

# Named Entity Recognition in Various Biomedical Domain

Hyun-Sook Lee<sup>1</sup>, Hyunchul Jang<sup>1</sup>, Jaesoo Lim<sup>1</sup>,  
Soo-Jun Park<sup>1</sup>, Seon-Hee Park<sup>1</sup>

**Keywords:** biological named entity recognition, domain independent system

## 1 Introduction

Until now, research in named entity (NE) recognition from bio-medical literature has focused on a limited domain. Most of biological NE recognition systems are designed to extract fixed semantic types of NEs and need domain-specific knowledge either by dictionaries and rules[1][2] or by some supervised learning techniques[3][4]. Porting to a different domain of these systems may require a lot of time and cost for non-trivial tasks, generating large corpus or rules by experts.

In this paper, we will discuss two major requirements of NE recognizing system to overcome these domain portability issues. And this paper presents an automated system that recognizes biological entities using UMLS metathesaurus[5] that is a comprehensive and huge resource to represent biological concepts.

## 2 Method

The first fundamental requirement for the domain independent system recognizing biological NE is selecting appropriate semantic types in new domain. Major semantic types for each domain can be different. For example, NEs assigned to “Mental or Behavioral Dysfunction”, one of UMLS semantic types, occur hardly in apoptosis domain but often in Alzheimer disease domain. It is important to select domain-specific semantic types because it determines what useful information from literature is. Additional requirements are resource construction and rule generation. The existing systems require large corpus or well defined dictionaries and rules by experts. To solve these problems in a automated way makes the system adapt to new domain easily.

Our system builds language resources by extracting meaningful information from concept names of UMLS (Unified Medical Language System) metathesaurus statistically. Then, it proposes major semantic types in a specific biological domain automatically using these resources. This system extracts various features from concept names and makes rules by combining feature-extracted tokens. Finally, NEs are recognized by using these rules from the given text. The proposed system consists of four major modules of Resource Builder, Semantic Type Selector, Rule Generator and Named Entity Recognizer.

In Resource Builder module, concept names of UMLS metathesaurus are divided into several subsets by using semantic categories. Then, Singleterm(NE just composed of one word) and Keyterm(word that plays an important role to constitute NEs in a certain category) are extracted. Semantic Type Selector is a module to select domain-specific semantic types automatically from UMLS semantic types. Because UMLS semantic types provide a broad scope, allowing for the

---

<sup>1</sup> Bioinformatics Research Team, Electronics and Telecommunication Research Institute, 161, Gajeong-dong, Daejeon, 305-350, Korea, E-mail: {lhs63473, janghc, jslim, psj, shp}@etri.re.kr

semantic categorization of a wide range of terminology in multiple domains, it can be used to determine semantic types for various domains. To select semantic types for a particular domain, we assume that semantic types with high frequencies from texts associated with the domain are important. Semantic type Selector extracts nouns from Medline abstracts associated with a certain domain and determine their semantic categories by matching Singleterm and Keyterm. Frequent categories of nouns are selected as major semantic types in the domain. Rule Generator module makes primary rules by combining features extracted from concept names, such as capital letter, numeric character, Greek letter and alphabet etc., including Singleterm and Keyterm. And it constructs rule database by filtering rules using weights. Named Entity Recognizer extracts NE candidates from input text. Then rule generation steps are followed for these candidates. Finally we match the candidates' rules with rule database data and determine their semantic categories.

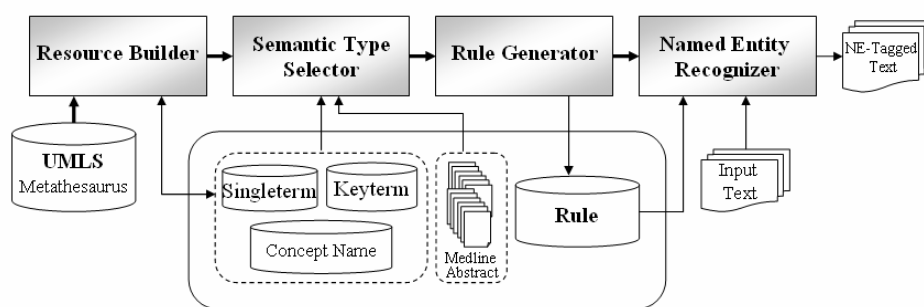


Figure 1: System architecture

### 3 Discussions

To recognize NE from bio-medical literature, rules or corpus is needed and a kind of semantic types assigned to NEs in a certain domain must be determined. This paper proposed NE recognizing system applicable to a new domain effectively using UMLS metathesaurus. It minimizes the cost of building resource and rules and doesn't require experts' help for domain-specific semantic types.

### References

- [1] Fukuda, K., Tamura, A., Tsunoda, T. and Takagi, T., Toward IE: Identifying protein names from biological papers, *Proc. Pacific Symp. Biocomputing '99*, 1999.
- [2] Denys Proux, Francois Rechenmann and Laurent Julliard, Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction, *Genome Informatics*, 9:72-80, 1998.
- [3] Jun'ichi Kazama, Takaki Makino, Yoshihiro Ohta and Jun'ichi Tsujii, Tuning Support Vector Machines for Biomedical Named Entity Recognition, *Proc. ACL-02 Workshop on Natural Language Processing in the Biomedical Domain*, pp. 1-8, 2002.
- [4] Irena Spasic, Coran Nenadic and Sophia Ananiadou, Using Domain-Specific Verbs for Term Classification, *Proc. ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pp. 17-24, 2003.
- [5] <http://www.nlm.nih.gov/research/umls/>