Joint Compound/Named Entity Recognition and POS Tagging for Serbian: Preliminary Results (*WG3 poster session*)

Matthieu Constant
Univ. Paris-Est, LIGM, CNRS
France
mconstan@univ-mlv.fr

Cvetana Krstev
Univ. Belgrade
Serbia
cvetana@matf.bg.ac.rs

Dusko Vitas
Univ. Belgrade
Serbia
vitas@matf.bg.ac.rs

1 Introduction

We present preliminary results on joint compound/named entity (NE) recognition and POS tagging for Serbian. Our approach is inspired by our previous work (Constant and Sigogne, 2011) on French on a very similar joint task (compound recognition + POS tagging). We used a supervised statistical approach based on linear Conditional Random Fields (CRF) and a corpus that was semi-automatically annotated by exploiting large-scale linguistic resources (Tufiş et al., 2008).

2 Data

For the experiment we have used the Serbian translation of Vernes novel "Around the World in Eighty Days". The text was analyzed using Serbian lexical resources in Unitex system. These resources are: (a) Serbian e-dictionaries of simple words (b) Serbian e-dictionaries of compounds; (c) Serbian dictionary graphs for the recognition of some specific cases of compounds like multiword numerals and compounds with numerals; (d) Serbian system for named-entity recognition that consists of a large cascade of finite-state transducers. This system recognizes all basic named-entities: temporal expressions, amount expressions, persons, locations and organizations. All Serbian resources were enriched as a result of this analysis.

The annotated text was prepared in two steps. In the scope of the project SEE-ERA.NET - Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages (ICT 10503 RP), the text was analyzed with e-dictionaries of simple words (a) and then manually disambiguated (Tufiş et al., 2008). Later, the text was analyzed with remaining resources (b-d), results were manually disambiguated where necessary. Finally, both texts were automatically merged into one. The resulting text uses annotation codes applied in the Serbian system of e-dictionaries.

Below is an example of annotated sentence (found in the corpus):

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

(The gentleman fully realized all this when he consulted his Bradshaw, which contained details of cross-ocean navigation for each day.)

The character _ delimits tokens in a multiword unit. Named entities are tagged with the symbol NE. Other symbols stand for POS.

3 Approach

The integration of compound/named entity recognition in the process of POS tagging can be seen as integrating a segmentation task. It is somewhat similar to tasks like chunking or Named Entity Recognition, that identify the limits of chunk or Named Entity segments and classify these segments. By using a BIO-like scheme, it is then equivalent to labelling simple tokens. Each token is labeled by a tag in the form B-X or I-X, where X is the POS labelling the lexical unit (simple word, compound or named entity) the token belongs to. Suffix B indicates that the token is at the beginning of the lexical unit. Suffix I indicates an internal position. Suffix O is useless as the end of a lexical unit corresponds to the beginning of another one (suffix B) or the end of a sentence. Such scheme therefore determines lexical unit limits, as well as their POS. A simple approach is therefore to relabel the training data in the BIO-like scheme and to learn a CRF model from it. Using such scheme, the example shown above is relabeled as follows:

O/B-PREP svemu/B-PRO tome/B-PRO džentlmen/B-N se/B-PAR podrobno/B-ADV obavestio/B-V pregledajući/B-V svoj/B-PRO Bredšo/B-NE ,/B-PONCT koji/B-PRO je/B-

V sadržavao/B-V red/B-N vožnje/I-N prekomorske/B-A plovidbe/B-N za/B-PREP svaki/B-NE dan/I-NE ./B-PONCT

We trained our CRF models by incorporating features used in (Constant and Sigogne, 2011). For instance, we included very standard ones for POS tagging (word forms, suffixes, prefixes, and so on), as well as lexicon-based ones. We also incorporated features specific to multiword units: e.g. word bigram, features computed fro compound e-dictionaries.

4 Preliminary results

We split the data in two sections: a training set (around 56,000 lexical units including 2,365 multiword ones) and an evaluation set (around 8,700 lexical units including 465 multiword ones). The data contains 14 POS tags plus the named entity one. We provide the preliminary results in the following table. The evaluation was performed by computing the overall F-score and the F-score limited to multiword units.

	without lex	with lex
overall	94.6	96.8
multiword	83.1	88.9

Table 1: Preliminary results

We can observe that the results obtained without lexicons are comparable with the ones obtained for French with the same configuration (Constant and Sigogne, 2011). The results with lexicon are stronger but are biased as the used lexicons were updated during the corpus annotation validation. For future experiments, we shall run the same experiments on a new evaluation corpus, and therefore removing the bias. In this poster session, we will also show that such a joint approach reaches comparable results with a pipeline approach.

References

Dan Tufiş and Svetla Koeva and Toma Erjavec and Maria Gavrilidou and Cvetana Krstev. 2008. Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages. In Marko Tadi, Mila Dimitrova-Vulchanova and Svetla Koeva (eds.). *Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL 2008)*, pp. 145-152. Dubrovnik, Croatia.

Matthieu Constant and Anthony Sigogne. 2011. MWU-aware Part-of-Speech Tagging with a CRF model and lexical resources. *ACL Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE'11)*. Portland, Oregon, USA.