Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

BioProber: Software System for Biomedical Relation Discovery from PubMed

Hyunchul Jang, Jaesoo Lim, Joon-Ho Lim, Soo-Jun Park* and Kyu-Chul Lee

Abstract—The numbers of articles and journals that are published are increasing at a considerable rate, and the published information is growing continuously and fast. Because of this, researches to acquire knowledge automatically have been carried out in the areas of information retrieval, information extraction and text mining. Information retrieval approaches are good for specific topics that the number of related articles is small. But, if the number is bigger, searching skill and knowledge acquisition ability are useless. Though many efforts have been made to extract information from literature, many approaches have concentrated on specific entities, such as proteins, genes and their interactions, and much information is still remained in unstructured text. So, we have developed a system that discovers relations between various categories of biomedical entities. Our system collects abstracts from PubMed by queries representing a topic and visualizes relationship from the collection by automatic information extraction.

I. Introduction

MOST biologists visit sites such as PubMed to obtain documents which key facts are written in. There are many web-based applications for information retrieval (IR), but IR approaches return too many documents commonly except for the very specific topics.

Entity recognition (ER) is to find the biological entities that are mentioned within a text. ER is useful on its own for cross-linking the literature that is related to certain proteins or genes. ER helps to exclude unrelated documents and to find sentences that biomedical entities are mentioned.

In contrast to IR systems, information extraction (IE) systems aim to extract pre-defined types of fact, particularly, relationships between biological entities. The simplest IE approach is to identify entities that co-occur within abstracts or sentences. If two entities are repeatedly mentioned together, it is likely that they are somehow related, although the type of relationship is not known [1],[2]. Co-occurrence based methods tend to give better recall but worse precision than natural language processing (NLP) methods [3]. The main drawback of NLP approach is that a large number of extraction rules are needed to cover the many slightly different ways of expressing a certain relationship. These rules can either be developed manually or learned automatically from a corpus. Both methods are labor intensive, as the latter requires the prior manual tagging of a large training corpus [3].

There are many manually verified databases. They are very reliable, easily searchable and well structured. But they contain specific entities, such as proteins, genes and their interactions. And though facts extracted automatically from literature by IE can be stored in a database, with the option of being verified by a curator, much information remains in unstructured text. Most approaches have focused on extracting few types of relationship including physical protein-protein interactions and unspecified molecular mechanisms between proteins [3],[4].

Recently, NLP methods have been developed for extracting information on gene regulation, protein phosphorylation and tissue specificity. Because of the inherent complexity of this task, only a few systems have been designed that are able to extract multiple types of relationship [5]-[8]. So, we have developed a system that extracts relations between various categories of biomedical entities without limits in types of relation.

II. Methodology

Our system consists of the following three parts: PubMed collector, relation extractor and relation analyzer. The PubMed collector asks abstracts with a query given by a user

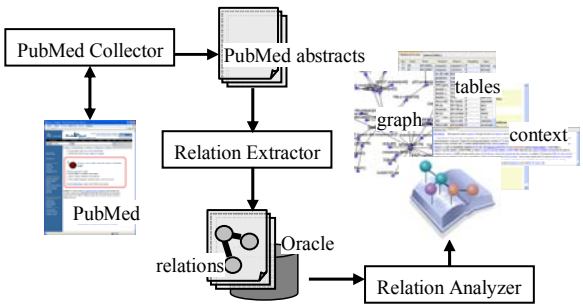

Fig. 1. Overview of BioProber

Manuscript received April 3, 2006. This work was supported in part by the Korean Institute for Information Technology Advancement (IITA) under Korean Ministry of Information and Communication.

Hyunchul Jang, Jaesoo Lim, Joon-Ho Lim, and Soo-Jun Park* are with the Bioinformatics Team, Electronics and Telecommunication Research Institute, Gajeong-Dong, Yuseon-Gu, Daejeon, 305-700, Korea (*corresponding author to provide phone: +82-42-860-6899; fax: +82-42-860-1208; e-mail: psj@etri.re.kr).

Kyu-Chul Lee is with the Department of Computer Engineering, Chungnam National University, Gung-Dong, Yuseon-Gu, Daejeon, 305-764, Korea (e-mail: kclee@cnu.ac.kr).

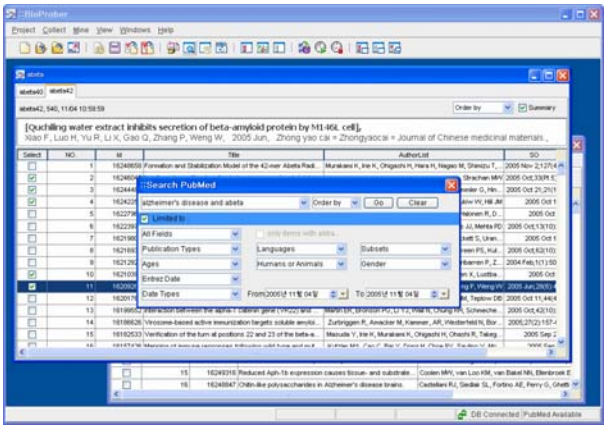

Fig. 2. The screenshot of the PubMed collector showing the PubMed search interface and searched results.

and fetches them. The relation extractor divides abstracts into sentences and recognizes biomedical named entities in sentences. Then, it extracts relational events among recognized entities. The relation analyzer accumulates extracted relations and visualizes in graphs and tables. This series is managed as a project.

A. PubMed Collector

Our system supports the same search interface and method as the original PubMed search. The PubMed collector searches, summarizes and fetches abstracts by the Entrez Programming Utilities. For one view which means a project, a number of searches can be executed.

Whenever one search is executed, a tab is added. A tab represents one query and a project is composed of tabs. The order of tabs is changeable. Names of tabs are given automatically by query string and users may rename them.

Query information, such as query string, query restriction, query time, the number of result abstracts and sorted order, are displayed on tab and result abstracts are listed in table as shown in Fig. 2. The PubMed collector gathers summaries of each abstracts and lists summaries in each columns of table. Users can select parts of document summaries and designate the order of columns.

The PubMed collector may import lists from external files, like Microsoft excels or plain text files. Reversely, all lists can be stored in a file. Into MS excel, each list is stored in each sheet. Users can store selected entries only and define contents in order and columns.

All lists or a selected list are updated by refresh command and new entries are marked in tables. Users can select parts of collected abstracts for the relation extraction step.

B. Relation Extractor

Selected abstracts are processed in background procedure. Extracted entities and relations are shown by the relation analyzer.

We use natural language processing techniques including part-of-speech (POS) tagging and syntactic parsing. We use a

```
Entity
  Virus
  Bacterium
  Anatomical_Abnormality
  Body_Part_Organ_Organ_Component
  Tissue
  Cell
  Cell_Component
  Gene_Genome
  Chemical
    Chemical_Viewed_Functionally
      Pharmacologic_Substance
        Antibiotic
      Biologically_Active_Substance
        Neuroreactive_Substance_Biogenic_Amine
        Hormone
        Enzyme
        Vitamin
        Immunologic_Factor
        Receptor
    Organic_Chemical
      Nucleic_Acid_Nucleoside_Nucleotide
      Organophosphorus_Compound
      Amino_Acid_Peptide_Protein
      Carbohydrate
      Lipid
    Inorganic_Chemical
    Element_Ion_Isotope
  Body_Substance
  Body_System
Event
  Organism_Function
    Mental_Process
  Organ_Tissue_Function
  Cell_Function
  Molecular_Function
    Genetic_Function
  Pathologic_Function
    Disease_Syndrome_Neoplastic_Process
      Mental_Behavioral_Dysfunction
    Cell_Molecular_Dysfunction
  Injury_Poisoning
```

Fig. 3. Hierarchical semantic categories of named entities.

statistical named entity recognition method using the Maximum Entropy model. Our named entity recognizer extracts 40 categories of named entities as shown in Fig. 3. Relations are extracted by syntactic analysis [9],[13] not by co-occurrence information. Relation types and entity categories are not limited to proteins.

We parse sentences syntactically in forms of the Penn Treebank syntactic tags [10] and extract relations by analyzing parsing results. Our rules are simple and small

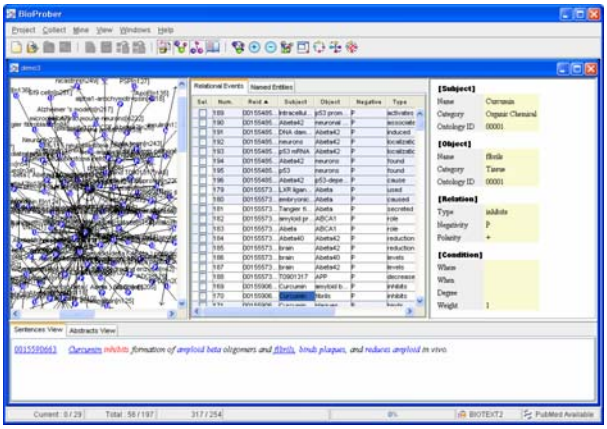

Fig. 4. The screenshot of the relation analyzer showing extracted relations in graph visualization, relation or entity list, detail specification and literature source.

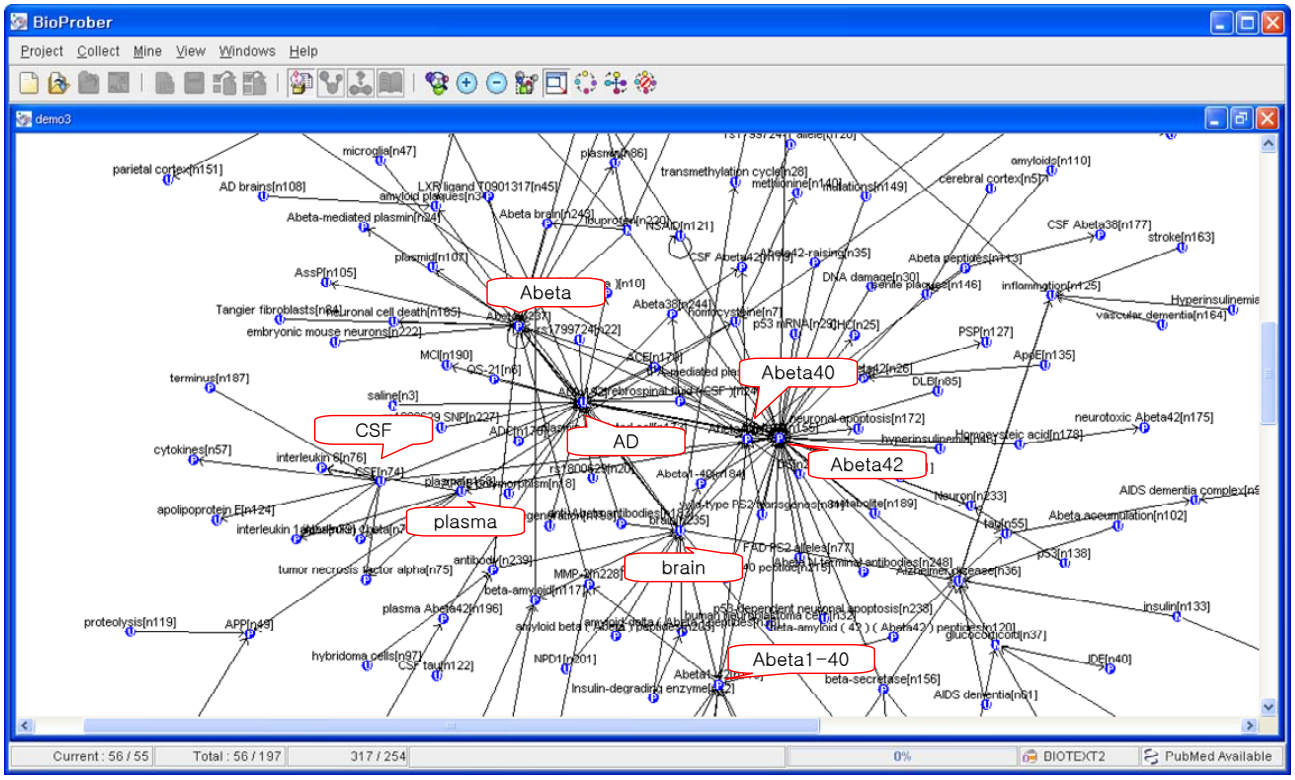

Fig. 5. The screenshot of graph visualization in full size showing there are meaningful relation between beta amyloid protein (Abeta), Abeta40, Abeta42, Alzheimer's Disease (AD) and brain and these entities are key material in AD.

because the syntactic tag set have fewer number of tags than the POS tag set, but not limited to relation types.

C. *Relation Viewer*

The relation viewer accumulates extracted relations and visualizes in graphs and tables as shown in Fig. 4.

When a relation is selected, sentences and abstracts which the relation is extracted from are printed. Recognized entities are displayed in blue font. Extracted relations are displayed in red font and related entities are underlined. If an entity or a relation is extracted from more than one sentence, all those sentences or abstracts are printed. By PubMed ID, original PubMed abstracts can be browsed with an internet browser, like Microsoft Internet Explorer.

III. Implementation

All parts are implemented in JAVA programming language. For collecting PubMed abstracts, we use Entrez Programming Utilities under its user requirements [11]. Entrez Programming Utilities are tools that provide access to Entrez data outside of the regular web query interface and may be helpful for retrieving search results for future use in another environment.

For tagging and parsing sentences, we used Brill's transformation based part-of-speech tagger [12] and Stanford Parser [13]. For our named entity recognition, we used Maximum Entropy models and applied OpenNLP's maxent package [14].

To train our named entity recognizer, we built two sub-domain corpora, one is related to Alzheimer's disease and the other is related to diabetes mellitus, which were collected from MEDLINE database.

IV. Discussion

In visualization of extracted relations, the frequency and each type is not displayed yet. The number of edge linked a node means the quantity of relation for an entity. But when an edge represents a relation between two entities, current graph does not show how many articles mention this relation. We can show its types and frequency in table.

To edit and navigate the network of extracted relations, we have developed another application, called 'bioINET'. Our system exports extracted relations to a file in XML scheme and bioINET imports it and support additional functions, such as searching, editing, navigating and inferring relational networks. Immigration of some basic functions will be done in the near future.

Patent documents and full-texts are valuable resource as PubMed abstracts. Full papers mention more relations than abstracts and patent documents have much information that can not be found in journals.

V. Conclusion

Our system assists biologists and doctors for retrieving

PubMed and acquiring biomedical relation information from PubMed. Our named entity recognizer extracts 40 categories of entities and our relation extractor is not limited to relation types.

REFERENCES

[1] M. Stephens, M. Palakal, S. Mukhopadhyay, R. Raje and J. Mostafa, "Deecting gene relations from Medline abstracts," *Pacific Symposium on Biocomputing,* 6, 2001, pp. 483-495.

[2] T.K. Jenssen, A. Lægreid, J. Komorowski and E. Hovig, "A literature network of human genes for high-throughput analysis of gene expression," *Nature Genetics* 28, 2001, pp.21-28.

[3] L. J. Jensen, J. Saric and P. Bork, "Literature mining for the biologist: from information retrieval to biological discovery," *Nature Reviews Genetics*, 7, 2006, pp. 119-129.

[4] M. Krallinger, R. A. Erhardt and A. Valencia, "Text-mining approaches in molecular biology and biomedicine," *Drug Discovery Today*, v.10, 6, 2005, pp. 439-445.

[5] J. Saric, L. J. Jensen, R. Ouzounova, I. Rojas and P. Bork, "Extraction of regulatory gene/protein networks from Medline," *Bioinformatics* 22, no.6, 2005, pp. 645-650.

[6] N. Daraselia, A. Yuryev, S. Egorov, S. Novichkova, A. Nikitin and I. Mazo, "Extracting human protein interactions from MEDLINE using a full-sentence parser," *Bioinformatics* 20, 2004, pp. 604− 611.

[7] C. Friedman, P. Kra, H. Yu, M. Krauthammer and A. Rzhetsky, "GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles," *Bioinformatics* 17, 2001, S74− S82.

[8] A. Rzhetsky et al., "GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data," *Journal of Biomedical Informatics*, 37, 2004, pp. 43− 53.

[9] H. Jang, Jaesoo Lim, J. Lim, Soo-Jun. Park, S. Park and K. Lee, "Extracting Protein-Protein Interactions in Biomedical Literature Using an Existing Syntactic Parser," *KDLL, LNBI,* 3886, 2006, pp. 78-90.

[10] M. P. Marcus, B. Santorini, M. A. Marcinkiewicz, "Building a large annotated corpus of English: the Penn Treebank," *Computational Linguistics* Vol. 19, 1994, pp. 313-330.

[11] Entrez Utilities. Available: http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html

[12] E. Brill, "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging," *Computational Linguistics* v.21 n.4 (2002) 543-565

[13] Stanford Natural Language Processing Group - Stanford Parser: Available: http://www-nlp.stanford.edu/software/lex-parser.shtml

[14] OpenNLP maxent package. Available: http://maxent.sourceforge.net/