CoTri: extracting chemical-disease relations with co-reference resolution and common trigger words

Yu-De Chen1, Juin-Huang Ju2, Ming-Yu Chien3, Yu-Cheng Sheng4, Tsung-Lu Lee5, Jung-Hsien Chiang*6

12346Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 70101, Taiwan, 5Department of Information Engineering, Kun Shan University, Tainan 71070, Taiwan

12{chunjoe, jujh}@iir.csie.ncku.edu.tw;
34{imwilly37, ycsheng}@iir.csie.ncku.edu.tw;
5michaelee0407@gmail.com;
*6jchiang@mail.ncku.edu.tw;

Abstract. Drug discovery is an expensive and time-consuming process; however, these could be reduced if existing resources could be analyzed to identify candidates for drug repurposing. Moreover, due to the rapid growth of biomedical literature and the labor-intensive manual text-mining annotation, the process of drug repurposing discovery remained challenging. In this study, we explore the potential chemical-disease relations from biomedical literatures. Our proposed approach includes the following tasks: 1) the NER processing; 2) a co-reference resolution method; 3) a trigger word learning; 4) classification. A great number of novel chemical-disease relations have been discovered using our proposed approaches, which enabled greater insights into drug discovery and drug repurposing for further exploration of chemical-disease mechanisms.

Keywords. Relation extraction; Drug repurposing; Co-reference resolution;

1 Introduction

Chemical and disease are important elements in pathogenesis and treatment of the human body. For example, the cholinergic hypothesis of Alzheimer's disease is simply depicted in Figure 1. The neurotransmitter, acetylcholine, localized in brain which transmits signals across the synapse. The concentration decrease of acetylcholine could lead to Alzheimer's disease. Acetylcholinesterase is an enzyme that hydrolyzes the neurotransmitter namely acetylcholine. Therefore, acetylcholinesterase could cause a concentration decrease of

acetylcholine leading to Alzheimer's disease. According to the above chain reaction, the interactions between chemicals and diseases play an important role in drug discovery.

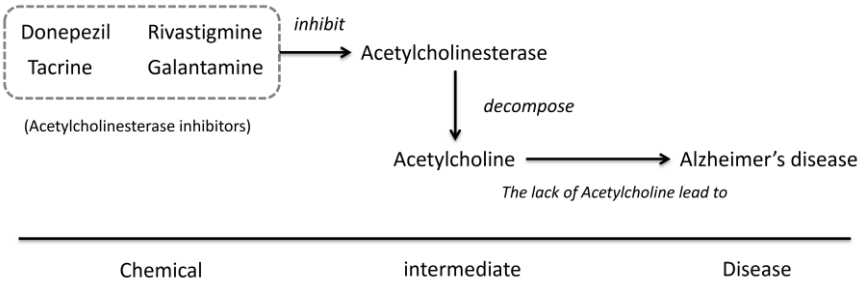

Fig. 1. Mechanism of action of a chemical-disease interaction

Identification of the chemical-disease interactions is one of the key aims in drug discovery. Although there are databases [1, 2, 3] preserving substantial interactions of biomedical entities, many other unidentified interactions could be buried in the biomedical literature. Hence, automated identification of the interactions through biomedical literature repositories could enable the discovery of the hidden interactions.

Yu *et al*. [4] integrated the information of drugs, protein complexes, and diseases from databases to form a tripartite network describing weighted relationships between drugs and diseases to discover their hidden associations. Zhang *et al*. [5] designed a database of structured knowledge extracted from MEDLINE citations to construct potential new drug-drug interactions. Bui *et al*. [6] proposed syntactic features trained with a support vector machine classifier to extract drug-drug interactions from biomedical text. However, none of them was focused on extracting "chemical-disease" relations. The aim of this paper was to design a syntactic feature-based approach along with a trigger learning method to extract chemical-disease relations from the biomedical literature.

2 System Modules

There are two main parts in our approach: 1) pre-processing whole abstract text to sentences data; 2) predicting potential relations in these sentences. In the first part, we processed the text data by conducting

named entity recognition (NER), co-reference resolution, and sentence segmentation sequentially. Then, we generated feature vectors and classified the sentence data by support vector machine (SVM) in the second part. Figure 2 shows an overview of the proposed system. More details of each part will be described below.

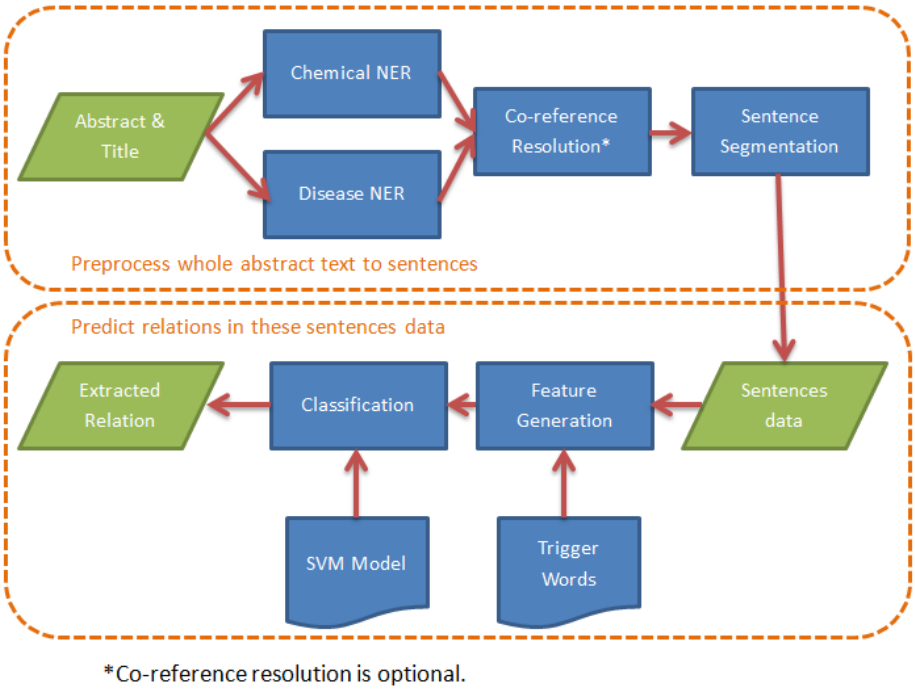

Fig. 2. Overview of the proposed system

2.1 Named Entity Recognition

The first step of the pre-processing is to identify the named entities in an article. We utilized tmChem tool [7] for identifying chemical mentions, and DNorm tool [8] for identifying disease mentions.

2.2 Co-reference Resolution

People may use varying referents to refer to the same things, namely co-reference. The referents may appear in different sentences. To further improve the accuracy of relation extraction, we utilized the Stanford NLP tool [9] to solve these co-reference issues.

2.3 Sentence Segmentation

First, we replaced new line characters ('LF' in ASCII code) with whitespace characters. Then, we divide each abstract into several sentences by using the built-in Java function for sentence segmentation.

2.4 Feature generation

We adapted the feature generation proposed by Bui *et al*. [6]. We extracted the common drug, gene, and disease relation terms as trigger words [10] instead of the original trigger words that were used to extract protein-protein and drug-drug interactions. In addition to these original features, we added another feature: negative expression. We utilized NegEx [11] to determine whether a sentence included a negation expression.

2.5 Classification

To predict a potential chemical-disease pair, we used LIBSVM classifier [12] with RBF kernel in this system. We trained the SVM model with the training set and the development set provided by the CDR task of BioCreative V [13].

3 Experiment

As shown in Table 1, we can see the results of the proposed system evaluated on the testing set of the CDR task of BioCreative V. There are only a few chemical and disease entities detected by the Stanford NLP co-reference resolution tool. The performance improvement is insignificant by co-reference resolution. Because of the limited submission time (30 seconds), we set two thresholds for the number of sentences and the number of words in a sentence. In our additional experiments, we used the training set and the development set of the CDR task of BioCreative V as training data and testing data respectively. It shows that the result with our common trigger words is better than other trigger words in Table 2. Finally, Table 3 shows that the performance of sentence-level approach is better than the abstract-level approach. it is probably because use of tool for co-reference resolution [9] is unable to extract most cross-sentence relations.

Table 1. Performance comparison of co-reference resolution methods

	Precision	Recall	*F*-score
Co-occurrence	0.164	0.765	0.271
Without co-reference resolution	0.417	0.412	0.414
With partial co-reference resolution*	0.418	0.414	0.416
With co-reference resolution**	0.414	0.401	0.407

*We conducted the co-reference resolution only when the number of sentences of an abstract was less than 15 and the number of words in each sentence was less than 30.
**Spend too much time and failed in submission in several test cases.

Table 2. Performance comparison of different trigger word lists

	Precision	Recall	*F*-score
Our trigger words[10]	0.466	0.584	0.518
Bui *et al*. trigger words[6]	0.465	0.543	0.501
Training set trigger words*	0.482	0.270	0.346

*We considered a term as a trigger word if the term occurred in the chemical-disease relations (CDR) more than five times.

Table 3. Performance of abstract-level and sentence-level approaches

	Precision	Recall	*F*-score
Abstract-level	0.466	0.584	0.518
Sentence-level	0.560	0.585	0.572

4 Discussion

In this work, we describe a syntactic feature-based approach to extract chemical-disease relations with co-reference resolution and trigger word detection. The co-reference resolution is slightly helpful in extracting the hidden CDRs. However, it still has a room for improvement. The trigger word learning [10] could further improve the performance. It can be used in different relation extraction tasks. The significant fall in performance of the abstract-level approach indicates that a well-designed co-reference resolution is needed for this task.

5 Acknowledgment

This work was supported by the Ministry of Science and Technology of Taiwan (MOST103-2221-E-006-254-MY2).

REFERENCES

1. Davis, A. P., Murphy, C. G., Johnson, R., Lay, J. M., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B. L., Rosenstein, M. C., and Wiegers, T. C., "The comparative toxicogenomics database: update 2013," Nucleic acids research, p.gks994, 2012.
2. Whirl-Carrillo, M., McDonagh, E., Hebert, J., Gong, L., Sangkuhl, K., Thorn, C., Altman, R., and Klein, T. E., "Pharmacogenomics knowledge for personalized medicine," Clinical Pharmacology & Therapeutics, Vol.92, No.4, pp.414-417, 2012.
3. Zhu, F., Shi, Z., Qin, C., Tao, L., Liu, X., Xu, F., Zhang, L., Song, Y., Liu, X., and Zhang, J., "Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery," Nucleic acids research, Vol.40, No.D1, pp.D1128-D1136, 2012.
4. Yu, L., Huang, J., Ma, Z., Zhang, J., Zou, Y., & Gao, L. (2015). Inferring drug-disease associations based on known protein complexes. BMC Medical Genomics, 8(Suppl 2), S2.
5. Zhang, R., Cairelli, M. J., Fiszman, M., Rosemblat, G., Kilicoglu, H., Rindflesch, T. C., ... & Melton, G. B. (2014). Using semantic predications to uncover drug–drug interactions in clinical data. Journal of biomedical informatics, 49, 134-147.
6. Bui, Q. C., Sloot, P. M., van Mulligen, E. M., & Kors, J. A. (2014). A novel feature-based approach to extract drug–drug interactions from biomedical text. Bioinformatics, btu557.
7. Leaman, Robert, Chih-Hsuan Wei, and Zhiyong Lu. "tmChem: a high performance approach for chemical named entity recognition and normalization." *Journal of cheminformatics*, 2015.
8. Leaman, Robert, Rezarta Islamaj Doğan, and Zhiyong Lu. "DNorm: disease name normalization with pairwise learning to rank." *Bioinformatics*, 2013.
9. Recasens, Marta, Matthew Can, and Daniel Jurafsky. "Same Referent, Different Words: Unsupervised Mining of Opaque Coreferent Mentions." HLT-NAACL, 2013.
10. Jiun-Huang Ju and Jung-Hsien Chiang (2014), "Identifying drug-gene-disease interactions and an application to explore the undiscovered networks," The 25th International Conference on Genome Informatics (GIW 2014), Tokyo, Japan, December 15-17, 48.
11. Chapman, Wendy W., et al. "A simple algorithm for identifying negated findings and diseases in discharge summaries." *Journal of biomedical informatics*, 34.5 (2001): 301-310.
12. Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: A library for support vector machines." ACM Transactions on Intelligent Systems and Technology (TIST), 2011.
13. Li J, Sun Y, Johnson R. et al. (2015) Annotating chemicals, diseases, and their interactions in biomedical literature, in Proceedings of the fifth BioCreative challenge evaluation workshop, Sevilla, Spain