Language Resource and Rule Construction for Biological Named Entity System Using UMLS

Hyun-Sook Lee lhs63473@etri.re.kr
Tae-Hyun Kim heemang@etri.re.kr
Soo-Jun Park psj@etri.re.kr
Seon-Hee Park shp@etri.re.kr

Bioinformatics Research Team, Electronics and Communication Research Institute,
161 Gajeong-dong, Daejeon 305-350, Korea

Keywords: biological named entity recognition, UMLS, metathesaurus, language resource construction

1 Introduction

Biological named entity recognition means to extract names of biological entities which are used as important information such as protein name, gene name from bio-medical literatures. This paper proposes the method of language resource construction as well as rule generation as a basic step for bio-text mining. In the early researches, well defined dictionaries and rules by experts are used for named entity recognition [2, 3]. There have been researches that use various biological information resources such as SWISS-Prot, UMLS as a dictionary [5]. Extracting rules from corpus is one of the currently used methods. In these methods to make rules for named entity recognition, some useful additional information like contextual information or verbs adjacent to named entity is extracted from various corpora [1, 4].

In order to solve domain portability issue that is the main limitation of named entity recognition, this paper suggests methods that minimize the cost of resource construction and rule generation by extracting useful information from UMLS automatically without the help of experts and curated large corpus.

2 Method

In this paper basic language resources are constructed from biological information that is obtained automatically by using statistical methods from Metathesaurus of UMLS. Then, rules are generated with these resources. UMLS (the Unified Medical Language System) is a project to retrieve information of various bio-medical information resources and to integrate them effectively. It provides huge Metathesaurus with over two million bio-medical vocabularies. Because Metathesaurus provides information of concept names and their semantic types for classification, it can be used to recognize biological named entities. In this paper, in order to build basic language resources, concept name and semantic type are mapped to named entity and semantic category respectively. The proposed system consists of three major modules of Resource Builder, Feature Extractor, and Rule Generator.

In Resource Builder module, concept names are divided into several subsets by using semantic categories. Then, in order to obtain information that characterizes each semantic category following steps are taken: A concept name is divided into tokens, then by calculating weight value for each token, Single Term and Keyterm are extracted. Single Term means a word that becomes a named entity. Keyterm is a word that occurs in a certain category and plays an important role to constitute a named entity.

Feature Extractor module extracts various features from concept names in order to generate rules for named entity recognition. In this paper, Single Term and Keyterm are used to obtain semantic

characteristics of each token, and capital letter, numeric character, alphabet and Greek letter features are used to get surface characteristics. Also, in order to obtain structural characteristics of named entity composed of many tokens, we use preposition, conjunction and special character features. Feature extraction is done by tokenizing named entity into a word unit and extracts various features for each token.

Rule Generator module constructs rules by combining feature-extracted tokens in previous step. A rule can include many feature-extracted tokens, and each token can have many subtypes. Thus, to express the rule with the form of having one subtype for each feature in a token, we generate rules considering all of the possible combinations. After rules are generated, filtering is performed by calculating weights.

Language resources and rules constructed through proposed steps transferred to named entity recognition module and play an important role that affects to the performance of named entity recognition.

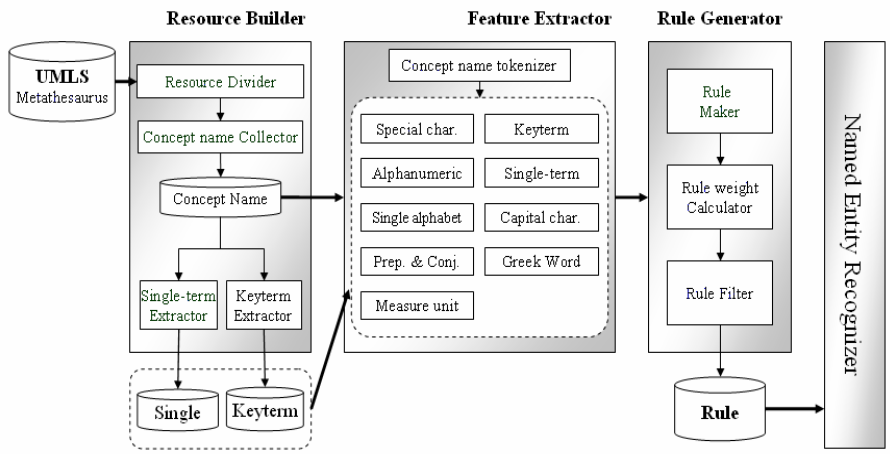

Figure 1: System Architecture.

3 Discussion

Resource construction and rule generation methods as basic steps for biological named entity recognition are proposed in this paper. The proposed methods generate resources and rules automatically from the Metathesaurus of UMLS statistically. It can reduce the cost of building resource and rules. It can also be applicable to a new domain effectively. After all, the proposed methods can contribute to solve the domain dependency problem that occurs frequently in bio-text mining research.

References

[1] Campbell, D.A. and Johnson, S.B., A technique for semantic classification of unknown words using UMLS resources, *Proc. AMIA Symp'99.*, 716–720, 1999.

[2] Proux, D., Rechenmann, F., and Julliard, L., Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction, *Genome Informatics*, 9:72–80, 1998.

[3] Fukuda, K., Tamura, A., Tsunoda, T., and Takagi, T., Toward information extraction: identifying protein names from biological papers, *Pac. Symp. Biocomput.*, 4:707-718, 1999.

[4] Spasic, I., Coran nenadic and sophia ananiadou, using domain-specific verbs for term classification, *Proc. ACL 2003 Workshop on Natural Language Processing in Biomedicine*, 17–24, 2003.

[5] Gaizauskas, R., Demetriou, G., and Humphreys, K., Term recognition and classification in biological science journal articles, *Proc. Computational Terminology for Medical and Biological Applications Workshop of NLP-2000,* 37–44, 2000.