Joint Compound/Named Entity Recognition and POS Tagging for Serbian: Preliminary Results

Matthieu Constant*, Cvetana Krstev◇ and Dusko Vitas◇
*Univ. Paris-Est - LIGM - CNRS, France, ◇ Univ. Belgrade, Serbia

Summary (WG3 poster)

- **MWE-aware Part-of-Speech tagging** in Serbian with an **hybrid approach**
- Based on (Constant and Sigogne 2011) applied to French
- Statistical dimension: **Conditional Random Fields** (CRF) learned on an annotated corpus
- Symbolic dimension: features are computed from **external linguistic resources**
- Comparison of joint and pipeline strategies

Serbian Data (Tufiș et al. 2008)

Annotated text

- Serbian translation of Verne's novel "Around the World in Eighty Days"
- MWE limitations: compounds and multiword Named Entities
- 15 Part-of-Speech labels
- 68,476 tokens and 4,537 sentences
- 64,829 lexical units including 2,830 multiword ones
- Split: train (3,883 sentences/2,365 MWEs); test (654/465)

Example of annotated sentence

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.

Preparation

- Automatic application of several lexical resources
  - Compound and POS annotation: e-dictionaries (simple words and compounds); application of dictionary graphs for the recognition of some specific cases of compounds like multiword numerals and compounds with numerals
  - Named entities: system for named-entity recognition that consists of a large cascade of finite-state transducers; the system recognizes all basic named-entities: temporal expressions, amount expressions, persons, locations and organizations
- Manual disambiguation and correction

Statistical supervised approach

Sequential labelling with CRF

- MWE recognition = a chunking task → use of BIO-like annotation scheme
- Hydridity : external MWE lexicons can be used as a source of additional features
- Software: lgtagger (Constant and Sigogne 2001)

Joint strategy

- The annotation scheme combines MWE segmentation and POS tagging
- The exemple sentence is annotated 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

Preliminary experiments

	without lex	with lex
overall	94.6 (83.1)	96.8 (88.9)
unknown	76.0 (56.6)	87.2 (71.7)

Table: F-score for the joint approach with or without use of external lexicons. Scores in parenthesis correspond to MWE only.

System	Overall	MWE
baseline	90.9	32.6
joint	96.8	88.9
pipeline	96.9	89.7

Table: F-score comparison of the best systems (with use of lexicons). The baseline coresponds to an MWE segmentation based on lexicon lookup, followed by CRF-based POS tagging. The pipeline approach consists of a CRF-based MWE segmentation (BI annotation scheme) followed by POS tagging.

References

- Dan Tufiș 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 Tadić, 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