Development of Construction and Management Tools for Biological Named Entity Dictionary

Hyunchul Jang janghc@etri.re.kr
Taehyun Kim heemang@etri.re.kr
Hyunsook Lee lhs63473@etri.re.kr
Soojun Park psj@etri.re.kr
Seonhee Park shp@etri.re.kr

Bioinformatics Research Team, Electronics and Telecommunications Research Institute, 161 Gajeong-dong, Yuseong-gu, Daejeon 305-350, Korea

Keywords: biological named entity, dictionary, name identification, UMLS, GO

1 Introduction

A system must identify biological entities to extract information for text mining on biological literature. The characteristics of biological entity nomenclature are very various and easily influenced by authors' personality, it is not easy to recognize the names by rules or training methods only. And it's impossible actually to list all possible names of entities and their all variations that appear in biological literature. To resolve this problem, it's effective to lookup dictionaries for already known named entities basically and to recognize unknown terms by rules and training methods. The quality of the used dictionaries is essential for the success of the matching procedure.

But it's not only very difficult but also expensive to construct high-quality dictionaries to one's satisfaction. Even though dictionaries are constructed, it's never easy to maintain them and it costs a great deal in the same manner. Therefore when we apply a process that we extract required information from well constructed resources and construct proper dictionaries automatically, we may obtain a good result with cutting down on dictionaries construction and maintenance expenses.

2 Method and Results

We propose an automatic construction method for biological named entity dictionaries and management tools to construct and manage them for search, maintenance and repair.

We use UMLS(unified Medical Language System) resources and GO(Gene Ontology) additionally. We concentrate on UMLS Metathesaurus that integrates biological concept vocabularies. We analyze concept names and semantic network system of UMLS Metathesaurus and classify biological vocabulary automatically. And then we build a dictionary from classified vocabulary. We add GO terms for recognizing relationships between extracted entities. GO terms are commonly used to label the processes and functions of organisms.

Because the stored formats, structures and classification systems of UMLS and GO are different, indices and building methods are considered for them basically and the tools are developed to integrate and manage these two resources. These tools are prepared to be used in java packages and GUI imports them to provide users with convenience. Web applications can do dictionary services using these packages, like dictionary searching and editing.

Dictionary entities are stored in relational database tables basically and some columns are indexed. Frequently searched fields may be indexed in external structure, like B-Tree for fast answering in name

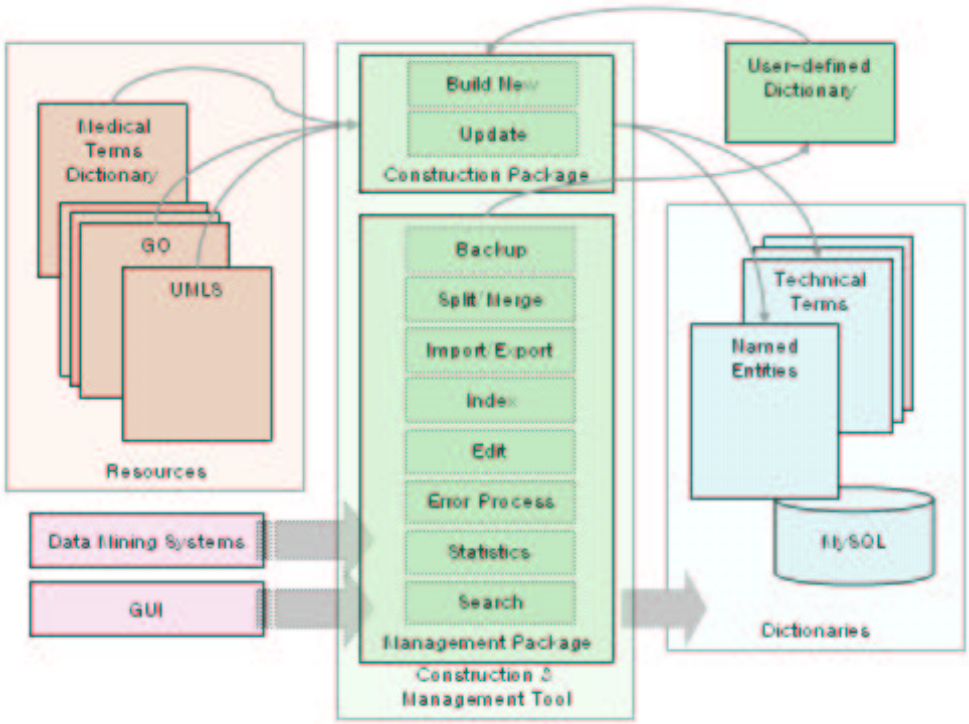

Figure 1: Scheme of construction and management tool.

extraction step. For useful categorization or filtering, some tables of UMLS Metathesaurus are joined into a view or new table in database. We import/export dictionary entities in XML files for effective exchange and share of information of dictionaries between biological data mining systems.

3 Discussion

A biological text mining system needs relationship terms dictionary and biological related keywords dictionary to extract relations between entities. Common terms used in biological meaning and natural language must be distinguished between them.

References

[1] Dehoney, D., Harte, R., Lu, Y., and Chin, D., Using natural language processing and the gene ontology to populate a structured pathway database, *IEEE CSB'03* Poster paper, 646–647, 2003.

[2] Fukuda, K., Tsunoda, T., Tamura, A., and Takagi, T., Toward IE: identifying protein names from biological papers, *Pac. Symp. Biocomput.*, 3:707–718, 1998.

[3] Hanisch, D., Fluck, J., Mevissen, H., and Zimmer, R., Playing biology's name game: identifying protein names in scientific text, *Pac. Symp. Biocomput.*, 8:403–414. 2003.

[4] 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.

[5] `http://www.geneontology.org/` Gene Ontology Consortium Web Site.

[6] `http://www.nlm.nih.gov/research/umls/` UMLS Project Web Site