Semantic and Syntactic Patterns of Bulgarian Multiword Names

Svetla Koeva, Tsvetana Dimitrova
Department of Computational Linguistics, Institute for Bulgarian Language

1. The Task

Persons, locations and organisations are the three major classes of proper names with a rigid designator called named entities (NEs) (Grishman and Sundheim, 1995). As person, location and organisation NEs often consist of more than one word, they can be classified as types of multiword expressions (MWEs) (Sag et al., 2002).

In our work, we aim at extending the resources for named entity recognition and classification (NERC) for Bulgarian by developing a set of rules, expanding available gazetteers and building a corpus with manually annotated NEs. We focus on multiword names (MWNs) – persons, locations and organisations, and their extensions with triggers. Both proper nouns and triggers are classified in a set of predefined categories – semantic classes – with respect to their general semantics and the patterns they evoke.

2. Towards a Classification of Bulgarian Multiword Names

We follow Searle in assuming that proper nouns do not describe or specify characteristics of objects, but are logically connected to characteristics of the object to which they refer (Searle, 1958). Thus, a proper noun refers to a specific person or object and inherits its characteristics which may be morphologically expressed if the language structure allows it. For example, Bulgarian personal names are marked for noun gender (masculine and feminine) and noun case (nominative and vocative). We assume that the triggers semantically depend on the referent of the proper noun.

Persons, locations, organisations and their triggers are grouped into semantic classes sharing common semantic patterns (language independent, to a great extent). For example, the triggers denoting academic position are related to the semantic pattern: in-something; at-somewhere (*profesor po himiya v Sofiyskiya universitet* 'professor in chemistry at the Sofia University'). The triggers of person NEs are: (1) Legislative job titles: *ministăr-predsedatel* ‘prime minister'; (2) Executive job titles: *izpălnitelen direktor* ‘executive officer’; (3) Aristocratic title: *princ* ‘Prince'; (4) Judicial position: *sădiya* ‘judge’; (5) Academic position: *glaven asistent* ‘senior assistant’; (6) Academic title: *doktor na filolologičeskite nauki* ‘Doctor of Philology’; (7) Military rank: *general mayor* ‘Major General’; etc. Personal names that select triggers – academic titles, can be only first names modified with or substituted by family names, while the personal names that select aristocratic titles, can be only first names. This condition exemplifies the constraints we impose over NE structure through the rules.

3. Syntactic Patterns of Bulgarian Multiword Names

The syntactic patterns (rules) define the NE class, part-of-speech and grammatical subclass of the head noun, part-of-speech of NE constituents, dependencies between constituents, cliticization (possessive and interrogative clitics), contiguity, and word order permutations.

Rules are abstractions of contiguous elements and do not define a hierarchical (linguistic) structure. The rule writing formalism (Karagiozov et al., 2014) imposes specification on each constituent in a sequential order. Thus, the rules have to exhaust all lexical, grammatical and word order permutation combinations, specific for Bulgarian personal names. The empirical knowledge for lexical compatibility results in construction of different lexicons, each containing a set of words with common semantic, morphological and syntactic features, i.e., nobility particles, feminine family names, military titles etc. In some cases, the sets of words can be defined by means of their grammatical class and/or subclass, i.e., 'NH' – all proper nouns, 'NHFsof' – feminine proper nouns, singular and indefinite; etc.

The word order permutations are expressed as different paths in the rules. Since the proper nouns show relatively low variation of components, the formalism is appropriate for their description. The rules account for: single or compound proper nouns, proper noun phrases without triggers, and proper noun phrases extended with triggers. Although the rules have not been tested in real texts so far, they are constructed such as to recognise a set of artificially created examples.

4. The Corpus

The annotated corpus is excerpted from the Bulgarian National Corpus (texts of genres such as news, fiction, popular science, subtitles have been selected), and, at the moment, it consists of 101,507 tokens. Sentence boundaries, PoS tags and lemmas are automatically annotated. The MWNs are manually annotated with their major constituents – proper names, triggers, proper name heads, triggers heads, and semantic classes of proper names and triggers. We follow tag-for-meaning principles of annotation, e.g., if an organisation name involves a person name (as *Gianni Versace S.p.A.*), we apply nested annotation where *Gianni Versace* is annotated as a person name, and together with *S.p.A.* – as an organisation.

5. Related work

A set of general NER rules with reasonable accuracy was developed for rule-based annotation of NEs (Karagiozov et al., 2012). Several machine learning methods are also applied for NER in Bulgarian. Georgiev et al. (2009) offer feature-rich NER covering persons, organisations, and locations. Kim et al. (2012) automatically label NEs in Bulgarian and Korean texts with information obtained through English-Bulgarian/Korean parallel sentences from Wikipedia. Stoyanova (2014) works on automatic categorisation of MWEs with a focus of MNEs based on idiomaticity.

6. Conclusion

The resources we are developing: the semantic classification of MNEs, syntactic patterns, gazetteers and the annotated corpus, provide training data and feature sets for machine learning methods and rules and a test corpus for our rule-based approach.

References:

Georgiev et al. 2009: Georgiev, G., P. Nakov, K. Ganchev, P. Osenova, K. Simov. Feature-rich named entity recognition for Bulgarian using conditional random fields. – In: *International Conference RANLP 2009 – Borovets, Bulgaria,* pp. 113-117.

Grishman and Sundheim 1995: Grishman, R., B. Sundheim. Design of the MUC-6 evaluation. – In: *Proceedings of MUC-6*, Stroudsburg, PA: ACL, pp. 1–12.

Karagiozov et al. 2012: Karagiozov, D., A. Belogay, D. Cristea, S. Koeva, M. Ogrodniczuk, P. Raxis, E. Stoyanov, C. Vertan. i-Librarian – Free online library for European citizens. – In: *INFOtheca*, no. 1, vol. XIII, May, BS Print: Belgrade, pp. 27-43.

Karagiozov et al. 2014: Karagiozov, D., A. Belogay, A. Genov. Izvlichane na semantichna informaciya v sistemata za upravlenie na sadarzhanie ATLAS. – In: *Ezikovi resursi i tehnologii za balgarski ezik*. Sofia: Academic Publishing House, pp. 258-297.

Kim et al. 2012: Kim, S., K. Toutanova, H. Yu. Multilingual named entity recognition using parallel data and metadata from Wikipedia. – In: *Proceedings of the 50th Annual Meeting of the ACL*, pp. 694–702.

Sag et al. 2002: Sag, I., T. Baldwin, F. Bond, A. Copestake, D. Flickinger. Multiword expressions: A pain in the neck for NLP. – In: *Proceedings of CICLing-2002*, Mexico City, pp. 1–15.

Searl 1958: Searle J. R. Proper names. – In: *Mind*, 67 (266), pp. 166-173.

Stoyanova 2014: Stoyanova, I. Automatic categorisation of multiword expressions and named entities in Bulgarian. – In: *Proceedings of CLIB 2014*. Sofia: Institute for Bulgarian Language, pp. 40-48.