Lemmatising Serbian as Category Tagging with Bidirectional Sequence Classification

Andrea Gesmundo, Tanja Samardžić

Computer Science Department, Linguistics Department
University of Geneva, Switzerland
Andrea.Gesmundo@unige.ch, Tanja.Samardzic@unige.ch

Abstract

We present a novel tool for morphological analysis of Serbian, which is a low-resource language with rich morphology. Our tool produces lemmatisation and morphological analysis reaching accuracy that is considerably higher compared to the existing alternative tools: 83.6% relative error reduction on lemmatisation and 8.1% relative error reduction on morphological analysis. The system is trained on a small manually annotated corpus with an approach based on Bidirectional Sequence Classification and Guided Learning techniques, which have recently been adapted with success to a broad set of NLP tagging tasks. In the system presented in this paper, this general approach to tagging is applied to the lemmatisation task for the first time thanks to our novel formulation of lemmatisation as a category tagging task. We show that learning lemmatisation rules from annotated corpus and integrating the context information in the process of morphological analysis provides a state-of-the-art performance despite the lack of resources. The proposed system can be used via a web GUI that deploys its best scoring configuration.

Keywords: lemmatisation, morphological analysis, tagging, sequence classification, Serbian, corpus

1. Introduction

Rapidly increasing availability of digital texts all over the world has facilitated creation of textual resources for languages that are traditionally considered as low resource languages. Tools for automatic processing of these newly available resources, however, are often missing. An example of such a language is Serbian, with several electronic corpora created recently. Apart from the referential corpus and a historical corpus created at the University of Belgrade (Vitas et al., 2003; Kostić, 2001), there is a web corpus of Serbian (available at `http://www.sketchengine.co.uk`, as well as parallel corpora included in the SE Times Corpus (available at `http://elx.dlsi.ua.es/~fran/SETIMES`) and ParaSol (available at `http://parasol.unibe.ch`). None of these corpora contain any kind of linguistic annotation, which is crucial for these resources to be adequately exploited.

Lemmatisation and morphosyntactic tagging[1] are basic steps in automatic processing of language corpora. Texts containing this information can be directly used in lexicographic work and in linguistic research. This annotation is also often a prerequisite for developing systems for more sophisticated automatic processing such as syntactic and semantic parsing, information retrieval, etc. The work on processing morphologically rich languages suggests that using comprehensive morphological dictionaries is necessary for achieving good results (Hajič, 2000; Erjavec and Džeroski, 2004). However, such dictionaries are constructed manually and they cannot be expected to be developed quickly for many languages.

Despite different projects that resulted in resources that can be used for developing tools for automatic morphological processing, such as electronic dictionaries and manually annotated texts, there are no tools that can process Serbian texts with a state-of-the-art performance. Applying standard or general tools to Serbian results in a performance well below the performance achieved for other languages (Juršič et al., 2010; Popović, 2010). This is explained by scarce resources and rather complex morphology.

In this paper, we present a new approach to the task of lemmatisation and morphosytactic tagging that reaches much better performance than the existing approaches, using the limited resources that are available.

2. Lemmatisation as a Category Tagging Task

In the task of lemmatisation, each instance of a word in a text is assigned a lemma so that different inflected word forms are identified as instances of the same lemma. The task of morphosyntactic tagging is to assign to each instance of a word in a text a label that represents the morphosyntactic categories that the instance realises. Automatic lemmatisation and tagging require defining models that can determine the lemma and the morphosyntactic label for a given word. These two tasks are usually resolved separately, but morphosyntactic information (usually the part-of-speech label) is often used in lemmatisation.

The only lemmatisation system that has been tested on Serbian is the multilingual automatic lemmatiser by Juršič et al. (2010). This system learns the rules of transforming words to lemmas from a list of examples, reversely sorted word forms and their corresponding lemmas. It first defines the most common suffix for all the examples and the most frequent transformation of the suffix. More specific rules (regarded as exceptions to the general rule) are learned by iterative grouping of the examples with increasingly longer

[1] Morphologically rich languages, such as Serbian, require complex morphosyntactic tagging, where the tags encode multiple morphosyntactic categories of words including lexical category (part-of-speech), gender, aspect, case, tense etc.

Word	Lemma	MSD	PoS	Lem-tag
propagirao	propagirati	Vmps-sman-n—p	Vm	1+ti
je	jesam	Va-p3s-an-y—p	Va	0+sam
svoju	svoj	Ps-fsa	Ps	1+
jeres	jeres	Ncfsa–n	Nc	0+
,	#	#	#	#
zanosio	zanositi	Vmps-sman-n—p	Vm	1+ti
se	se	Q	Q	0+
njome	ona	Pp3fsi	Pp	5+ona

Table 1: An example of an annotated sentence with induced lemma category tags.

Lemmatisation			
	LemmaGen	TnT	BTagger
All	86.10	—	**97.72**
Known	—	—	99.51
Unknown	—	—	84.98
Morphosyntactic tagging			
	LemmaGen	TnT	BTagger
All	—	85.47	**86.65**
Known	—	**93.86**	90.00
Unknown	—	58.38	**62.19**

Table 2: The accuracy rates of BTagger compared with other tools.

suffixes. To deal with the ambiguity of word forms, the examples are distributed into separate lists so that each list contains the examples with the same morphosyntactic tag and the transformation rules are learned for each list separately.

In our approach, the lemmatisation task is redefined as a category tagging task. Each word in a text is assigned a tag that encodes the transformation details needed to transform the word into its lemma. The transformation from a word to a lemma is done in two steps: 1) remove a suffix of length *n* from the word form; 2) add a new lemma suffix. Therefore, each label consists of two parts: the first is the integer *n* and the second is the new suffix. For example, to transform the Serbian conjugated verb *zanosi-o* into its lemma *zanosi-ti* we have to remove a suffix of length 1 and add the suffix "-ti". This means that the word form *zanosio* is assigned the label "1+ti".

With this approach, it is possible to cluster the words that follow a regular pattern of transformation into a single class, while a specialised label is learnt for frequent irregular words. The advantage of structuring the lemmatisation task as a tagging task is that it allows us to apply successful tagging techniques and use the context information in assigning transformation labels to the words in a text.

3. Bidirectional Sequence Classification for Tagging

Having reformulated the lemmatisation task, we perform both morphosyntactic tagging and lemmatisation using the same tagger based on the Bidirectional Tagger presented in Shen et al. (2007). We chose this model since it has been shown to be easily adaptable for solving a wide set of tagging and chunking tasks obtaining state-of-the-art performances with short execution time (Gesmundo, 2011; Gesmundo, 2009a; Gesmundo, 2009b). Furthermore, this model has consistently shown good generalisation behaviour reaching significantly higher accuracy in tagging unknown words than other systems.

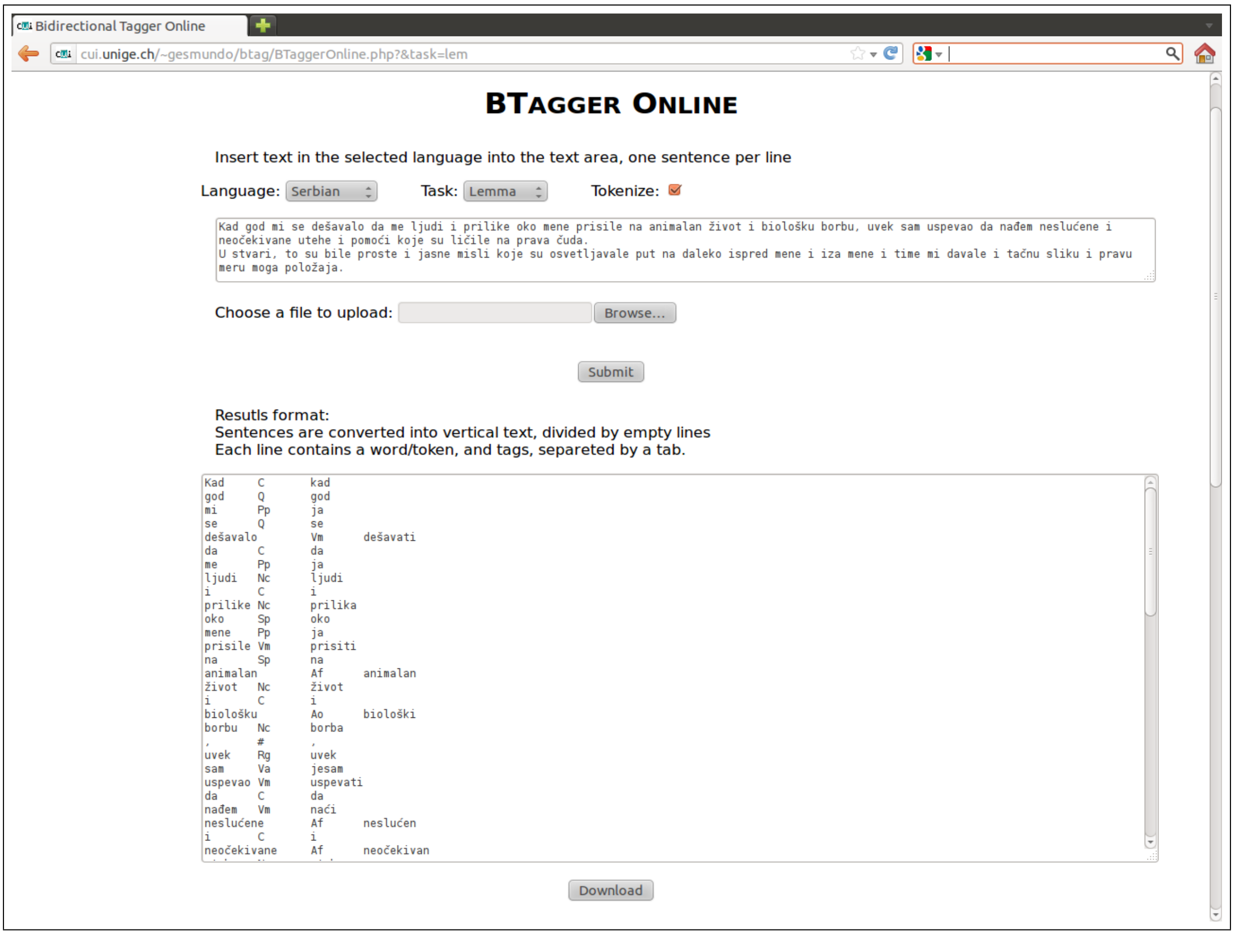
The training framework applied is Guided Learning. This framework is more complex than simple supervised learning. The system learns the parameters for the local classifier from examples of word forms and their labels given in the gold standard annotation, but it has no indications on the order of inference. The specificity of this framework lies in the ability to learn the order of inference together with the parameters of the local classifier instead of using a predefined order (usually left-to-right). Guided Learning follows an "easiest-first" approach allowing any order of inference. The task of assigning a sequence of labels to a sequence of words is performed iteratively, starting from easy decisions. The easily identifiable labels are first assigned to the words in the sequence for which such labels are available. These labels are then used in later iterations to disambiguate postponed difficult tagging decision. In this way right-context, left-context and bidirectional-context features can be used at little extra cost.

The training algorithm is based on the Averaged Perceptron Algorithm (Collins, 2002; Freund and Schapire, 1999). Basing the learning algorithm on the perceptron scheme allows us to keep a low system complexity and moderate execution time, without sacrificing learning capability and quality of the results.

4. Experimental Evaluation

We perform an experimental evaluation of our system using the same corpus that has been used in the other reported experiments, the Serbian translation of G. Orwells "1984" manually annotated within the Multext-East project (Erjavec, 2010; Krstev et al., 2004). It contains 108805 tokens, with 8392 annotated lemmas and 906 morphosyntactic labels. We use the corpus not only to train and test the tagging system but also to induce the lemmatisation category labels as described in Section 2.. We use 80% of the sentences in the corpus for training and 20% for testing.

An example of an annotated sentence is given in the first three columns of Table 1. The morphosyntactic labels, called morphosyntacic definitions (MSD, the second column in Table 1), are compact representations of a number of lexical, morphological, and syntactic categories realised in each word form. Each category is encoded by a single character in the label. The first character encodes the part-of-speech (verb, noun, adjective, etc.). The second character encodes a subclassification for each main category (e.g. main, auxiliary, modal, copula for verbs, common, proper for nouns etc.). Other characters specify morphological features that are marked in the word form such as case, number, tense, voice, mood etc. For example, the MSD label "Ncfsa–n" denotes that the word *jeres* is a common noun, feminine gender, used in singular accusative case, not marked for definiteness and the status of a clitic, not animate. Detailed specifications of the labels are provided in the Multext-East project documentation.


Bidirectional Tagger Online
cui.unige.ch/~gesmundo/btag/BTaggerOnline.php?&task=lem

BTagger Online

Insert text in the selected language into the text area, one sentence per line

Language: Serbian Task: Lemma Tokenize: ☑

```
Kad god mi se dešavalo da me ljudi i prilike oko mene prisile na animalan život i biološku borbu, uvek sam uspevao da nađem neslućene i
neočekivane utehe i pomoći koje su ličile na prava čuda.
U stvari, to su bile proste i jasne misli koje su osvetljavale put na daleko ispred mene i iza mene i time mi davale i tačnu sliku i pravu
meru moga položaja.
```

Choose a file to upload:

Resutls format:
Sentences are converted into vertical text, divided by empty lines
Each line contains a word/token, and tags, separeted by a tab.

```
Kad	C	kad
god	Q	god
mi	Pp	ja
se	Q	se
dešavalo	Vm	dešavati
da	C	da
me	Pp	ja
ljudi	Nc	ljudi
i	C	i
prilike	Nc	prilika
oko	Sp	oko
mene	Pp	ja
prisile	Vm	prisiti
na	Sp	na
animalan	Af	animalan
život	Nc	život
i	C	i
biološku	Ao	biološki
borbu	Nc	borba
,	#	,
uvek	Rg	uvek
sam	Va	jesam
uspevao	Vm	uspevati
da	C	da
nađem	Vm	naći
neslućene	Af	neslućen
i	C	i
neočekivane	Af	neočekivan
```

Figure 1: Sample screenshot of the Serbian morphological analyser on-line GUI.

Training and testing is performed using a set of features extracted from a window context of size 5 centred on the target word. In the task of lemmatisation, we use the following features: two words before and two after the current word, the predicted part-of-speech labels (the fourth column in Table 1) and the predicted lemma category labels (the fifth column in Table 1) for the words in the window for which these labels are available. The part-of-speech labels are extracted from the morphosyntactic tags provided in the manual annotation in the corpus. They correspond to the first two characters of the tags (the part-of-speech and the subcategorisation characters). Both the lemma category labels that specify the transformation of the word form into its lemma, as defined in our approach, and the part-of-speech labels are learnt from the examples in the training corpus. The learning resulted in 364 different lemma labels and 28 part-of-speech labels.

The features used in the task of morphosyntactic tagging are two words before and two after the current word, and the complete morphosyntactic labels of these words (the third column in Table 1).

Table 2 shows the accuracy rates achieved by our system (BTagger) in the tasks of lemmatisation and morphosyntactic tagging. The results are compared with the LemmaGen lemmatiser (Juršič et al., 2010), the only freely available lemmatizer tested on Serbian, and with the TnT tagger, which has been reported to have the best performance on Serbian (Popović, 2010).

As it can be seen in Table 2, BTagger achieves the accuracy rates that are higher than those of the other tools on both tasks. There is a big improvement on the lemmatisation task. BTagger's accuracy rate of 97.73% is compared to the 86.10% accuracy rate of LemmaGen. The error reduction rate on this task is 81%. Even though this information is not available for LemmaGen, we also report the performance of BTagger evaluated on unknown (84.98%) and known (99.51%) words separately, since these two categories are often considered in evaluation of lemmatisers. The performance of BTagger on this task is not just much better than previously achieved results on Serbian, but it is also comparable with the state-of-the art performances for other languages. It is important to stress that this is achieved using only a small manually annotated corpus, with no language specific external sources of data such as independent morphological dictionaries, which have been considered necessary for efficient processing of morphologically rich languages.

In the morphsyntactic tagging task, we compare the results of BTagger with the results of the TnT tagger. The error reduction rate on this task is 8.12% overall. While the performance on known words is below the TnT tagger, the improvement is obtained on unknown words (error reduction rate 9.15%). The overall improvement in this task is much smaller than in the case of lemmatisation, but the results are still interesting because they confirm the good generalisation behaviour of BTagger and show the advantage of this tagger over the best performing alternative system.

5. On-line Interactive Graphic User Interface

A pre-trained version of the system is accessible on-line at `http://cui.unige.ch/~gesmundo/btag/BTaggerOnline.php`. The system deployed for this application is trained on the full corpus with the feature set described in Section 4.. An interactive web-browser user interface enables morphological analysis of an unlimited number of sentences typed in by the user.

A sample screenshoot of the web GUI is shown in Figure 1. It displays an example of the user input and the result of the morphological analysis.[2] The input sentences are typed into the interface form one sentence per line with words/tokens separated by spaces. Alternatively, the user can choose to separate the tokens automatically by checking the Tokenize option, as shown in the upper part of Figure 1. The input sentences can also be uploaded from a file using the upload box. The user can view the result of the morphological analysis directly in the text area that appears below the Submit button, as shown in the lower half of Figure 1 or download it as a text file file by clicking on the Download button at the bottom of the web page.

The output is displayed in the standard column format: sentences are separated by empty lines, each line contains a word, predicted part-of-speech tag, and predicted lemma separated by a tab.

6. Conclusion

We have shown that redefining the task of lemmatisation as a category tagging task and using an efficient tagger to perform it on Serbian results in a great improvement in the performance compared to the previous approaches. The adaptive general classification model used in our approach to Serbian makes use of rich contextual features overcoming the lack of resources, which presented an obstacle for the other approaches. For this reason, it can be expected to be easily portable across languages enabling good quality processing of languages with complex morphology, with no need for comprehensive, manually constructed linguistic resources.

7. Acknowledgements

The work described in this paper was partially funded by the Swiss National Science Foundation grants CRSI22 127510 (COMTIS) and 122643.

8. References

Michael Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of The 2002 Conference on Empirical Methods on Natural Language Processing*, pages 1–8, Philadelphia, USA. Association for Computational Linguistics.

Tomaž Erjavec and Sašo Džeroski. 2004. Machine learning of morphosyntactic structure: lemmatizing unknown Slovene words. *Applied Artificial Intelligence*, 18:17–41.

Tomaž Erjavec. 2010. MULTEXT-East version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 2544–2547, Valletta, Malta. European Language Resources Association (ELRA).

Yoav Freund and Robert E. Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine Learning*, 37(3):277–296.

Andrea Gesmundo. 2009a. Bidirectional sequence classification for part of speech tagging. In *Proceedings of Evaluation of NLP and Speech Tools for Italian*.

Andrea Gesmundo. 2009b. Bidirectional sequence classification for named entities recognition. In *Proceedings of Evaluation of NLP and Speech Tools for Italian*.

Andrea Gesmundo. 2011. Bidirectional sequence classification for tagging tasks with guided learning. In *Actes de Traitement Automatique des Langues Naturelles (TALN 2011)*.

Jan Hajič. 2000. Morphological tagging: data vs. dictionaries. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 94–101, Seattle, Washington. Association for Computational Linguistics.

Matjaž Juršič, Igor Mozetič, Tomaž Erjavec, and Nada Lavrač. 2010. LemmaGen: Multilingual lemmatisation with induced ripple-down rules. *Journal of Universal Computer Science*, 16(9):1190–1214.

Djordje Kostić. 2001. Quantitative description of Serbian language structure – Corpus of Serbian language.

Cvetana Krstev, Duško Vitas, and Tomaž Erjavec. 2004. MULTEXT-East resources for Serbian. In *Proceedings of 8th Informational Society - Language Technologies Conference, IS-LTC*, pages 108–114, Ljubljana, Slovenia.

Zoran Popović. 2010. Taggers applied on texts in Serbian. *INFOtheca*, 2(11):21–38.

Libin Shen, Giorgio Satta, and Aravind Joshi. 2007. Guided learning for bidirectional sequence classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 760–767, Prague, Czech Republic. Association for Computational Linguistics.

Duško Vitas, Cvetana Krstev, Ivan Obradović, Ljubomir Popović, and Gordana Pavlović-Lažetić. 2003. An overview of resources and basic tools for the processing of Serbian written texts. In *Workshop on Balkan Language Resources and Tools*, pages 97–104, Thessaloniki, Greece.

[2] The version of the system that we have trained and tested on the corpus of Serbian can also be used for Croatian and other varieties of former Serbo-Croatian.