Hybrid Lexical Tagging in Serbian

Matthieu Constant[∗∓], Cvetana Krstev[†], Duško Vitas [†]

[∗]Université Paris-Est, LIGM, Marne-la-Vallée, France, Mathieu.Constant@u-pem.fr
[∓] Alpage project, INRIA, Paris, France
[†]University of Belgrade, Belgrade, Serbia, {cvetana,vitas}@matf.bg.ac.rs

Abstract

This paper describes a joint approach to lexical tagging in Serbian, combining three fundamental natural language processing tasks: part-of-speech tagging, compound and named entity recognition. The proposed system relies on conditional random fields that are trained from a newly released annotated corpus and finite-state lexical resources used in an existing symbolic Serbian tagging system. Experimental results show that a joint strategy outperforms pipeline ones, in particular on out-of-domain texts.

1. Introduction

Lexical tagging is a key preprocessing stage for Natural Language Processing (NLP) applications as it maps a sequence of tokens into a sequence of tagged lexical units. Lexical units are semantic units that can be made of several tokens like multiword expressions (MWE). In this paper, we present methods to integrate three fundamental NLP tasks related to lexical tagging and applied to Serbian: part-of-speech (POS) tagging, Named Entity (NE) recognition and compound[1] recognition. Given a sequence of tokens in Serbian, our goal is to provide a tagged sequence of lexical units, a lexical unit being either a simple word, a compound or a multiword named entity. For instance,

- **Input:** *O svemu tome džentlmen se podrobno obavestio pregledajući svoj Bredšo, koji je sadržavao red vožnje prekomorske plovidbe za svaki dan.*
  (Mr. Fogg learned all this in consulting his Bradshaw, which gave him the daily movements of the trans-Atlantic steamers)
- **Output:** *O/PREP svemu/PRO tome/PRO džentlmen/N se/PAR podrobno/ADV obavestio/V pregledajući/V svoj/PRO Bredšo/NE ,/PONCT koji/PRO je/V sadržavao/V red_vožnje/N prekomorske/A plovidbe/N za/PREP svaki_dan/NE ./PONCT*

In particular, we experiment a supervised hybrid strategy, proposed by (Constant and Tellier, 2012), integrating information coming from finite-state linguistic resources (e-dictionaries and local grammars) into a statistical model trained on a reference annotated corpus. The contributions of the paper are the following:

- Release of new datasets for Serbian, with fine-grained annotations of NE and compounds, as well as POS tags;
- Experimental comparisons between different integration methods;
- Release of a new lexical tagger for Serbian

2. Background

2.1. Compound and NE recognition and POS tagging

Our paper focuses on the combination of three fundamental tasks of NLP: POS tagging, compound recognition (CR) and Named Entity Recognition (NER). These three tasks have been widely studied in the literature, but research on their combination is much less common. NER and POS tagging (as well as CR and POS tagging) are traditionally combined, as POS tagging is very frequently used to provide linguistic information to NER (CR) in the form of features in statistical approaches. CR and NER can also be combined in different ways as shown in (Vincze et al., 2011): either CR is informed by NER, or NER is informed by CR. Some have implemented limited joint strategies:[2] for instance, joint CR and POS tagging (Constant and Tellier, 2012). Furthermore, many studies have shown that statistical models could be efficiently trained by combining annotated corpora and pieces of information coming from linguistic/knowledge databases (ex. lexicons, gazeeters) like in (Denis and Sagot, 2009) for POS tagging or (McCallum and Li, 2003) for NER, or (Constant et al., 2012) for CR.

2.2. Lexical tagging in Serbian

The lexical tagger presented in this paper builds on an existing Serbian symbolic system implemented via the Unitex platform (Vitas and Krstev, 2012). This system relies on large-coverage and fine-grained e-dictionaries of simple and compound words, as well as on local grammars (cf. next section). In particular, the NE hierarchy in Serbian NER system consists of five top-level types: persons, organizations, locations, amounts, and temporal expressions, each of them having one or more levels of sub-types. For instance, locations have sub-types: hydronyms, oronyms, regions, cities, etc. The tagging strategy allows nesting, which means that a named entity can be nested within another named entity, e.g. an organization name can be nested within a person's role (or function) which is nested within a personal NE, like in *<pers><role>prvi premijer <org>Savezne*

[1] For the purpose of this paper, we define a compound as a contiguous sequence of tokens that has a non-compositional meaning. We exclude multiword named entities from it.

[2] Note that these tasks can also be jointly combined with parsing: e.g. CR (Nivre and Nilsson, 2004; Green et al., 2011) or NER (Finkel and Manning, 2009).

vlade</org></role><persName.full>Milan Panić</persName.full></pers> 'The first prime minister of the Federal Government Milan Panić'. However, for the purpose of our experiment we have kept only the NEs with the longest span.

The Serbian NER system is a handcrafted rule-based system that relies on comprehensive lexical resources for Serbian described in Subsection 3.1. For recognition of some types of named entities, e.g. personal names and locations, e-dictionaries and information within them is crucial; for others, like temporal expressions, local grammars in the form of Finite-State Transducers (FST) that try to capture a variety of syntactic forms in which a NE can occur had to be developed. However, for all of them local grammars were developed that use wider context to disambiguate ambiguous occurrences as much as possible. These local grammars were organized in cascades that further resolve ambiguities (Maurel et al., 2011). NER system was evaluated on a newspaper corpus and results reported in (Krstev et al., 2014) showed that F-measure of recognition was 0.96 for types and 0.92 fot tokens.

Note that the Serbian system we are relying on is not the only tagging system available for Serbian. For instance, (Agić et al., 2013) have experimented different statistical models and configurations for POS tagging and lemmatization. A TreeTagger was also used to PoS tagging of the Corpus of Contemporary Serbian (Utvić, 2011).

3. Data

3.1. Lexical Resources

The resources for NLP of Serbian consist of electronic dictionaries and local grammars. They are being developed using the finite-state methodology (Courtois and Silberztein, 1990), (Gross, 1989). The role of e-dictionaries, covering both simple words and compounds, and dictionary finite-state transducers (FSTs) is text tagging. Each e-dictionary of forms consists of a list of entries supplied with their lemmas, morphosyntactic, semantic and other information. The forms are, as a rule, automatically generated from the dictionaries of lemmas containing the information that enables production of forms. Compounds are assigned the same POS as simple words; however, compound verbs are not covered yet. The system of Serbian e-dictionaries covers both general lexica and proper names and all inflected forms are generated from 135,000 simple forms and 13,000 compound lemmas. Approximately 28.5% of these lemmas represent proper names: personal, geopolitical, organizational, etc.

Dictionary FSTs are used for recognition and tagging of some open classes of compounds, multiword numerals written with digits, words and their combinations (e.g. *2,52 milijarde* '2.52 billions'), and multiword nouns, adjectives and adverbs derived from numerals and written with digits (e.g. adjective *18-dnevni* '18 days long'), interjections with freely repeating parts (e.g. *hi-hi-hi-ho-ho-ho-ha-ha-ha*), etc. (Krstev and Vitas, 2006). The output format of these FSTs follows exactly the format of e-dictionaries; thus, from the recognized sequence an e-dictionary entry is formed and added to the used e-dictionaries. For instance, if the recognized sequence is the form *18-dnevnu*, a dictionary FST produces a dictionary entry *18-dnevnu,18-dnevni.A:aefs4q* which gives the form's lemma *18-dnevni*, its POS *A* (adjective), and a set of morphosyntactic categories *aefs4q*.

3.2. Annotated Corpus

For the experiment we used two texts: one for training and development and another one for testing. For training and development we used the Serbian translation of Verne's novel "Around the World in Eighty Days". The text was analyzed using Serbian lexical resources presented in previous sections in Unitex system.[3] The annotated text was prepared in two steps. First, the text was analyzed with e-dictionaries of simple words and then manually disambiguated (Tufiş et al., 2008). In the next step, the text was analyzed with remaining resources (e-dictionaries of compounds, dictionary FSTs and NER system), and results were manually disambiguated and corrected where necessary. Finally, both texts were automatically merged into one. The resulting text uses annotation codes applied in the Serbian system of e-dictionaries.

For testing, we prepared another text that comprises parts coming from three different sources: (i) the first chapter of the novel "The Good Soldier Švejk" (translation to Serbian) (referred to as *Švejk*); (ii) a few news articles dealing with floods in Serbia in 2014 (referred to as *Floods*); (iii) a few chapters of the History manual for elementary schools (referred to as *History*). First, the text was processed by e-dictionaries of simple words and compounds, dictionary FSTs, and at the end NER was applied. In the next step, all NE tags were manually checked and corrected. Finally, POS tags and lemmas of all simple words and compounds were manually disambiguated and necessary corrections were done (e.g. missing tags for words not covered by e-dictionaries were added).[4] The size of the training and testing texts are presented in Table 1.

	tokens	**simple**	**compounds**	**NE**
Verne	64,829	51,845	3,054	3,036
Švejk	2,953	3,104	108	192
Floods	4,272	3,232	237	395
History	5,193	4,859	471	531
Test	13,418	11,195	816	1,118

Table 1: The size of the training and testing texts. Tokens comprise words and punctuation marks. NEs can be both simple words and multiword units.

Processing of training and testing texts revealed that some entries were missing in dictionaries and they were added to them for future use (see Table 2). Entries added from the training text were used during the training phase, while entries added from the testing texts were not used

[3] The Unitex software system: `http://www-igm.univ-mlv.fr/~unitex/`.

[4] The disambiguation was done by a special tool integrated into Unitex system that facilitates manual disambiguation `http://tln.li.univ-tours.fr/Tln_Colloques/Tln_JUnitex2014/Communications/Vitas.pdf`.

in the testing phase, in order not to bias the experimental results.

	simple	**compounds**
Verne	294	143
Švejk	16	1
Floods	36	33
History	8	36
Test	60	70

Table 2: New entries added to e-dictionaries during text processing.

4. Approach

Given a sequence of tokens, our goal is to provide a tagged sequence of lexical units: a lexical unit being either a simple word, a compound or a multiword named entity; a tag being either a POS or a NE class. This involves the integration of three different tasks: POS tagging, NE recognition and compound recognition. Each of the three intended tasks are usually considered as sequential tagging tasks. Indeed, multiword NE tagging and MWE recognition can be seen as segmentation tasks (like chunking). By using a IOB-like scheme, it is equivalent to labeling simple tokens. Each token is labeled by a tag in the form B-X or I-X, where X is the label of the lexical unit the token belongs to. Prefix B indicates that the token is at the beginning of the lexical unit. Prefix I indicates an internal position. label O indicates an element that corrsponds to a simple word.

The three different tasks on the same sentence should produce independently the first three annotations (columns NER, CR and POS) in Table 3.

As depicted in Section 2., there are several possible orchestrations to reach our goal: either using a joint approach, or using a pipeline one. The joint approach consists in performing the three tasks in one step using a single sequential tagger (one model) by using a tagset, the labels of which combine the three annotations.[5] The corresponding output is provided in the last colomn of Table 3.

The pipeline approach consists in applying sequentially different tagging tasks. In particular, we tested two possibilities:

- POS → SEG: POS tagging is first performed on the token sequence, and the predicted POS are then provided as an input to a standalone MWE/NE recognition system
- SEG → POS : A standalone MWE/NE recognizer provides a sequence of lexical units as an input of a POS tagger.

For each module of these different orchestations, we used Linear chain Conditional Random Fields (CRF). They are discriminative probabilistic models introduced in (Lafferty et al., 2001) for sequential labelling and have been shown to be very accurate for segmentation tasks.

token	NER	CR	POS	JOINT
O	O	O	PREP	B-PREP
svemu	O	O	PRO	B-PRO
tome	O	O	PRO	B-PRO
džentlmen	0	O	N	B-N
se	O	O	PAR	B-PAR
podrobno	O	O	ADV	B-ADV
obavestio	O	O	V	B-V
pregledajući	O	O	V	B-V
svoj	O	O	PRO	B-PRO
Bredšo	B-NE	O	NE	B-NE
,	O	O	PONCT	B-PONCT
koji	O	O	PRO	B-PRO
je	O	O	V	B-V
sadržavao	O	O	V	B-V
red	O	B-N	?	B-N
vožnje	O	I-N	?	I-N
prekomorske	O	O	A	B-A
plovidbe	O	O	N	B-N
za	O	O	PREP	B-PREP
svaki	B-NE	O	?	B-NE
dan	I-NE	O	?	I-NE
.	O	O	PONCT	B-PONCT

Table 3: An example

5. Experiments

5.1. Setup

The various CRF models used in our experiments were trained on 80% of the Verne Corpus. The remaining 20% were used as development (dev) dataset (e.g. for feature tuning). As mentioned in Section 3.2., the test set is composed of the texts *Svejk*, *Floods* and *History*. We therefore performed out-of-domain evaluation in the sense that the dataset used for training/dev belong to a domain different from the one used for testing. The models were trained and tested with the software *lgtagger* (Constant and Sigogne, 2011) that allows easy incorporation of information coming from lexical resources into CRF in the form of features.

For our experiments, we set two parameters: (a) orchestration strategy; (b) use of lexicon-based features. Parameter (a) offers three possible values: one joint strategy and two pipeline ones. Parameter (b) is binary-valued (NO LEX or LEX). In the latter case, the lexicon-based features are computed as follows. We first applied the Serbian e-dictionaries and cascades of FSTs described in Section 3.1. on the whole corpus presented in 3.2., in order to create a single lexicon containing all recognized forms. *lgtagger* uses this lexicon to construct a preliminary "naive" segmentation to be used as a source of features (for more details, see (Constant et al., 2012)).

5.2. Results

Experimental results on development and test datasets are given in Table 4. Results are evaluated with the standard F-score (F) that is the harmonic mean of precision (P) and recall (R). Whereas all strategies reach comparable scores on the lexical segmentation task alone, it appears

[5] Note that it does not correspond to a strict combination of the three types of annotations, as we do not tag the internal elements of the multiword lexical units.

that the joint strategy is more robust on the tagging task on out-of-domain texts (test dataset). The experimental difference between the two tagging approaches is statistically significant with p-value < 0.01 computed from χ^2 score. This strategy has also the advantage of being easy to implement (a single model to train and to apply), although it is slightly slower to train than the ones used in pipeline strategies. For instance, the model used in the best joint strategy is trained in 1229s, instead of 484s for the longer training in a pipeline strategy on the same machine (Intel(R) Xeon(R) CPU E5640 @ 2.67GHz 8 core).

One can observe that the use of lexicon-based features greatly improves the accuracy of the lexical tagger, especially for out-of-domain texts: a gain of 6.5 pts in terms of F-score as compared with 2.5 pts for in-domain text (dev dataset). It also appears that lexical resources have a significant impact on precision first (+7 pts on the joint system) and then on recall (+4.5 pts).

	global	UC		UL		U	
	F	cov.	F	cov.	F	cov.	F
ALL	90.35	28.5	79.12	26.8	91.87	4.8	60.32
MW	63.72	90.6	61.39	37.4	7.45	37.1	6.49

Table 5: Scores on unknown units on the TEST set with the joint strategy and the use of lexicon-based features. We have investigated three sets of unknown units: lexical units absent from the training corpus (UC), units absent from lexical resources (i.e. not in dictionary and not recognized by NE transducers) (UL), units absent from training corpus or lexical resources (U). Raws ALL (resp. MW) correspond to all lexical units (resp. the multiword lexical units). For each set, column *cov.* displays its coverage on the tested text; column F displays the lexical tagging F-score. The column "global F" indicates the overall F-scores on the TEST set.

6. Discussions

This section is devoted to go deeper in the analysis of the results in order to have a better understanding what really happens with the joint system. We first make an error analysis on the development set, in order to obtain the main kinds of errors produced by the system. We then discuss how good unknown units are tagged.

6.1. Error analysis

There were 182 differences between the reference text and the output text. The differences can be described in the following way:

- **POS** is wrongly attributed. There were 107 differences of this kind. Adjectives, particles and adverbs had the most wrongly attributed POS (23, 22 and 19, respectively). Verbs and conjunctions were assigned wrongly in most of the cases (27 and 17, respectively). Prepositions were always correctly tagged; numerals were never wrongly assigned. The most confusions were between pairs: adjective/verb (19), particle/conjunction (17), noun/verb (13). Many cases of adjective/verb confusion come from the past participle of a verb and an adjective derived from it, e.g. *zatvoren* 'close/closed'.

- **NE recognition** A simple word NE was not recognized (instead a correct POS was assigned), or a simple word was wrongly recognized as a NE. There were 18 differences of this kind. Example: a time NE *uveče* 'in the evening' was assigned a POS *ADV*. The second example: the noun *Kina* 'China' (the name of a ship) was recognized as a toponym.

- **NE type** A NE type was wrongly attributed or it was not attributed at all. There were 25 differences of this kind. Example: a money NE *dve hiljade dolara* 'two thousand dollars' was recognized as an amount NE. The most differences included time, amount, money and measure NEs.

- **NE span** A NE span was not correctly established. There were 26 differences of this kind. Example: a time NE *osam časova i četrdeset i dva minuta* 'eight o'clock and forty two minutes' was recognized as two separate time NEs: *osam časova* and *četrdeset i dva minuta*.

- **compound recognition** compound not recognized, or a simple word sequence wrongly recognized as a compound. There were 3 differences of this kind. Example: a compound *dobro delo* 'a good deed' was recognized as a sequence *dobro ADV delo N*, where a POS *ADV* is wrongly assigned (it should be *A*).

- **foreign words** A foreign word assigned an incorrect POS. There were 3 of this kind. For instance, *of* in *Siti of Pariz* 'City of Paris' was assigned PONCT tag (for punctuation marks and special characters) instead X (unknown/foreign words).

6.2. Unknown units

We have also explored results for unknown units in order to picture how the system is able to behave on unseen units. These results are displayed in Table 5. One first striking observation is that the lexicon has a very high impact for the prediction of multiword lexical units: the recognition of multiword units absent from the lexicon is a disaster, reaching an accuracy lower than 10%. Furthermore, we can deduce from the results that the impact of the lexicon-based features for simple units is mitigated: the model tends to favour other features.

7. Conclusions and Future Work

This paper describes three methods to integrate POS tagging, NE recognition and compound recognition into a lexical tagging system for Serbian: two pipeline strategies involving a POS tagger and a NE/compound recognizer; a joint strategy performing the three tasks at the same time. All strategies were based on CRF models trained from a new annotated corpus and existing lexical resources. The experimental results showed that the joint strategy appears to be the more robust to tag out-of-domain texts. The lexical resources showed to greatly improve the accuracy of the system, especially for multiword unit tagging. This paper opens new perspectives. In particular, a neural network

		DEV (in-domain)						TEST (out-of-domain)					
		NO LEX			LEX			NO LEX			LEX		
		R	P	F	R	P	F	R	P	F	R	P	F
JOINT	SEG	98.37	97.69	98.03	99.33	99.09	99.21	95.58	91.46	93.48	97.61	95.65	96.62
	TAG	95.07	94.41	94.74	97.36	97.12	97.24	85.86	82.15	83.96	91.28	89.44	90.35
POS → SEG	SEG	98.75	97.53	98.13	99.49	99.09	99.29	96.15	91.26	93.64	97.55	94.96	96.24
	TAG	95.18	94.00	94.58	97.44	97.04	97.24	86.15	81.76	83.90	91.06	88.64	89.83
SEG → POS	SEG	98.35	97.21	97.77	99.45	98.94	99.19	95.86	90.91	93.32	97.72	95.59	96.64
	TAG	94.87	93.77	94.31	97.36	96.86	97.11	85.99	81.55	83.71	90.71	88.73	89.71

Table 4: Overall scores on DEV and TEST datasets. We provide two kinds of evaluation: (a) lexical segmentation alone (SEG); (b) segmentation + tagging (TAG).

could be experimented in order to get freed from feature-engineering.

8. Acknowledgments

This research was partly supported by the Serbian Ministry of Education and Science under grant #47003.

9. References

Agić, v., N. Ljubešić, and D. Merkler, 2013. Lemmatization and morphosyntactic tagging of croatian and serbian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*. Sofia, Bulgaria: Association for Computational Linguistics.

Constant, M. and A. Sigogne, 2011. MWU-aware part-of-speech tagging with a CRF model and lexical resources. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE'11)*.

Constant, M., A. Sigogne, and P. Watrin, 2012. Discriminative strategies to integrate multiword expression recognition and parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL'12)*.

Constant, M. and I. Tellier, 2012. Evaluating the impact of external lexical resources into a crf-based multiword segmenter and part-of-speech tagger. In *Proceedings of LREC 2012*. Istanbul, Turkey.

Courtois, B. and M. Silberztein, 1990. *Dictionnaires électroniques du français*. Paris: Larousse.

Denis, P. and B. Sagot, 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC'09)*. Hong Kong.

Finkel, J. R. and C. D. Manning, 2009. Joint parsing and named entity recognition. In *Proceedings of the conference of the North American Chapter of the Association for Computational Linguistics (NAACL'09)*.

Green, S., M.-C. de Marneffe, J. Bauer, and C. D. Manning, 2011. Multiword expression identification with tree substitution grammars: A parsing tour de force with french. In *Proceedings of the conference on Empirical Method for Natural Language Processing (EMNLP'11)*.

Gross, M., 1989. The use of finite automata in the lexical representation of natural language. In M. Gross and D. Perrin (eds.), *Electronic Dictionaries and Automata in Computational Linguistics*, volume 377 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, pages 34–50.

Krstev, C., I. Obradović, M. Utvić, and D. Vitas, 2014. A System for Named Entity Recognition Based on Local Grammars. *J Logic Computation*, 24(2):473–489.

Krstev, C. and D. Vitas, 2006. Finate State Transducers for Recognition and Generation of Compound Words. In T. Erjavec and J. Žganec Gros (eds.), *Proc. of the 5th Slovenian and 1st International Conference Language Technologies*. Institut "Jožef Stefan".

Lafferty, J., A. McCallum, and F. 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)*.

Maurel, D., N. Friburger, J.-Y. Antoine, I. Eshkol, D. Nouvel, et al., 2011. Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement Automatique des Langues*, 52(1):69–96.

McCallum, A. and W. Li, 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*.

Nivre, J. and J. Nilsson, 2004. Multiword units in syntactic parsing. In *Proceedings of Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA)*.

Tufiş, D., S. Koeva, T. Erjavec, M. Gavrilidou, and C. Krstev, 2008. Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages. In *Proc. of the 6th International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL 2008)*. Dubrovnik, Croatia.

Utvić, M., 2011. Annotating the Corpus of Contemporary Serbian. *INFOtheca*, 12(2):36a–47a.

Vincze, V., I. Nagy, and G. Berend, 2011. Multiword expressions and named entities in the wiki50 corpus. In *Proceedings of the conference on Recent Advances in Natural Language Processing (RANLP'11)*.

Vitas, D. and C. Krstev, 2012. Processing of Corpora of Serbian Using Electronic Dictionaries. *Prace Filologiczne*, LXIII:279–292.