Adapting a State-of-the-Art Tagger for South Slavic Languages to Non-Standard Text

Nikola Ljubešić[1,2], Tomaž Erjavec[1], and Darja Fišer[3,1]

[1]Dept. of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
[2]Dept. of Information and Communication Sciences, University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb, Croatia
[3]Dept. of Translation, Faculty of Arts, University of Ljubljana
Aškerčeva cesta 2, SI-1000 Ljubljana, Slovenia
{nikola.ljubesic,tomaz.erjavec}@ijs.si
darja.fiser@ff.uni-lj.si

Abstract

In this paper we present the adaptations of a state-of-the-art tagger for South Slavic languages to non-standard texts on the example of the Slovene language. We investigate the impact of introducing in-domain training data as well as additional supervision through external resources or tools like word clusters and word normalization. We remove more than half of the error of the standard tagger when applied to non-standard texts by training it on a combination of standard and non-standard training data, while enriching the data representation with external resources removes additional 11 percent of the error. The final configuration achieves tagging accuracy of 87.41% on the full morphosyntactic description, which is, nevertheless, still quite far from the accuracy of 94.27% achieved on standard text.

1 Introduction

With the rise of social media, the potential from automatically processing the available textual content is substantial. However, there is a series of problems connected to processing Computer Mediated Communication (CMC) due to frequent deviation from the norm (Miličević and Ljubešić, 2016), such as omission of diacritics, non-standard word spellings and frequent use of colloquial expressions. For example, experiments on English part-of-speech tagging showed a drastic loss in accuracy when shifting from Wall Street Journal text (97%) to Twitter (85%) (Gimpel et al., 2011).

Part-of-speech (PoS) tagging is a crucial step in the text processing pipeline, as it gives invaluable information about the grammatical properties of words in context and thus enables, e.g., better information extractions from texts, high quality lemmatization, syntactic parsing, the use of factored models in machine translation etc.

This paper concentrates on adapting a state-of-the art tagger of standard Slovene (Ljubešić and Erjavec, 2016), Croatian and Serbian (Ljubešić et al., 2016) to CMC texts on the example of Slovene language by experimenting with in-domain training data and additional external resources and tools such as word clusters and word normalization.

The rest of the paper is structured as follows: Section 2 gives an overview of the related work on this problem, Section 3 introduces the dataset used, Section 4 describes the tagging experiments we performed, Section 5 reports on the error analysis of the results and Section 6 gives some conclusions and directions for further research.

2 Related Work

Early work on PoS tagging social media was, as usual, mostly focused on English (Gimpel et al., 2011; Owoputi et al., 2013). Recently there has been more work on other languages, primarily through the organization of shared tasks, such the EmpiriST on German (Beißwenger et al., 2016) and PoSTWITA on Italian.[1]

There are two main approaches to processing non-standard data: normalization and domain adaptation (Eisenstein, 2013). Most approaches nowadays follow the domain adaptation path al-

[1] http://corpora.ficlit.unibo.it/PoSTWITA/

Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, pages 60–68,
Valencia, Spain, 4 April 2017. ©2017 Association for Computational Linguistics

though the literature still lacks a detailed comparison of the two strategies on specific tasks.

In domain adaptation there are, again, two main strategies (Horsmann and Zesch, 2015): adding more labeled data (Daumé III, 2007; Hovy et al., 2015) and incorporating external knowledge (Owoputi et al., 2013). Horsmann and Zesch (2015) show that (1) adding manually annotated in-domain data is highly effective (but costly) and (2) adding out-of-domain training data or machine-tagged data is less effective than adding more external knowledge, especially word clustering information.

The contribution of our paper is the following: First, we perform the first experiments in annotating Slavic non-standard texts with part-of-speech and morphosyntactic information, therefore dealing with several hundreds of tags. Next, we investigate the impact of strategies that were proven to be most successful on English, German and Italian on a new language group and level of tag complexity. Last but not least, we release a split of a freely available dataset, as well as the tagger as a useful tool and a strong baseline for other researchers to improve on.

3 CMC Dataset

As the primary resource for training and evaluating our tagger of non-standard language we used the publicly available Janes-Tag v1.2 dataset (Erjavec et al., 2016c), which contains Slovene CMC texts, with the text types being tweets, forum posts, comments on blog posts and comments on news articles. The texts were sampled from the Janes corpus (Fišer et al., 2016), a large corpus (9 million texts with about 200 million tokens) of Slovene CMC. The texts in the Janes corpus are, inter alia, annotated with language standardness scores for each text. These scores were assigned automatically (Ljubešić et al., 2015) and classify texts into three levels of technical and linguistic standardness. Technical standardness (T1, quite standard – T3, very non-standard) relates to the use of spaces, punctuation, capitalization and similar, while linguistic standardness (L1 – L3) takes into account the level of adherence to the written norm and more or less conscious decisions to use non-standard language with respect to spelling, lexis, morphology, and word order. The texts for the Janes-Tag dataset were sampled so that they contain, for each text type, roughly the same number of T1L1, T1L3, T3L3, and T3L3 texts, except for tweets, where only T1L3 and T3L3 texts were included in order to maximize twitter-specific deviations from the norm.

The texts in Janes-Tag were first automatically annotated and then manually checked for the following levels of linguistic annotation: tokenization, sentence segmentation, normalization, part-of-speech tagging and lemmatization. Here normalization refers to giving the standard equivalent to non-standard word-forms, e.g., *jaz* (*I*) assigned to the source *jst*, *js*, *jest* etc., while tagging and lemmatization is then assigned to these normalized forms. It should be noted that two (or more) source word tokens can be normalized to one token or vice versa.

The tagset used is defined in the (draft) MULTEXT-East morphosyntactic specification Version 5[2] for Slovene, which are identical to the Version 4 specifications (Erjavec, 2012), except that four new tags have been added for CMC specific phenomena, such as hashtags and mentions. Version 5 tagset for Slovene defines all together 1900 different tags (morphosyntactic descriptions, MSDs), i.e., it is a fine-grained tagset covering all the inflectional properties of Slovene words.

The dataset is distributed in the canonical TEI encoding as well as in the derived vertical format used by concordancers such as CQP (Christ, 1994). Further details on the dataset can be found in (Erjavec et al., 2016a).

We split the dataset into training, development and testing subsets in a 80:10:10 fashion. We performed stratified sampling over texts with strata being text type and linguistic standardness in order for each subset to have the same distribution of texts given the two variables. This split is also available as part of (Erjavec et al., 2016c). Basic statistics of the dataset and subsets are given in Table 1.

Portion	Texts	Tokens
train	2,370	60,367
dev	294	7,425
test	294	7,484
Σ	2,958	75,276

Table 1: Janes-Tag dataset statistics.

It should be noted that in cases of $n : 1$ or $1 :$

[2] http://nl.ijs.si/ME/V5/msd/

n mappings between the original and normalized word token(s), we consider these in subsequent experiments as one token. The latter also means that one original token is assigned multiple PoS tags, e.g., *meuš → me boš* / `Pp1-sa--y Va-f2s-n`. These phenomena are, however, quite rare, occurring in our CMC dataset on only 0.4% of tokens.

4 Experiments

In this section we present experiments on introducing non-standard training data (4.1), adding word clustering information (4.2), measuring the impact of the standard inflectional lexicon (4.3), adding word normalization data (4.4) and combining standard and non-standard training data (4.5).

4.1 Impact of Non-Standard Data

In the first set of experiments we compare the state-of-the-art tagger for standard Slovene – the ReLDI tagger (Ljubešić and Erjavec, 2016) – with the same tagger implementation retrained on the training portion of the Janes-Tag dataset.

The ReLDI tagger is based on conditional random fields and uses the following features:

1. lowercased tokens at positions $\{-3, -2, ..., 3\}$;
2. focus token (token at position 0) suffixes of length $\{1, 2, 3, 4\}$;
3. tag hypotheses obtained from an inflectional lexicon for tokens at positions $\{-2, -1, ..., 2\}$;
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 uppercase followed by at least two lowercase characters at the start of the sentence.

For obtaining tag hypotheses for Slovene, we use, just as in the standard setting, the Sloleks lexicon (Dobrovoljc et al., 2015).

We evaluate each of the our configurations on the development portion of Janes-Tag via accuracy on two levels:

1. the fine-grained tagset, which contains the complete morphosyntactic descriptions (MSDs): the MSD tagset comprises 960 different labels in the Janes-Tag dataset; and
2. the coarse-grained tagset, comprising only the first two letters of the MSD, i.e., covering the part-of-speech and, typically, its type (e.g., common vs. proper noun): we term this the PoS tagset, and it comprises 42 different labels in Janes-Tag.

The results of this experiment are presented in the first part of Table 2. The standard tagger (configuration `reldi`) shows very poor performance, especially given its results on standard data (94.27% MSD accuracy and 98.94% PoS accuracy). Simply training the tagger on the ∼60k tokens of in-domain training data (configuration `reldi+janestag`), as opposed to the 500k tokens of training data in the standard configuration, improves the tagger drastically, although its performance still does not come near the performance on standard data.

We also experimented with extending the feature set with features encoding whether the token is a hashtag, mention or URL similar to Gimpel et al. (2011), but did not obtain any improvements.

In the following experiments we refer to the `reldi+janestag` configuration for brevity as the `janes` configuration.

At this point our experiments could continue in two directions: (1) combining standard and non-standard training data or (2) enriching the process with external knowledge. Given the non-negligible size of our non-standard training subset, we decided to first focus on enriching the process with external knowledge and focus on combining the two types of training data at a later stage.

Configuration	MSD	PoS
reldi	68.67	73.13
reldi+janestag	84.15	89.85
janes+brown.web	85.17	91.12
janes+brown.cmc	85.51	91.31
janes+brown.all	85.70	91.52
janes-lex	81.14	87.62
janes+brown.all-lex	84.18	91.04
hunpos+janestag	83.78	89.70
hunpos+janestag-lex	80.65	87.66
janes+brown.all+normlex	86.03	91.65
janes+brown.all+normcsmt	86.28	91.72
janes+brown.all+normgold	87.97	93.19

Table 2: Results in accuracy on the first four sets of experiments.

4.2 Adding Word Clustering Information

In this set of experiments we investigate the improvements that can be obtained by introducing knowledge from word clusters calculated on large amounts of non-annotated texts. The word clustering technique that has recently shown best results for enriching various decision processes (Turian et al., 2010; Owoputi et al., 2013; Horsmann and Zesch, 2015) are Brown clusters (Brown et al., 1992). We calculate this hierarchical clustering representation of words given their context on three different sources: (1) the 1 billion token slWaC v2.0 web corpus of Slovene (Erjavec et al., 2015) (`brown.web`), (2) the 200 million token Janes v0.4 corpus (Fišer et al., 2016) of Slovene CMC (`brown.cmc`) and (3) a concatenation of the two corpora (`brown.all`). On each resource we build 2000 clusters from words occurring at least 50 times.

We additionally experiment with four different and common ways of including the binary hierarchical clustering information in our tagger: adding the feature corresponding to the focus tokens' (1) whole binary path, (2) each length of the binary path prefix, (3) even lengths of path prefixes (Owoputi et al., 2013) and (4) path prefixes of length 2^n, $n \in \{1, 2, 3, 4\}$ (Plank et al., 2014). Among the four approaches, the one including even path lengths only (3) proved to yield just slightly (up to half percent), but consistently better results than the remaining three approaches (1, 2, 4).

We report the results of using Brown binary paths of even lengths with different resources (`brown.web`, `brown.cmc`, `brown.all`) in the second part of Table 2. When comparing the bare configuration trained on non-standard data (`reldi+janestag`) with the configurations extended with various Brown clusters, we measure an improvement on MSD accuracy of 1.02% to 1.55% and an improvement on PoS accuracy of 1.27% to 1.67%. The results across our experiments consistently show that Brown clusters improve PoS accuracy more than MSD accuracy. This is to be expected as the large number of different MSD tags comes close to the overall number of clusters.

The differences in the results given the source used to calculate Brown clusters are minor but consistent with an increase in quality (`brown.cmc`) and quantity (`brown.web`) of the underlying data. While the Janes clusters perform better than the slWaC ones regardless of the significantly bigger size of the slWaC corpus, the best results are obtained with clusters calculated from a concatenation of the two resources.

4.3 Impact of the Inflectional Lexicon

In this set of experiments we measure the impact of the inflectional lexicon on the tagging process. As stated before, the ReLDI tagger, as well as the `janes` configuration, use the Sloleks inflectional lexicon (Dobrovoljc et al., 2015) containing 100 thousand lexemes (lemmas) with 2.7 million word-forms. We perform the following experiments as it is not infrequent that even though large inflectional lexicons do exist for Slavic languages, they are not (freely) available.

We investigate two scenarios: (1) training the ReLDI tagger on non-standard data without an inflectional lexicon (`janes-lex`) and (2) training the ReLDI tagger on non-standard data and previously best-performing Brown clusters without the inflectional lexicon (`janes+brown.all-lex`). With the second scenario we investigate to what extent the lack of an inflectional lexicon can be compensated with word clusters.

To obtain a comparison with a configuration not relying on the ReLDI tagger, in this set of experiments we additionally report the results obtained with the HunPos tagger (Halácsy et al., 2007), a tagger giving very good results on Slavic languages (Agić et al., 2013), trained on the Janes-Tag training subset with (configuration `hunpos+janestag`) and without the inflectional lexicon (configuration `hunpos+janestag-lex`).

The results in the third section of Table 2 show that the lack of an inflectional lexicon (`janes-lex`) deteriorates MSD accuracy by 3% and PoS accuracy by 2.2%. Adding Brown clusters into the configuration (`janes+brownall-lex`) generates MSD accuracy as high as when using an inflectional lexicon (`reldi+janestag`) and even improves PoS accuracy by 1.2%, which is in line with our previous observation on a greater impact of Brown clusters on PoS accuracy than MSD accuracy. However, this configuration still performs worse than the one using both the inflectional lexicon and Brown clusters, loosing 1.5% MSD accuracy and

0.5% PoS accuracy.

The results obtained with the HunPos tagger are very much in line with the results obtained with the ReLDI tagger. In both configurations, with (`hunpos+janestag` is to be compared to `reldi+janestag`) and without the inflectional lexicon (`hunpos+janestag-lex` is to be compared to `janes-lex`), the ReLDI tagger is half a percent better on MSD accuracy and just slightly better on PoS accuracy. A similar but stronger trend was measured on standard data (Ljubešić et al., 2016). The better performance of the ReLDI tagger is probably due to its stronger modeling technique, while the smaller difference in comparison with the comparative experiments on standard Slovene is most likely the result of the nine times smaller training dataset.

4.4 Adding Normalization Data

Another potentially useful resource for tagging non-standard Slovene texts is the Slovene dataset of normalized CMC texts, Janes-Norm 1.2 (Erjavec et al., 2016b) which is a superset of Janes-Tag. In each of the following experiments we use only the part of Janes-Norm which is not included in Janes-Tag. This portion of Janes-Norm is slightly above 100 thousand tokens in size.

The following experiments investigate whether additional improvements can be obtained by introducing normalization information to our classification process.

In the first experiment (configuration `janes+brown.all+normlex`) we use the available normalization data as a normalization lexicon consisting of original word forms and their normalized counterparts. We extend the tagger feature set with MSD hypotheses of all normalized forms. The MSD hypotheses are obtained from the Sloleks inflectional lexicon.

In the second experiment (configuration `janes+brown.all+normcsmt` we train the cSMTiser.[3] normalization tool which was already been used for normalizing Slovene user-generated and historical data (Ljubešić et al., 2016) as well as Swiss dialectal data (Scherrer and Ljubešić, 2016). The tool is based on character-level statistical machine translation and is in this case trained on pairs of tokens, not pairs of sentences, as the two approaches yield very similar results on Slovene CMC texts (Ljubešić et al., 2016). Once the tool is trained, a lexicon similar to the one used in the first experiment is produced with the difference that (1) each token has just one normalization and (2) all tokens in the training and development set are covered in that lexicon. The feature set is extended as in the first experiment.

Given that we have the gold normalization available in our Janes-Tag dataset, we also calculated a ceiling for this tagger extension (configuration `janes+brown.all+normgold`) which uses the gold normalization for calculating the feature extension.

The results are presented in the final part of Table 2. Both automated approaches improve the previous best results (configuration `janes+brown.all`), the CSMT approach slightly outperforming the lexicon approach. However, the gold normalization approach shows that there is still room for improvement of 1.5% on both MSD and PoS levels. There are two possible reasons for this rather large gap: (1) in our two automated approaches we discard the context and (2) the same words that are hard to normalize are those that are hard to part-of-speech tag. The first issue could be partially resolved by training a sentence-level normalizer which is processing-wise much more costly, but does yield $\sim 10\%$ token error reduction as long as the texts are significantly non-standard (Ljubešić et al., 2016). The second issue could be only resolved with much more training data or better unsupervised techniques than Brown clustering.

4.5 Combining Standard and Non-Standard Training Data

In the final set of experiments we investigate the impact of combining existing standard training data with the newly developed non-standard data. We compare that impact on two configurations from our previous experiments: (1) the `reldi+janestag`, i.e., the `janes` configuration which is trained on Janes-Tag and does not use any external knowledge except the inflectional lexicon and (2) the `janes+brown.all+normdict` configuration which additionally uses Brown clusters and the normalization lexicon. We call the second configuration `janes+`.

We discard the configuration using cSMTiser (`janes+brownall+normcsmt`) since its improvement is minor and it makes the tagging pro-

[3] `https://github.com/clarinsi/csmtiser`

	janes		janes+	
nstd:std	**MSD**	**PoS**	**MSD**	**PoS**
-	84.15	89.85	86.03	91.65
1:10	86.05	90.51	87.38	91.77
1:5	85.98	90.49	87.70	91.97
1:3	86.32	90.77	**87.70**	**92.22**

Table 3: Results in accuracy on combining standard and non-standard training data.

cess dependent on one external tool.

We additionally investigate the impact of over-representing non-standard data by repeating the non-standard dataset once, twice and three times, yielding the ratio of non-standard and standard data of 1:10, 1:5 and 1:3. Further increases of the ratio of non-standard data did not generate any improvements, hence we do not report them.

The results of this set of experiments are given in Table 3. Adding standard training data has an overall positive impact, which is much greater on the basic configuration due to the lack of external resource supervision. However, the configuration using Brown clusters and the normalization lexicon always outperforms the basic configuration. Furthermore, over-representing non-standard data two or three times improves the results of the `janes+` configuration while the results of the `janes` configuration are rather constant. This makes sense as more non-standard data enables the tagger to properly weigh the features using non-standard external knowledge.

In the 1:3 ratio of non-standard and standard data, the `janes+` configuration outperforms the `janes` configuration by 1.4% for MSD accuracy and 1.5% for PoS accuracy. We tested whether these obtained differences are statistically significant with the McNemar's test for paired nominal data (McNemar, 1947). On the MSD level the obtained p-value was $2.57 * 10^{-9}$ while on the PoS level the p-value was $1.32 * 10^{-11}$.

Similarly, both the difference between the `janes` configuration not using and using standard data, as well as between the `janes+` configuration not using and using standard data have proven to be statistically significant with $p < 0.001$ on the MSD level. On the PoS level the difference between using and not using standard data gave $p = 0.001$ for the `janes` configuration and $p = 0.02$ for the `janes+` configuration.

5 Error Analysis

In order to gain more insight into the tagger behavior in various experimental settings, hence to better contextualize the results obtained in automatic evaluation as well as collect information useful for future improvements of the tagger, we performed manual evaluation of the erroneously tagged instances on the part-of-speech level.

Three types of the main sources of errors were observed: (1) non-standard lexis (e.g., *žvajzne* instead of the standard *udari*, Eng. *hit*), (2) non-standard word forms (e.g., *najsuperejši* instead of the standard *najbolj super*, Eng. *the greatest*), and (3) non-standard spelling (e.g., *uredu* instead of the standard *v redu*, Eng. *all right*).

In the manual error analysis, three experimental configurations were compared: (1) the original ReLDI tagger (`reldi`), (2) the ReLDI tagger trained on ssj500k and three times over-represented Janes-Tag (here referred to is `janes`) and (3) the ReLDI tagger trained on the same data as `janes` with the feature set extended with Brown clusters and the normalization lexicon (here referred to as `janes+`). The results of these three configurations on the test portion of the Janes-Tag dataset are presented in Table 4. We again check whether the difference between the `janes` and `janes+` configuration is statistically significant with the McNemar's test, obtaining a p-value of $1.53 * 10^{-10}$ on the MSD level and a p-value of $9.49 * 10^{-15}$ on the PoS level.

configuration	**MSD**	**PoS**
reldi	67.73	72.41
janes	85.85	90.22
janes+	87.41	91.98

Table 4: Results in accuracy of the three final configurations on the test portion of the dataset.

We first analysed the five most frequent errors in the `reldi` configuration, which represent 26% of all the errors of that configuration, and compared them with the `janes` and `janes+` configurations.

The most frequent error (which represented 7% of all the errors of that configuration) was the erroneous tagging of punctuation as abbreviations. An inspection of the erroneously tagged instances quickly revealed that this error was due to the non-standard multiplication of punctuation that was

not observed in the training data of standard language.

The second most frequent error (which represented nearly 7% of all the errors) was the mistagging of mentions of user accounts in tweets as foreign words, which is hardly surprising as they too did not exist in the standard training data.

On third place (representing 5% of all the errors) are verbs erroneously tagged as foreign language elements, which were mostly due to non-standard spelling (e.g., *prlezla* instead of *prilezla*, Eng. *climbed*) and lexis (e.g., *šprehal* instead of *govoril*, Eng. *spoke*).

Coming fourth (comprising 4% of all the errors) are the verbs mistagged as common nouns, which too is mostly due to non-standard spelling (e.g., *morm* instead of *moram*, Eng. *must*) and lexis (e.g., *fura* instead of *vozi*, Eng. *drives*).

The fifth, and last type of errors with a substantial 3% share of all the errors are misattributions of adverbs as common nouns, again mostly due to non-standard spelling (e.g., *lohk* instead of *lahko*, Eng. *easily*).

Next, we checked how these five most common errors in the original `reldi` configuration fare in the `janes` and `janes+` configurations. The analysis shows that the first two types of errors (non-standard punctuation and mentions) disappear in both settings because the phenomena were now adequately represented in the training data. In a similar vein, the error in mistagged verbs as foreign words and general adverbs as common nouns decreases 10-fold in both configurations. The mistagging of verbs as common nouns drops 3 times in `janes` and 5 times in `janes+`, the difference between the two going back to more observed examples of the non-standard spelling instances in the additional resources, the Brown clusters and the normalization lexicon.

In the third part of the manual error analysis we examined the most frequent errors in the `janes` and `janes+` configurations. The most frequent type of errors (which represents roughly 4% of all the errors in both configurations) was the mistreatment of proper nouns as common ones due to non-standard capitalization and Twitter-specific abbreviations. In `janes`, the second most frequent error type (which represents 4% of all the errors) was the mistagging of verbs as common nouns for the same reasons as in the `reldi` configuration explained above. The third error type in `janes` and second in `janes+` (comprising 3% of all the errors in both configurations) is the mistagging of adjectives as adverbs, which is a typical tagging error also for standard language. The fourth and fifth most frequent errors in `janes` are the erroneous tagging of foreign words as either proper or common nouns, which however sees a 25% decrease in `janes+` due to additional lexical supervision through Brown clusters.

6 Conclusions

The point of departure was the finding that applying a standard tagger to non-standard language results in a loss in accuracy almost comparable to results on English, more than doubling the amount of error. However, in the paper we have shown that retraining a standard tagger on 60 thousand tokens of non-standard data improves the results drastically.

Additional improvements can be made, primarily by (1) combining non-standard and standard training data (if a large amount of standard training data is available), (2) adding Brown clustering information and (3) adding any additional sort of relevant information, in our case word normalization information.

With a set of systematic experiments we have shown that Brown clusters improve coarse-grained tagging more than the fine-grained one, and that the tagging accuracy on PoS level improves more with Brown clusters than with adding 500k tokens of standard training data, while adding the given amount of standard training data achieves greater improvements on the MSD level. As future work, for enriching processes that have to distinguish between multiple hundreds of classes, a soft word clustering technique should be investigated.

We have observed a positive impact of both quality and quantity of the data used for calculating Brown clusters on the final tagging performance. While smaller amounts of in-domain data achieve better results than large amounts of out-of-domain data, merging these two yields the best results.

Using a large standard inflectional lexicon indirectly, through features, has a significant impact on the final tagging accuracy. A lack of such a resource can be compensated with Brown clusters, fully regarding MSD accuracy and even improving PoS accuracy. However, having both resources at ones' disposal generates the best results.

Finally, word normalization information can visibly improve the results by introducing MSD hypotheses of the normalized word forms in form of features.

While simply retraining the tagger on a combination of standard and non-standard training data removes more than half of the error of the standard tagger, adding additional features relying on external resources such as Brown clusters and word normalization removes additional 11% of the tagging error.

A practical contribution of the paper is that we make the data split[4] (Erjavec et al., 2016c) and the tagger[5] available. We expect the tagger to be used both as the currently best tagger for non-standard Slovene, as well as a strong baseline for future improvements on the problem.

We are currently finalizing datasets consisting of Croatian and Serbian tweets, prepared in a comparable fashion to Janes-Norm and Janes-Tag, and plan to add models for these two languages to the developed tagger in the near future.

Acknowledgments

The work described in this paper was funded by the Slovenian Research Agency national basic research project J6-6842 "Resources, Tools and Methods for the Research of Nonstandard Internet Slovene", the national research programme "Knowledge Technologies", by the Ministry of Education, Science and Sport within the "CLARIN.SI" research infrastructure and the Swiss National Science Foundation grant IZ74Z0 160501 (ReLDI).

References

Željko Agić, Nikola Ljubešić, and Danijela Merkler. 2013. Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 48–57, Sofia, Bulgaria, August. Association for Computational Linguistics.

Michael Beißwenger, Sabine Bartsch, Stefan Evert, and Akademie der Wissenschaften. 2016. EmpiriST 2015: A shared task on the automatic linguistic annotation of computer-mediated communication and web corpora. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task. Berlin, Germany*, pages 44–56.

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.

Oliver Christ. 1994. A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX 94: 3rd Conference on Computational Lexicography and Text Research*, pages 23–32.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Conference of the Association for Computational Linguistics*.

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, and Miro Romih. 2015. Morphological lexicon Sloleks 1.2. `http://hdl.handle.net/11356/1039`.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Tomaž Erjavec, Nikola Ljubesić, and Nataša Logar. 2015. The slWaC corpus of the Slovene Web. *Informatica*, 39(1):35.

Tomaž Erjavec, Jaka Čibej, Špela Arhar Holdt, Nikola Ljubešić, and Darja Fišer. 2016a. Gold-standard datasets for annotation of Slovene computer-mediated communication. In *Proceedings of RASLAN 2016: Recent Advances in Slavonic Natural Language Processing*, pages 29–40. Brno: Tribun EU.

Tomaž Erjavec, Darja Fišer, Jaka Čibej, and Špela Arhar Holdt. 2016b. CMC training corpus Janes-Norm 1.2. `http://hdl.handle.net/11356/1084`.

Tomaž Erjavec, Darja Fišer, Jaka Čibej, Špela Arhar Holdt, and Nikola Ljubešić. 2016c. CMC training corpus Janes-Tag 1.2. `http://hdl.handle.net/11356/1085`.

Tomaž Erjavec. 2012. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, 46(1):131–142. DOI: 10.1007/s10579-011-9174-8.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2016. JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin (Janes v0.4: Corpus of Slovene User Generated Content). *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research*, 4(2):67–99.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan,

[4] `http://hdl.handle.net/11356/1085`

[5] `https://github.com/clarinsi/janes-tagger`

and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics.

Péter Halácsy, András Kornai, and Csaba Oravecz. 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.

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

Dirk Hovy, Barbara Plank, Hector Martinez Alonso, and Anders Søgaard. 2015. Mining for unambiguous instances to adapt PoS taggers to new domains. In *Proceedings of NAACL*. Association for Computational Linguistics.

Nikola Ljubešić, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak, and Iza Škrjanec. 2015. Predicting the Level of Text Standardness in User-Generated Content. In *Proceedings of Recent Advances in Natural Language Processing*.

Nikola Ljubešić, Katja Zupan, Darja Fišer, and Tomaž Erjavec. 2016. Normalising Slovene data: historical texts vs. user-generated content. In *Proceedings of KONVENS*.

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.

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.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, Ryan T. McDonald, and Anders Søgaard. 2014. Adapting taggers to Twitter with not-so-distant supervision. In *COLING*, pages 1783–1792.

Yves Scherrer and Nikola Ljubešić. 2016. Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. In *Proceedings of Konferenz zur Verarbeitung natürlicher Sprache (KONVENS), Bochum, Germany*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.