the minimum occurrence parameter which was set to 5. The web text was previously lowercased and punctuations and newlines were removed from it.

For training the tagger, we exploited (1) the proximity of the Croatian and Serbian language, and (2) the fact that we have much more standard training data and much less Twitter training data. We sampled our final training data for each language in the following manner:

- for Slovene: we added to the Slovene standard training data ten times the available non-standard data, thereby reaching a similar amount of standard and non-standard data in our training set; from previous work we know that for CRFs oversampling in-domain data is the simplest and most effective method in merging out-domain and in-domain training data (Horsmann and Zesch, 2015; Ljubešić et al., 2017a)
- for Croatian: we merged the Croatian and the Serbian standard language training datasets, added to it ten copies of the Croatian Twitter training dataset and two copies of the Serbian training dataset, thereby putting emphasis on the Croatian training data, which is expected to be closer to the Croatian test data

[1] `http://nl.ijs.si/janes/english/`

- for Serbian: we merged the Croatian standard training data, two copies of the Serbian standard training data (as these are more than five times smaller than the Croatian ones), ten copies of non-standard Croatian training data, and four copies of non-standard Serbian training data, with the rationale that most non-standard elements in Croatian are present in non-standard Serbian as well, but with lower frequency; by oversampling non-standard Croatian in the Serbian dataset we emphasize the non-standard elements in the Croatian non-standard training data as the Serbian non-standard data is much closer to the standard language (Miličević and Ljubešić, 2016)

The system was implemented in CRFSuite (Okazaki, 2007), using the passive aggressive optimizer and 10 epochs, a setting which proved to yield best results in previous experiments (Ljubešić and Erjavec, 2016).

2.3 The JSI System

The JSI system (the name comes from the name of our current employer, the Jožef Stefan Institute) is an adaptation of the BiLSTM tagger written in pytorch[2], with some added modifications. The architecture of the submitted system is the following:

- a character-level subnetwork, consisting of a character embedding layer of 16 dimensions and a BiLSTM layer with 25 units
- the main network
  - concatenating the character-level representation of a word from the subnetwork described above (25*2, i.e., 50 dimensions), and the word embedding layer (100 dimensions)
  - feeding this concatenated 150-dimensional character- and word-level representation into a BiLSTM layer with 100 units
  - the per-token BiLSTM output being fed to a fully-connected layer with 256 units and a final softmax layer for prediction

While developing this architecture, we investigated the impact of various setups on the Slovene dataset. The results of experimenting with (1) different pretrained word embeddings, (2) the impact of adding different character-level representations, (3) fine-tuning the model on in-domain data and (4) pretraining the character-level encoder on a inflectional lexicon, are shown in Table 2. We performed our experiments on each of the above mentioned issues subsequently, always propagating to the next experiment set the setup achieving best results in the previous one. The setup we start with consists only of the main network, without the character-level subnetwork.

2.3.1 Word Embeddings

The first group of results considers different ways of pretraining word embeddings. The word embeddings were always pretrained on the web data available for each language.

We considered only two tools for pretraining word embeddings: word2vec (Mikolov et al., 2013) and fasttext (Bojanowski et al., 2017), and two architectures, CBOW and Skipgram. The results (*word2vec cbow* vs. *word2vec skipgram*) show for Skipgram to be significantly better suited for this task, which is in line with previous results (Reimers and Gurevych, 2017).

Comparing word2vec and fasttext (*word2vec skipgram* vs. *fasttext skipgram*), fasttext shows a slightly better performance, but the difference gets more obvious (almost half a point in token accuracy) once fasttext is used to generate representations for the words not present in the pretrained word embeddings (*fasttext skipgram generated*).[3]

[2] https://github.com/neulab/dynet-benchmark/blob/master/pytorch/bilstm-tagger-withchar.py

[3] We would expect the positive impact of generating embeddings for out-of-vocabulary words to diminish once character-level representations are added to the model. However, we did not investigate this.

setup	token accuracy with stdev
word2vec cbow	0.8407 ± 0.0025
word2vec skipgram	0.8550 ± 0.0041
fasttext skipgram	0.8578 ± 0.0041
fasttext skipgram generated	$\mathbf{0.8596 \pm 0.0031}$
added character-level encoding	0.8780 ± 0.0030
added bidirectional encoding	$\mathbf{0.8790 \pm 0.0032}$
additionally tuned on in-domain data	$\mathbf{0.8836 \pm 0.0026}$
pretrained character-level encoder on web data	$\mathbf{0.8855 \pm 0.0015}$

Table 2: Initial experiments on the JSI system, performed on the Slovene dataset. The standard deviation is calculated from ten evaluations performed during the last epoch.

2.3.2 Character-level Representations

The second group of experiments considers the impact of adding character-level representations of each token to the word representation via a dedicated character-level BiLSTM. Adding the character-level representation has shown the biggest impact among all the experiments, with ~ 2 accuracy points increase, and a minor difference between encoding the character sequence with a single-direction or a bi-directional LSTM.

2.3.3 Fine-tuning on the In-Domain Dataset

The third experiment considers the impact of not training the network on a simple merge of all the available relevant training data, but also fine-tuning the network exclusively on in-domain data.

Running three epochs on the concatenation of all datasets, and then additional two epochs only on the in-domain Twitter data, consistently improved the results for around half an accuracy point.

This method is somewhat similar to the oversampling method applied on the JANES system. It is, however, more elegant as it gives greater control over the amount and order of data fed into the system.

2.3.4 Pretraining the Character-level BiLSTM

Finally, in the last set of experiments we investigated whether there is positive impact if the character-level encoder was pretrained on a inflectional-lexicon-like resource. In this shared task the web data were automatically tagged with a CRF tagger relying on a lexicon (Ljubešić et al., 2016; Ljubešić and Erjavec, 2016), therefore we transformed the automatically-tagged web data into a lexicon by (1) picking only token-tag pairs occurring at least 100 times in the web data and (2) selecting only the most frequent token-tag pair per token. With the second criterion we lost some information on homonymous words, but also got rid of a lot of wrong automatic annotations of frequent words.

The results on pretraining the character-level encoder show that the improvement lies below half an accuracy point, but this improvement showed to be consistent across all the three languages.[4]

3 Results

In this section we report the results of the final setups of the JANES and the JSI system and compare it to the HunPos baseline (Halácsy et al., 2007) defined by the shared task organizers.

Additionally, we report the results of the JANES system using an inflectional lexicon for the specific language, namely Sloleks for Slovene (Dobrovoljc et al., 2015), hrLex for Croatian (Ljubešić et al., 2016a) and srLex for Serbian (Ljubešić et al., 2016b). We call this system JANES-lex. We compare to this system as it is very straightforward to add information from an inflectional lexicon as additional features to a CRF-based system.

[4]On Croatian data we ran an additional experiment not with the noisy web data, but the manually constructed inflectional lexicon hrLex (Ljubešić et al., 2016a), improving additionally for almost half an accuracy point. However, in this shared task we decided not to use resources that were not shared by the organizers as we (correctly) assumed that other teams will not use additional resources neither and that we would lower the comparability of the obtained results.

	Slovene	Croatian	Serbian
HunPos baseline	0.832	0.834	0.832
JANES	0.871	0.893	0.900
JANES-lex	0.877	0.897	0.901
JSI	0.883	0.890	0.900
JSI-simpler	**0.891**	**0.898**	**0.903**

Table 3: Results of the two systems, their two adaptations and the baseline on the test data. Reported metric is token-level accuracy.

We also report a modification of the JSI system that we implemented after the shared task was already concluded. Namely, we removed the fully connected layer between the main BiLSTM and the softmax layer, which is actually the most frequent setup for sequence labeling. The removed layer in the JSI system is a residue from the tagger we based our implementation on[5]. We call the simplified tagger JSI-simpler.

The results of the two taggers and the two variants are given in Table 3. The reported results are those obtained on the test data.

We can first observe that (1) all the systems outperform the HunPos baseline by a wide margin and that (2) the results of the four remaining systems are rather close.

The largest difference that can be observed between the four systems are 2 accuracy points on Slovene between the basic CRF implementation (JANES) and the simplified BiLSTM implementation (JSI-simpler). The same difference is not to be observed on the other two languages, with the same systems having a difference of 0.5 points on Croatian and 0.3 points on Serbian. The reason for the larger difference on Slovene data lies in the fact that the Slovene data is least standard (17% tokens being non-standard), followed by Croatian (13% non-standard tokens), with Serbian data deviating the least from the norm (10% non-standard tokens) (Miličević et al., 2017) as more complex modeling techniques pay off more as the language deviates stronger from the norm.

Adding lexicon information to the JANES system (JANES vs. JANES-lex) improves the results on all three languages, but just slightly, between 0.1% and 0.6%. Previous work on the problem (Ljubešić et al., 2017a) has shown that Brown clusters already provide to a large extent the information that was traditionally obtained through inflectional lexicons.

Comparing the JANES and JSI results by using the McNemar's statistical test (McNemar, 1947), the difference on Slovene is statistically significant at the $p < 0.001$ level, with an absolute difference in 1.2 points and an relative error reduction of 9.3%. The differences on the remaining two languages are not statistically significant.

When comparing the JSI and JSI-simpler results, it becomes obvious that the additional layer in the JSI system actually deteriorates the results. On all the three languages, the differences are statistically significant, on Slovene and Croatian on the $p < 0.001$ level, while on Serbian it is on the $p < 0.05$ level. The level of significance of difference between the JANES and JSI-simpler systems is identical to that of between JSI and JSI-simpler.

The most interesting observation from the final evaluation of the submitted and modified systems is that the difference between the traditional CRFs and the (probably over-hyped?) BiLSTMs is actually quite small, with relative error reductions being 15% on Slovene, 5% on Croatian and only 3% on Serbian. These results, as well as some preliminary results on standard test sets, suggest that there would be no significant difference in the results between CRFs and BiLSTMs on standard training and test data.

[5] https://github.com/neulab/dynet-benchmark/blob/master/pytorch/bilstm-tagger-withchar.py

JANES			JSI-simpler		
pred	true	freq	pred	true	freq
Xf	Npmsn	74	Xf	Npmsn	55
Cc	Qo	59	Cc	Qo	55
Ncmsan	Ncmsn	46	Npmsn	Xf	40
Ncmsn	Ncmsan	34	Ncmsan	Ncmsn	34
Ncmsn	Npmsn	31	Rgp	Cs	23
Xf	Npmsan	23	Xf	Ncmsn	24
Rgp	Cs	23	Sl	Sa	23
Npmsn	Xf	23	Ncmsn	Ncmsan	23
Xf	Ncmsn	22	Agpnsny	Rgp	20
Cs	Rgp	20	Sa	Sl	19

Table 4: The ten most frequent confusion pairs for the JANES and the JSI-simpler systems.

4 Error analysis

In this section we perform an analysis of confusion matrices of the JANES and the JSI-simpler system. We perform the analysis on the output of the system on the Croatian test set. We analyze and compare the 10 most frequent confusions for each system, which covers roughly 20% of all errors done by each of the systems. The confusion pairs are given in Table 4. Both systems make similar most frequent mistakes, some of which are typical for morphosyntactic tagging of standard varieties of South Slavic languages, other being more specific for the Twitter variety.

The typical mistakes on the standard language include confusing nominative masculinum common nouns (Ncmsn) for accusative masculinum common nouns (Ncmsan) and vice versa, confusing the word "i" (English "and") in its coordinating conjunction (Cc) and particle (Qo) usage, confusing adverbs (Rgp) for adjectives (Agpnsny for instance) and confusing the word "kada" (English "when") in its subordinative conjunction (Cs) and adverbial (Rgp) usage.

The errors that are more due to the specificity of the Twitter variety are confusing proper names (Np.*) or common nouns (Nc.*) for foreign residuals (mostly foreign words or foreign sequences of words, Xf) and vice versa.

When comparing the most frequent errors between the two systems, the JANES CRF-based system seems to have more problems with the traditional discrimination between different context-dependent cases of nouns, which points to the direction that BiLSTMs are better at modeling long-range dependencies as discriminating between the nominative and the accusative case often requires a very wide context. On the other hand, what the BiLSTM system seems to be worse at is discriminating between different cases for prepositions, which heavily depends on the following adjective or noun. While confusing an accusative preposition (Sa) for a locative one (Sl) the BiLSTM system did 23 times, this happened to the CRF system 17 times. In the opposite direction, the BiLSTM system did 19 mistakes while the CRF system did one mistake less, namely 18 of them. While it is clear why CRFs excel at predicting prepositional cases correctly as this dependence is in the scope of the local features, it seems that the BiLSTMs trade more mistakes in the local context for less mistakes in a wider one.

5 Conclusion

In this paper we have compared two popular sequence labeling techniques: conditional random fields (CRFs) and bidirectional long short-term memories (BiLSTMs) on the task of morphosyntactic annotation of tweets written in three closely related South Slavic languages: Slovene, Croatian and Serbian.

We have shown that CRFs with well defined features come very close to the performance of the stronger BiLSTM models, the difference between those two being bigger as the data are more non-standard. The relative error reduction between those two systems lies between 15% for Slovene, for which the Twitter variety deviates the most from the standard, and 3% for Serbian, for which the Twitter

variety deviates the least.

For the CRF system, we have shown that using contextual, suffixal and distributional features gives very good results. The latter make an inflectional lexicon mostly obsolete, with just minor improvements in accuracy if features from large inflectional lexicons are added.

For the BiLSTM system, we have shown that encoding a character-level representation of a word is the single most useful intervention, with minor improvements obtained through proper word embedding pretraining, fine-tuning on in-domain data and pretraining the character-level encoder on pairs of words and MSD tags from a large automatically tagged web corpus.

With an error analysis we have shown that the types of error performed by each of the systems are actually very similar, most of them still being typical tagger errors for languages with a rich inflectional morphology. However, there is evidence that BiLSTMs resolve long-range dependencies much better, such as discriminating between masculinum nouns in nominative and accusative singular, but yielding slightly more mistakes in the close-range dependencies such as the case of prepositions.

Acknowledgements

The work presented in this paper has been funded by the Slovenian Research Agency national basic research project "Resources, methods and tools for the understanding, identification and classification of various forms of socially unacceptable discourse in the information society" (ARRS J7-8280, 2017-2020), and by the Slovenian research infrastructure CLARIN.SI.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Peter F. Brown, Peter V. Desouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, and Miro Romih. 2015. Morphological lexicon sloleks 1.2. Slovenian language resource repository CLARIN.SI.

Tomaž Erjavec, Nikola Ljubešić, and Nataša Logar. 2015. The slWaC Corpus of the Slovene Web. *Informatica*, 39(1):35–42.

Tomaž Erjavec, Darja Fišer, Jaka Čibej, Špela Arhar Holdt, Nikola Ljubešić, and Katja Zupan. 2017. CMC training corpus Janes-Tag 2.0. Slovenian language resource repository CLARIN.SI.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. HunPos: An open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 209–212, Stroudsburg. Association for Computational Linguistics.

Tobias Horsmann and Torsten Zesch. 2015. Effectiveness of Domain Adaptation Approaches for Social Media PoS Tagging. *CLiC it*, page 166.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}wac - web corpora of bosnian, croatian and serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35. Association for Computational Linguistics.

Nikola Ljubešić, Filip Klubička, and Damir Boras. 2016a. Inflectional lexicon hrLex 1.2. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Filip Klubička, and Damir Boras. 2016b. Inflectional lexicon srLex 1.2. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Tomaž Erjavec, and Darja Fišer. 2017a. Adapting a state-of-the-art tagger for south slavic languages to non-standard text. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 60–68.

Nikola Ljubešić, Tomaž Erjavec, Maja Miličević, and Tanja Samardžić. 2017b. Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.0. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Tomaž Erjavec, Maja Miličević, and Tanja Samardžić. 2017c. Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.0. Slovenian language resource repository CLARIN.SI.

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).

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Maja Miličević and Nikola Ljubešić. 2016. Tviterasi, tviteraši or twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian tweets. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 4(2):156–188.

Maja Miličević, Nikola Ljubešić, and Darja Fišer. 2017. Birds of a feather don't quite tweet together: An analysis of spelling variation in Slovene, Croatian and Serbian Twitterese. In Darja Fišer and Michael Beisswenger, editors, *Investigating Computer-mediated Communication: Corpus-based Approaches to Language in the Digital World*, pages 14–43. Ljubljana University Press, Faculty of Arts.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Nils Reimers and Iryna Gurevych. 2017. Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. *CoRR*, abs/1707.06799.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank Jain. 2018. Language Identification and Morphosyntactic Tagging: The Second VarDial Evaluation Campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Santa Fe, USA.