Named entity recognition in the system for information extraction

One of the first tasks of information extraction is recognition of named entities. To solve this problem in both English and Western European languages a series of techniques that give good results were developed, while there is still plenty of room for development for the Slavic, typically morphologically rich languages. Up to now, the development of electronic morphological dictionary based on finite automata and development of algorithms based on statistical methods were the approaches to resolving this issue in the Serbian language. However, this paper will deal with the problem of identification of named entities[1] in Serbian texts using the GATE[2] system.

GATE (General Architecture for Text Engineering) is a open software developed by the University of Sheffield Natural Language Processing Group. The output from the GATE system is the text with tagged named entities. The process of tagging is done through several phases: 1) identification of tokens and segmentation, 2) association with the appropriate parts of speech (POS), 3) tagging the named entities identified by the list (Gazetteer). Work of each of the mentioned phases is based on the results of the previous phase.

The application of GATE on Serbian language texts does not provide equally good results in the individual phases, i.e. while the phase 1) gives satisfying results, the same can not be said for the phase 2), or for the phase 3) likewise. In fact, the problem of ambiguity which appears in the phase 2) remains present even after the recognition of named entities in the phase 3). One possible idea would be to use the existing statistical POS taggers adjusted for the Serbian language, but they have not given the results that are accurate enough yet.

In addition to ambiguity, there is also a problem of numerous morphological forms of named entities that would have to be in the lists. Such extended list could be generated on the basis of the electronic dictionary of named entities in the Serbian language, but that solution opens more problems: a) the size of the list would grow by adding all the types of named entities, b) the inability to contain the list of all named entities that appear in the Serbian texts.

There are several approaches to solving this problem. The first approach advocates that the lists should contain only the lemmas of named entities and that GATE would then, by incorporating some additional mechanisms, generate their derived forms and use them in tagging of named entities.

Another approach assumes that two different systems, GATE and Unitex[3], use finite automata to tag the text, and differ only in the representation of these automata. In both cases, the user creates a logical representation (graph in the Unitex system, and JAPE grammar in the GATE system), which the system automatically converts to the corresponding physical representation (a binary format). As there are a number of graphs for identifying named entities in Serbian developed for Unitex, instead of creating the equivalent JAPE grammar, this approach suggests the attempt of transformation of Unitex graphs to the physical representation of finite automata used by GATE.

The third approach, which we eventually chose, uses Unitex for all the phases of recognition and tagging of named entities, after which the resulting text is transformed by a special program to the output format of the GATE system. This ensures the possibility to load the resulting text into the GATE system, to clearly spot the detected named entities and categories, and then use other tools of the GATE system.

References

1. Satoshi Sekine and Elisabete Ranchhod (eds), ”Named Entities: Recognition, Classification and Use”, John Benjamins Publishing Company, Amsterdam, Philadelphia, 2009

2. http://gate.ac.uk/

3. http://igm.univ-mlv.fr/~unitex/