Semantic and Syntactic Patterns of Multiword Names (a Cross-language Study)

Cvetana Krstev, Svetla Koeva, Dusko Vitas, Tita Kyriacopoulou Claude Martineau, Tsvetana Dimitrova

1.The task

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

We focus on multiword names – persons, locations and organizations, and their extensions with triggers, which can be themselves nominal compounds. We aim at providing common template for classification of multiword NEs in different languages. The names and their triggers are classified in a set of predefined categories – semantic classes that capture the general semantics and to the great extend should be language independent.

The syntactic patterns of multiword names related to one semantic class show unique properties in different languages. The syntactic patterns (rules) in our model of representation provide information about the NE class, part-of-speech and grammatical subclass of the head noun, part-of-speech of constituents, dependencies between constituents, word order and contiguity, cliticization (possessive and interrogative clitics), and the class of embedded NEs, if any. The corresponding syntactic patterns for several languages: Bulgarian, English, French, Greek, and Serbian, are linked for the purposes of multilingual named entity recognition and classification (NERC).

2. General semantic classification of multiword names

There are several proposals to subdivide the three major classes of NEs (ACE, 2006; Fleischman and Hovy 2002; among others). Persons, locations, and organizations in our approach are classified into semantic classes according to their general meaning which (together with the language specific lexicalization and grammar) determines the corresponding syntactic structures. We justify the division in semantic classes according to their instances in different languages: first, middle and last name of people, locations, animals, fictional characters; feminine, masculine and neuter noun in singular or plural; non-compositional multiword NEs; foreign name; abbreviation. Triggers are common nouns and are divided into classes according to their general meaning and the type of their dependency to the head proper name. The semantic classification of triggers is related with the syntactic structure they evoke in different languages, and their position according to the head proper noun.

3. Syntactic patterns of multiword names in Bulgarian, English, French, Greek, and Serbian.

Multiword NEs are noun phrases but their structure and constituency show (combinatorial) constraints depending on the semantic class of the head proper noun. The permissible combinations are defined between: the instances of proper noun classes within a given major class; the instances of classes of a proper noun and its trigger; and the instances of classes of two or more triggers modifying one and the same proper noun, for example: Academic position + Academic title in Bulgarian and Serbian (*profesor doktor Ivan Ivanov*; **doktor profesor Ivan Ivanov*).

The triggers allow different arguments and modifiers – adjectives, adjectival phrases, nouns, noun phrases and prepositional phrases, as for example in Bulgarian *profesorăt po fizika Georgi Popov* ‘professor-the in physics Georgi Popov’, and in Serbian *professor fizike Džordže Popov* ‘professor physics$_{GEN}$ Džordže Popov’.

Further, the multiword names are describe according to their word order, contingency, and permissible clitization. For example, the Bulgarian person names may allow second post-determiner position of possessive clitics (*presidentăt ni Ivan Ivanov* ‘president-the our$_{POSS.CL}$ Ivan Ivanov’), and an interrogative clitic can focus each phrase constituent and the NE phrase as a whole. Bulgarian multiword location and organization names channel the person NE structure but interrogative clitics are always found at the end of the NE phrase, and possessive clitics are not always permissible.

4. Linking language specific syntactic patterns to a cross-lingual semantic class

The syntactic patterns are described in the finite state framework. Rules are abstractions of positive series where a transition can be a word, a lemma, and a grammatical tag. The rules for person, location and organization names formulated for different languages (Bulgarian, English, French, Greek, and Serbian) are linked to the semantic class they represent. Two finite state formalisms for rule representations are used - Unitex/GramLab (Paumier 2014) and Est (Karagiozov et al., 2014).

Several problems have to be handled: distinguishing and labeling the correct NE class in cases of

ambiguity, where one entity falls into several classes, and detecting the boundaries of NEs with complex structure and the embedded NEs.

5. Gazetteers

Both semantic and syntactic classifications can help in delineation of semantic classes, syntactic patterns, and further expansion of the gazetteers (applied as lexicons in the rules). Gazetteers can be amassed, for instance through internet search by filtering out proper names (personal, middle and family names, brand names), and triggers referring to particular semantic classes such as legislative positions (*minister, president*), positions titles, academic positions and titles. Also, names of organizations can be collected as a part of complex NEs, for instance in a Serbian NE *predsednik* ***Teniskog saveza Srbije*** *Petar Petrović* 'President of **the Serbian Tennis Federation** Petar Petrović'.

6. Annotated corpora

The manually annotated corpus have already been produced for some languages. For Bulgarian such corpus (about 200,000 tokens) contains texts from the Bulgarian National Corpus (of genres where NEs are of highest incidence such as news, fiction, popular science, subtitles). So far, 721 NEs are annotated. For Serbian, one text, Verne's novel *Around the World in 80 Days* was automatically annotated and manually checked. It contains 971 NEs. A multilingual annotated corpus (not necessarily aligned) will be produced in which each monolingual subcorpus would follow the same annotation principles (Constant et al. 2014). For instance, as NEs may consist of more than one word belonging to different semantic classes, different annotation can be used for each class. Tag-for-meaning principles can be followed for annotation – for example, if an organization name involves a person name (as in *Gianni Versace S.p.A.*), annotation can be where *Gianni Versace* is annotated as a person name, and together with *S.p.A.* - as an organization. A word, a MWE or a phrase can be annotated under one class only, excluding ambiguity. The head nouns (but not their dependents) can be assigned specific annotation. Coordination can be marked when two coordinated modifiers refer to the same head while the subordinate clauses and the coordination between separate NEs should be excluded from the scope of the annotation. The annotated corpus will provide significant number of occurrences for all NE classes and rule patterns and can serve as a test corpus for the multilingual NERC.

7. Related work

A set of general NER rules with reasonable accuracy was developed for rule-based annotation of NEs in Bulgarian (Karagyozov et al., 2012) and Serbian (Krstev et al., 2013). Several machine learning methods are also applied for the NER in Bulgarian texts. Georgiev et al. (2009) offer feature-rich NER focusing on morphological features and disambiguation of NEs. Kim et al. (2009) automatically label NEs in Bulgarian and Korean with information obtained through English-Bulgarian/Korean language parallel sentences from Wikipedia. Stoyanova (2014) shows automatic categorization of various types of MWEs with a focus of multiword NEs based on idiomaticity. The resources we are developing: the semantic classification of multiword NEs, the syntactic patterns of multiword names in Bulgarian, English, French, Greek, and Serbian, the gazetteers and the manually annotated multilingual corpus, provide training data and feature sets for machine learning methods and rules and a test corpus for our rule based approach.

References

ACE 2006: *Automatic Content Extraction. English Entity Guidelines* v6.6 15 2008.06.13 https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf

Constant et al. 2014: Constant, M., C. Krstev, D. Vitas. Joint Compound/Named Entity Recognition and POS Tagging for Serbian: Preliminary Results. At: *3rd General Parseme Meeting*. Abstract. http://typo.uni-konstanz.de/PARSEME/images/Meeting/2014-09-08-Frankfurt-meeting/WG3-CONSTANT-KRSTEV-VITAS-abstract.pdf

Fleischman and Hovy 2002: Fleischman, M., E. Hovy. Fine grained classification of named entities. In: *Proceedings of COLING 2002*, NJ:ACL, pp. 1–7.

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.* Academic Publishing House, pp. 258-297.
Kim et al. 2009: 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 Association for Computational Linguistics* (Vol. 1: Long Papers), pp. 694–702.
Krstev et al. 2013: Krstev, C., I. Obradović, M. Utvić, D. Vitas, A system for named entity recognition based on local grammars, In: *J Logic Computation* 24(2), pp. 473-489, 2014, Oxford Journals, doi:10.1093/logcom/exs079, first published online February 19, 2013,
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.
Stoyanova 2014: Stoyanova, I. Automatic categorisation of multiword expressions and named entities in Bulgarian. In: *Proceedings of CLIB 2014.* Institute for Bulgarian Language, pp. 40-48.
Paumier 2014: Paumier, S. 2014. Unitex 3.1beta User manual. http:// http://www-igm.univ-mlv.fr/~unitex/UnitexManual3.0.pdf.