The named entities marking scheme of biomedical texts for information extraction

Andrii Romaniuk[1], Tetiana Turchyn[2]

1. Department of computer-aided design, Lviv Polytechnic National University, Ukraine, Lviv, 12 S. Bandera street, E-mail: anrom7@gmail.com

2. Department of applied linguistics, Lviv Polytechnic National University, Ukraine, Lviv, 12 S. Bandera street, E-mail: tetiana.turchyn@gmail.com

Abstract – Named entity recognition (NER) is an important problem in text processing – particularly in information extraction (IE) – and is useful in many practical applications in the semantic Web context. The goal of NER is to identify all occurrences of specific types of named entities in the given document collection.

Key words: text mining (TM), named entity (NE), named entity recognition (NER), biomedical corpus, information extraction (IE)

I. Introduction

Within an overwhelming amount of biomedical knowledge recorded in texts, there is high research interest in techniques that can identify, extract, manage, integrate and exploit this knowledge. In the past few years, there has been an explosion of research papers on text mining for biomedical literature.

II. Named Entity Recognition: benefits

An ability to automatically perform named entity recognition (NER) has a lot of benefits, such as improving the expressiveness of queries and quality of research results

NE may be divided into several categories [1]:

- Generic NEs: consist of names of persons (PERSON), organizations (ORG), locations (LOCATION), dates, times, etc.
- Domain-specific NEs: consist of, for example, name of proteins, enzymes, organisms, cells etc., in the biological domain.

III. Information Extraction

Information Extraction (IE) Extract instances of predefined categories from unstructured data, building a structured and unambiguous representation of the entities and relations between them [1].

The main NER challenge is related with terminology, due to the complexity of the used terms for biomedical entities and processes [1, 2]:

- Non-standardized Naming Convention
- Ambiguous Names
- Abbreviations
- Descriptive Naming Convention

- Conjunction and Disjunction
- Nested Names
- Names of Newly Discovered Entities

IV. Implementation

The global work flow of a NER system is composed of the following [2]:

- **Corpus** Collection of (usually related) texts
- **Preprocessing** Perform tasks over natural language texts to simplify the recognition process
- **NER** Recognizes specific entity names
- **Postprocessing** Refinment of already recognized names, solving problems of the recognition process or extending it to recognize more entity names
- **Entity Names** Recognized Names from text

Conclusions

The primary goal of text mining is to retrieve knowledge that is hidden in the text and to present it in a concise and simple form to the final users.

As far as NER is developed mostly for English, further study will concern to Ukrainian medical and biomedical texts, since there is need in it.

References

1. U.Leser and J.Hakenberg. What makes a gene name? named entity recognition in the biomedical literature. Briefings in Bioinformatics, 6(4): 357–369, 2005
2. Zhou, G., Zhang, J., Su, J., Shen, D., & Tan, C. (2004). Recognizing names in biomedical texts: a machine learning approach. Bioinformatics, 20, 1178–1190.