Named Entity Recognition for Slovak Language

Ondrej KAŠŠÁK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
ondrej.kassak@gmail.com

Nowadays we are literally overwhelmed with information. It is impossible for us to process all the information we find. Various approaches have been proposed for the information overload problem as personalized recommendations based on the content or searching methods using key entities from the texts. They assume the named entities appearing in the text for their working as an input. Based on them, recommendation algorithms can search and work more efficiently in comparison with other methods working only with the text titles or with the most frequent words in the text.

In our research we propose the method for recognizing and extraction of named entities in texts. The aim of our proposed method is to recognize entities in the text and then place them into the proper categories. We primarily focus on texts in Slovak language because a comprehensive tool for this language that would identify all entities classified according to the MUC-6 (6th Message understand conference) [1] is missing. We also describe possibilities of application for other flective languages.

The proposed method consists of two parts - the initial part of pre-processing of the text and the recognition of the named entities.

For text pre-processing our method uses a form of stemming. We remove the word suffix caused by inflection. Suffixes are identified by comparison each word with the set of Slovak word endings. The result is a form of words which is not the word formation root but, with only a few exceptions, we get the uniform forms of words with which can be used in further computation [2].

Process of the recognizing named entities is composed from identifying potential entities occurring in the processed text, determine its scope and consequently identify the category to which they belong. We use Slovak[1] and English version of Wikipedia to identify new entities, database for fast recognizes of entity that we found before and Slovak National Corpus[2] for filtering common words with first capital letter from

* Supervisor: Michal Kompan, Institute of Informatics and Software Engineering
[1] http://sk.wikipedia.org/
[2] http://korpus.juls.savba.sk/

Spring 2012 PeWe Workshop, April 1, 2012, pp. 65–66.

entities. We propose the named entity extraction (Figure 1), which consists of several steps.

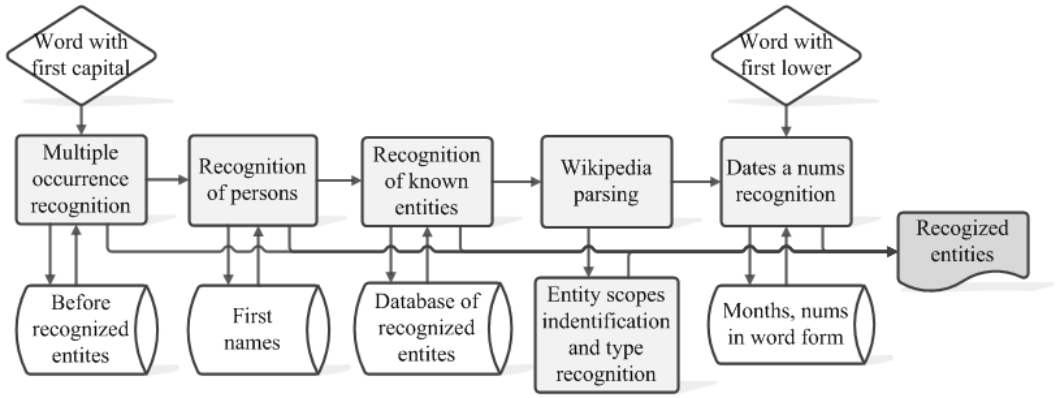

Figure 1. Sequence of steps describing proposed method.

Entities usually start with a capital letter in Slovak. We have just to search for the capital letters in the text. If we then sort out the beginnings of sentences or quotations that are not entities at the same time, we get a set of entity beginnings. Thus we are able to identify persons, organizations, locations and miscellaneous entities.

When found the beginning of the entity we compare it with list of before recognized entities, so we can simply identify entities that have already been recognized before without the long process of standard recognition of a new entity.

If don't find agreement, we compare it with the set of first names and possibly the database of recognized entities. If still don't identify the entity we recognize its scope through web parsing. If found the scope we try to recognize entity type.

In addition to mentioned entity types identifies our method also numeric entities and dates. The numeric ones distinguish between money amount, percentage and a number itself. In identifying the type of numerical entities the context words are the most significant.

To evaluate proposed approach we processed 60 articles from 3 various Slovak news servers. Texts were manually annotated by human expert. We obtained result of 79% F-measure (84% precision, 74% recall). Total we correctly recognized 1204 entities of 1620. We wrong identified 232.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11.

References

[1] Grishman, R., Sundheim, B.: Message understanding conference-6: A brief history. In: Proceedings of COLING, 96, (1996), pp. 466–471.

[2] Przepiórkowski, A.: Slavonic Information Extraction and Partial Parsing. In: Computational Linguistics, (June 2007), pp. 1-10.