Smart Searching System for Biomedical Information

Hong-Woo Chun [1]*, Chang-Hoo Jeong [1], Sa-Kwang Song [1],
Yun-Soo Choi [1], Sung-Pil Choi [1] and Hanmin Jung [1]

[1]Korea Institute of Science and Technology Information,
245 Daehangno, Yuseong-gu, Daejeon, 305-806, South Korea

ABSTRACT

Interactions between Biomedical entities provides meaningful information to detect and invent new drugs for diseases. Natural Language Processing-based Biomedical interaction extraction approach have shown encouraging results in the previous studies. While interaction extraction research has been a popular topic, research about searching and browsing methods for the extracted information has not been an attractive topic relatively.

This demonstration presents a smart searching system that provides various analysis tools for Biomedical interactions in whole PubMed. We expect that researchers can discover and develop new research outcomes through the proposed searching system.

1 INTRODUCTION

Many automatic Information Extraction (IE) approaches using Natural Language Processing (NLP) and Text Mining technologies have been proposed to extract automatically meaningful information in Biomedical domain. Biomedical entity recognition (Song *et al.*, 2011) and relation extraction research with respect to disease-gene association (Chun *et al.*, 2004) and protein-protein interaction (Chun *et al.*, 2011) are examples of the IE research in Biomedical domain. The IE system recognizes and extracts knowledge from a massive literature and the extracted knowledge is accumulated in a knowledge base.

In order to decide research topics, not only IE techniques but also effective searching methods for the extracted information are very important. While information extraction research has been one of the favorite topics, research about searching methods for the extracted information has not been an attractive topic relatively. In other words, it is overlooked even though various specialized searching and browsing methods are necessary to express the extracted information appropriately.

The demonstration will show a smart searching system for Biomedical entities and their interactions. Four types of searching services are included: Smart slide, Semantic network browsing, Top5, and Find it.

2 BIOMEDICAL INFORMATION IN PUBMED

Biomedical information in PubMed contains semantic triples extracted from 21 million PubMed abstracts. A semantic triple consists of two Biomedical entities and one verb, and two Biomedical entities are syntacticly a subject and an object for a verb appeared in a sentence. To extract the semantic triples, various NLP techniques are applied as the following two steps:

- To recognize entities, a machine learning-based named entity recognizer (Yoshida *et al.*, 2004) is used. The target concepts in the named entity recognition are the following six: genes/proteins, diseases, enzymes, drugs, symptoms and chemical compounds.
- To extract semantic triples, ENJU full syntactic parser (Miyao *et al.*, 2009), and GENIA event ontology (Kim *et al.*, 2006) are used. GENIA event ontology can cover biomedical relations.

Table 1 describes the statistical information of the extracted semantic triples from the whole PubMed abstracts.

Table 1. Statistical information of VSB

Objects	Frequency
Semantic triples	18,032,232
Entities	24,336,926
Event verbs	17,925,550

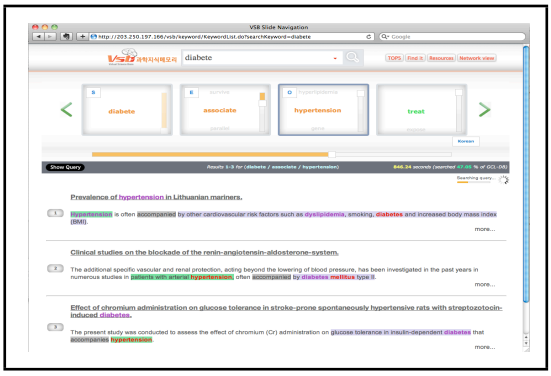

Fig. 1. Search result view with auto query generator

3 SMART SEARCHING SYSTEM

3.1 Smart slide

The proposed searching method has a familiar user interface. Searching process is started with a query, and the auto completion function recommends candidate queries.

Search results contain ranked documents as similar as those of the common searching systems (Figure 1). However, three differentiated functions are included in the proposed searching system as follows:

- First, six biomedical entities are highlighted in the results, and the information about entities are shown if mouse pointer is

*Corresponding author : hw.chun@kisti.re.kr

Fig. 2. Semantic network browsing views

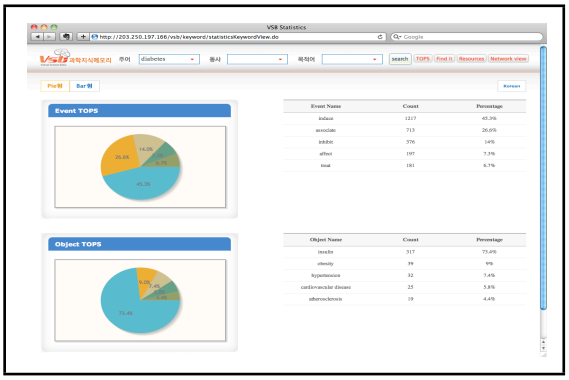

Fig. 3. Top5 views

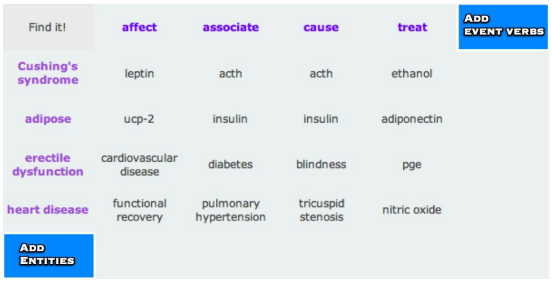

Fig. 4. FIND IT view

positioned over an entity. The information about the entities contains the corresponding concepts.

- Second, it is easy to use other services by selecting titles and highlighted entities. Titles are links to websites of the original PubMed articles. Once an entity is selected, a popup window shows up another services.
- Third, a query is easily constructed by selecting an entity or a verb from candidates. The candidate entities and verbs are all possible entities or event verbs related to the previously selected entity or verb. For the first input query, candidate verbs are listed in the next pane. Once a verb is selected, the next candidate entities are listed in the next pane.

This service might be helpful for more specific search with more specific query, and the search results are displayed immediately when a query is changed like the Google instant searching service.

3.2 Semantic network browsing

The proposed searching method describes relations among entities. A vertex and an edge indicate an entity and a verb (relation), respectively. All entities in the network contain the identifiers of external public databases such as UMLS, UniProt, BioThesaurus, KEGG and DrugBank. Thus, network involves not only the extracted information from texts but also information of other external databases. If a vertex is selected, synonyms are listed based on the frequency, and if an edge is selected, all relations between two entities are shown with the evidence sentences. Moreover, links to websites for the original documents are also provided (Figure 2).

3.3 Top 5

The proposed searching method shows popular relations for a query. As for a query, the results show list of entities or relational verbs based on frequency of co-occurrences (Figure 3). The co-occurrence indicates a sentence that contains both two entities, or both an entity and a verb. A Pie type and a bar type are the way to show the results.

3.4 Find It

The proposed searching method can provide latent attributes for diseases. *Erectile dysfunction*, an example in Figure 4, affects *cardiovascular disease*, associates with *diabetes*, can cause *blindness*, and is popularly treated by *PGE*. Evidence sentences can be shown by selecting corresponding entity.

4 CONCLUSION

Researches about information extraction from literature and construction of a knowledge basehave been actively conducted. However, researches about various searching and browsing methods for the extracted information have been relatively neglectful.

In the proposed approach, a smart searching system is introduced, and it contains four useful searching methods to utilize a multi-faceted scientific knowledge effectively. We expect that the proposed searching system provides various opportunities for researchers to detect and invent new products such as drugs conveniently.

REFERENCES

Song S. K., Choi Y. S., Chun H. W., Jeong C. H., Choi S. P., Sung W. K. (2011). *Multi-word Terminology Recognition Using Web Search*. UNESST 2011, CCIS 264, 233–238.

Chun H. W., Tsuruoka Y., Kim J. D., Shiba R., Nagata N., Hishiki T., Tsujii J. (2006) *Automatic recognition of topic-classified relations between prostate cancer and genes using MEDLINE abstracts*. BMC Bioinformatics, **7** (Suppl 3):S4.

Chun H. W., Jeong C. H., Song S. K., Choi Y. S., Choi S. P., Sung W. K. (2011). *Composite Kernel-based Relation Extraction using Predicate-Argument Structure*. UNESST 2011, CCIS 264, 269–273.

Yoshida K., Tsujii J. (2004). *Reranking for Biomedical Named-Entity Recognition*. BioNLP.

Miyao Y., Sagae K., Stre R., Matsuzaki T., Tsujii J. (2009). *Evaluating Contributions of Natural Language Parsers to Protein-Protein Interaction Extraction*. Biomedical Informatics, **25**(3) 394–400.

Kim J. D., Ohta T., Teteisi Y., Tsujii J. (2004). *GENIA Ontology*. Technical Report(TR-NLP-UT-2006-2). Tsujii Laboratory, University of Tokyo.