Overview of Biomedical Relations Extraction using Hybrid Rule-based Approaches

Rabiah A.Kadir
Department of Computer Science, UPM, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 Serdang, Selangor
rabiah@fsktm.upm.edu.my

Behrouz Bokharaeian
Natural Interface based on Language (NIL) Group, Universidad Complutense de Madrid, Spain
bokharaeian@gmail.com

***Abstract*—Unstructured text documents are the major source of knowledge in biomedical fields. These huge amounts of information cause very difficult task of extraction or classification. Therefore, there is a need for knowledge discovery and text mining tools in this field. A lot of works have been done on relation extraction in biomedical field. However, each of them was implemented in three major types of techniques separately i.e. co-occurrence, kernel based and rule based methods. There are many variants of these algorithms have been developed but the combination of it has not been verified yet. In this paper we will compare each of those three methods and propose a new combination of relation extraction method between medical and biological entities from biomedical documents. Furthermore, a lot of researches have been done on biomedical binary relation such as protein-protein and gene-protein relations and few researches were on complex relations such as metabolic pathways. However, in this work we will discuss the overview a combination of three methods called as hybrid rule-based to extract complex and simple relations.**

***Index Terms*—text mining, biomedical relation extraction, kernel methods, Rule-base methods**

I. INTRODUCTION

Unstructured text documents are the main means of publishing research in many fields such as biomedical sciences. The MEDLINE 2010 database contains over 19 million records, and the database is currently growing at the rate of 500,000 new citations each year. At the other hand traditional keyword and indexing search methods cannot satisfy researches, and an emergency need for more advanced searching methods that based on natural language processing techniques.

Relation extraction is a subfield of text mining task. It can be categorized as unnamed relations. This provides the associated biomedical terms but does not specify the actual relation; such as, *relation class* - does not specify the relation either but indicate which predefined classes that the relation may fall; *named- relation* - the actual relation among terms and pathways. The course of the cellular or metabolic process that the biomedical substances may affect. Fig. 1 shows the typical framework of biomedical relation systems that involves several steps such as preprocessing; relation extraction based on relevant or irrelevant information before it can extracts the actual relation.

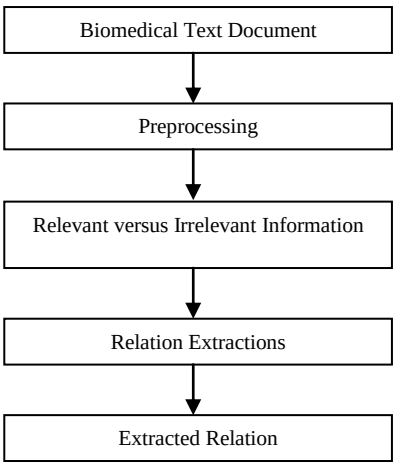

Figure 1. A typical framework of a biomedical relation extraction system

With such explosive growth in biomedical knowledge, it is extremely challenging to keep up-to-date with all of the new discoveries and theories. Since simple indexing and keyword searching cannot satisfy complex searching requirement, automatic methods that can understand human language and identify interesting information are becoming essential in biomedical information management.

The task includes identifying individual terms of biomedical substances, such as diseases, drugs, genes and gene productions and also extracting relevant information of what is expresses or predicted about specific terms, including relations and other hidden information. Some relations can be gene-disease, gene-gene, protein-protein interactions and Disease-Treatment relation.

Manuscript received June 24, 2013; revised August 21, 2013.

doi: 10.12720/jiii.1.3.169-173

A number of works have been done on biomedical relation extraction field, which can be categorized in three major types of techniques: co-occurrence, kernel based and rule based methods. There are many variants of these algorithms which have been developed and studied separately but the combination of them together has not been studied yet.

The aim of this research is to develop a relation extraction method through combination of co-occurrence, rule-based and kernel-based methods to extract biomedical relations from biomedical text documents, and capable to extract complex biomedical relations

Furthermore, there are many research have been done on biomedical binary (simple) relation, such as protein-protein and gene-protein relations, and several research also have been done on complex relations such as metabolic pathways.

II. Related Study

Relation extraction in biomedical relatively has a long history, most of the work done on protein-protein interaction which is interaction of two proteins bind together often to carry out their biological function [1, 2, 3]. Gene-disease [4], disease-symptoms, drug-disease are the other popular relations in this area.

Various algorithms were used but 4 general categories are most popular techniques in relation extraction, such as rule-base [5, 6], kernel methods [3, 7], co-occurrence [8, 9] and natural language parser [2, 10].

Relation extraction can be used at abstract level or full-text level. Abstract level is faster but does not have enough information. Whereas, full-text have much information but increase the time processing. It is also can be implemented in sentence, paragraph and document. In such co-occurrence driven approaches, associations have a higher chance to be true when the co-occurrence of entities is observed in a small amount of text such as a sentence. Whereas, a lower chance to be true when observed in larger amounts such as paragraphs. The results were reported varied, however most of the results is above 50%.

Some researches focus only in one special disease, where [6] is an interesting work that is about extracting mutation relations in HIV virus from biomedical documents. Most of biomedical articles concentrate on two relations but in other area, hierarchy relations extraction also been studied.

III. Relation Extraction

There is no standard terminologies in applying this word in relation extraction tasks, however we include two parts of terms:

Named entity recognition: subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. Named entities are phrases that contain the names of persons, organizations, locations, times and quantities; example:

> *[PER Wolff], currently a journalist in [LOC Argentina], played with [PER Del Bosque] in the final years of the seventies in [ORG Real Madrid].*

Inter-Species Normalization: Normalization helps to link objects of potential interest, such as genes, to detailed information not contained in a publication; it is also key for integrating different knowledge sources. From an information retrieval perspective, normalization facilitates indexing and querying. Gene Normalization (GN) is particularly challenging, given the high ambiguity of gene names: they refer to orthologous or entirely different genes, are named after phenotypes and other biomedical terms, or they resemble common English words.

Two major categories of the relation extractions (*with prior knowledge or without prior knowledge*):

A. Relation Extraction without Prior Knowledge

Three major categories of these methods are co-occurrence, rule base and kernel methods. Co-occurrence methods are one group of techniques that consider how two entities occur with each other. And rule based method uses some handmade rules to find relations from text documents. Kernel based algorithms maps the data into a high dimensional feature space, where each coordinate corresponds to one feature of the data items. Below is the detail description of each major category:

- *Co-occurrence:* use two entities occurrences statistics to detect whether their co-occurrences is due to chance. Such as pointwise mutual information (PMI), ch-square, log-likelihood ratio (LLR) and co-citation.

  Pointwise (Normalized) mutual information, PMI: If X and Y are random variables, the PMI between two possible outcomes, where X=x and Y=y is:

$$PMI(x,y) = \log \frac{\Pr(X=x, Y=y)}{\Pr(X=x)\Pr(Y=y)}$$

  This quantity is zero if x and y are independent, positive if they are positively correlated, and negative if they are negatively correlated.
- *Rule base:* Many efficient and scalable methods for statistical information extraction from large corpora are based on the use of simple extraction patterns. Two rules for extracting presidents of United States: NP1 and the other as President. NP1 was selected President of the United States. Then, three rule for extracting the home states of US presidents will shown as: homeStateOf (Senator, State) , NP1 of NP2; George Washington's home state of NP2; George Washington of NP2

  In the rule base method, a set of rule patterns are designed to help filling the slots. Each rule pattern is a regular expression.

- *Kernel methods:* In many cases, data cannot be easily expressed via features. For example, in most NLP problems, feature based representations produce inherently local representations of objects, for it is computationally infeasible to generate features involving long-range dependencies. Kernel methods are an attractive alternative to feature-based methods. Kernel methods retain the original representation of objects and use the object in algorithms only via computing a kernel function between a pair of objects. A kernel function is a similarity function which satisfies certain properties. More precisely, a kernel function K over the object space X is binary function $K : X \times X \rightarrow [0, \infty]$ mapping a pair of objects $x,y \in X$ to their similarity score $K(x;y)$.

  Convolution kernel: trees, graphs kernels and sequences (relation) kernel are the kernels that we consider represent trees in Terms of their substructures (fragments). These latter define feature spaces, which are mapped into vector spaces. The associated kernel function measures the similarity between two trees by counting the number of their common fragments. For example, Fig. 2 illustrates the syntactic parse of the sentence "Mary brought a cat to school".

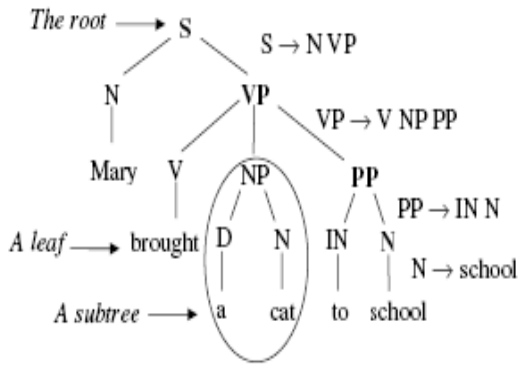

Figure 2. Sample of a semantic parse tree.

The main idea of tree kernels is to compute the number of the common substructures between two trees T1 and T2 without explicitly considering the whole fragment space. Kernel function for two sub tree can be defined as:

$$K(T1,T2) = \sum_{n1 \in N_{T1}}^{n} \sum_{n2 \in N_{T2}}^{n} \Delta(n1,n2)$$

where N_{T1} and N_{T2} are the sets of the T_1's and T_2's nodes, and the delta is equal to the number of common fragments rooted in the n_1 and n_2 nodes.

B. Relation Extraction with Prior Knowledge

In fact humans in recognizing relations are not thus constrained and rely on an abundance of implicit world knowledge or background information. Some propose methods for using knowledge and resources that are external to the target sentence, as a way to improve relation extraction. Exploiting background knowledge such as relationships among the target relations, as well as by considering how target relations relate to some existing knowledge resources such as domain model, kernel PCA, kernel LSI or GVSM/Wordnet.

IV. Propose Method - Hybrid Rule Based

Experimental procedures: We select a standard corpus and build our corpus from biomedical articles. The proposed algorithm will train with human ground true data. The algorithm will test with experimental data. At the first phase we will work on more standard relations such as protein-protein relation or drug-drug relation and then will introduce new biomedical relations.

Variables involve:

- *Independent variables: Relevant relation:* number of relevant relations in corpus. *Relevant sentences:* number of sentences in the corpus that is relevant to relation *Number of sentences:* number of sentences in the corpus
- *Dependent variables: Retrieve relation:* number of retrieved relations. *Retrieved sentence:* number of relevant sentences.

In the propose framework there are some components that we introduced as below:

Combiner: The Combiner (feature generator and selector and extractor) component will select and extract the different features of corpus and sentence with different scores and relations from different relation extractor methods and will normalize the features for further usage.

Classifier: based on normalized data which is obtained from previous stage (combiner) will determine the relation based on normalized features from combiner some suggested features are:

- Size of article or text document in word.
- Average (and variance) of paragraph length in word.
- Average (and variance) of sentence length in word.
- Relative and total frequency of expected items
- Entropy rate of the text
- Bag of words
- Classification of documents

The framework for the proposed method has shown in Fig. 3.

Data Analyzing: This part used to shows that the algorithm has significant improvement compare to traditional co-occurrence and rule pattern system. In this part, we will implement some experiments with two different algorithms. We also will run all three algorithms with different corpuses to compare the performance of them. The relevant and retrieved relations are used parameters in this part.

V. Conclusion

This paper introduced a combination of three methods (i.e.: rule-base, kernel methods and co-occurrence) called as hybrid rule-based to extract complex and simple relations. We presented information about relation extraction in biomedical with unstructured text document

using three major approaches. Then, followed by the description of the implementation of new approach with additional components known as combiner and classifier. Each component has its own function as presented in previous section.

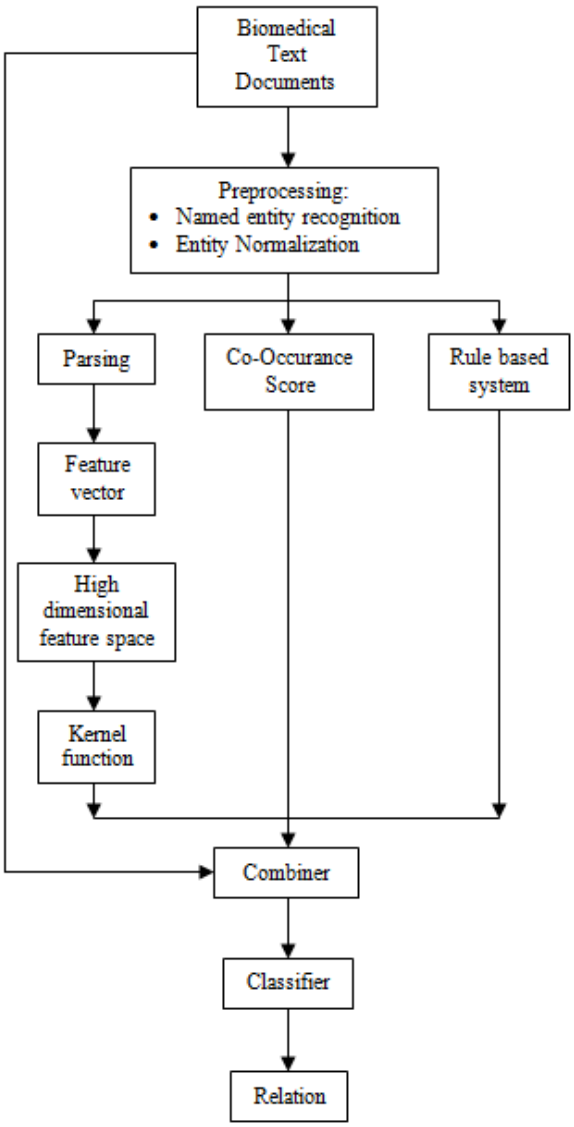

Figure 3. Framework of the proposed relation extractor

Our initial cognitive studies on mechanisms of relation extraction task by human being have proved this hypothesis that human being use these different methods in different situations. Usage of basic statistical rules in addition to language based processes for extracting biomedical relations from biomedical articles is an understandable and investigated usage of our proposed mentioned method utilizes by biomedical researchers.

The proposed method may never be completely implemented, however, further studies to identify potential problems and solution, as well as better experiment, should result in a much more relevant and improve the relations extraction.

ACKNOWLEDGMENT

This work is supported in part by the Ministry of Higher Education (MoHE), Malaysia under Grant LRGS/TD/2011/UITM/ICT/03.

REFERENCES

[1] J. X. Li, "Kernel-Based learning for biomedical relation extraction," *Journal of The American Society For Information Science And Technology*, vol. 59, no. 5, pp. 756–769, 2008.
[2] M. Krallinger, *et al.*, "Evaluation of text-mining systems for biology: Overview of the second biocreative community challenge," *Genome Biology*, vol. 9, no. 2, S1, 2008.
[3] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi, "Banner: An executable survey of advances in biomedical named entity recognition, pacific symposium on biocomputing," vol. 13, pp. 652-663, 2008.
[4] H. M. Müller and A. Rangarajan, "Textpresso for neuroscience: searching the full text of thousands of neuroscience research papers," *Journal of Neuroinform*, vol. 3, pp. 87-94, 2008.
[5] T. Barnickel, J. Weston, R. Collobert, H. W Mewes, Stu ¨Mpflen V, "Large Scale Application Of Neural Network Based Semantic Role Labeling For Automated Relation Extraction from Biomedical Texts," *Plos ONE*, vol. 4, no. 7, E6393. Doi: 10.1371, 2009.
[6] H. J. Dai, Y. C. Chang, T. RTH, *et al.*, "New challenges for biological text-mining in the next decade," *Journal of Computer Science and Technology*, vol. 25, no. 1, pp. 169–179, 2010.
[7] R. Bunescu, R. F. Ge, R. J. Kate, E. M. Marcotte, R. J. Mooney, A. K. Ramani, and Y. W. Wong, "Mutationfinder: A high-performance system for extracting point mutation mentions from text," *Bioinformatics Applications Note*, vol. 23, no. 14, pp. 1862-1865.
[8] T. C. Rindflesch, *et al.*, "EDGAR: Extraction of drugs, genes and relations from the biomedical literature," in *Proc. Pac. Symp. Biocomput.*, vol. 5, pp. 514-525.
[9] Y. Miyao, K. Sagae, R. Sætre, T. Matsuzaki, and J. Tsujii, "Evaluating contributions of natural language parsers to protein–protein interaction extraction," vol. 25, no. 3, pp. 394–400, 2009.
[10] H. Shatkay *et al.*, "Integrating image data into biomedical text categorization," *Bioinformatics*, vol. 22, no. 14, 2006.

Rabiah A. Kadir was borned in East Malaysia, Sarawak year 1969. Sheenrolled her diploma in Computer Science in 1987 after finishing her schooling at College Science Datuk Patinggi Abang Haji Abdillah, Kuching Sarawak, Malaysia. Rabiah furthered her study in first degree of Computer Science year 1990 at Universiti Pertanian Malaysia. In year 1997, she graduated her Masters in Computer Science at Universiti Kebangsaan Malaysia. After several years, she enrolled her PhD in Computer Science with a major field in computational linguistic on December 2003 and completed her study on May 2007. She graduated her PhD from Universiti Kebangsaan Malaysia.During her study in Masters and PhD, she was attached with Universiti Putra Malaysia as a tutor and lecturer respectively. She is a senior lecturer in Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia with a specialization in computational linguistics in Artificial Intelligence Research Group. Currently, she is seconded with Najran Universiti, Kingdom of Saudi Arabia as an Assistant Professor. Her research interests include information retrieval and expert system. Rabiah Abdul Kadir joins Malaysian Information Technology Society (MITs) since year 2008 as Vice Treasurer. She is also a member to the Malaysian Information Retrieval and Knowledge Management Society. She was awarded two gold medals for her PhD research work in BIS and Eureka Exhibition in year 2007. Currently, she had published more than 30 journals and international proceedings.

Behrouz Bokharaeian, BSc in software Engineering, Sharif University Technology (1997-2001), M.Sc in Software engineering, Amirkabir University of Technology, (2001-2003). And also Master in Biomedical Informatics from Amirkabir University of Technology (2005-2008), And Ph.D candidate in Complutense university of Madrid

(2011) Faculty of Computer Science and Information Technology. He worked in artificial intelligence group in Amirkabir University of Technology; Now Mr Behrouz Bokharaeian is a member of Natural interaction based on language research group in Complutense university of Madrid. He worked as a software developer in Iran Gostar Company. He had published several journals and international proceedings.