Chemical-disease Relations Extraction Based on The Shortest Dependency Path Tree

Huiwei Zhou*1, Huijie Deng1, Jiao He2

1 School of Computer Science and Technology, Dalian University of Technology, China
2 Library , The First Affiliated Hospital of Dalian Medical University, China

*1 zhouhuiwei@dlut.edu.cn;
1 denghuijie@mail.dult.edu.cn;
2 hejiao_barton@163.com

Abstract. Identifying chemical-disease relations (CDR) from biomedical literature could improve chemical safety and toxicity studies. This paper proposes a Shortest Dependency Path Tree (SDPT) to capture the most direct syntactic and semantic relationship between chemical and disease. Based on SDPT, structured dependency features (SDF), structured phrase features (SPF) and flattened dependency features (FDF) are proposed to represent syntactic information between two entities, which are all effective for CDR. Experiments on the CDR training and developing dataset show that our method achieves 55.05% F1-score.

Keywords: Chemical-disease Relations Extraction; Syntactic Information; Shortest Dependency Path Tree

1 Introduction

The BioCreative V proposes a challenge task of automatic extraction chemical-disease relations (CDR) from the biomedical literature in support of new drug discovery and drug safety surveillance. There are two specific subtasks: (1) Disease Named Entity Recognition and Normalization; (2) Chemical-induced diseases relation extraction. This paper focuses on the subtask (2).

Relation extraction (RE) aims at identifying instances of pre-defined relation type in text [1-5]. Generally, machine-learning based RE approaches can be divided into two categories: feature-based and kernel-based methods. Feature-based methods focus on defining flattened features ranging from lexical to syntactic and semantic information. Kernel-based methods exploit structured representations of instances. Tree kernel [6] is one of the most commonly used kernels, which could capture the structured syntactic connection information between the two entities. The effective representations of relation

instances have been studied [7-9]. Zhang et al. [7] investigate five tree spans of a phrase tree for general RE task of ACL, among which the Path-enclosed Tree (PT) achieves the best performance. Phrase tree represents constituents of neighbours, which is suitable for capturing local syntactic information. For two entities over long distance, phrase representations will carry noisy and influence the performance of relation extraction.

The chemical and disease entities in a sentence are usually over long distance. Dependency structure reflects semantic modification relationships of words in a sentence, which compactly represent global syntactic information. To grasp global syntactic information connecting chemical and disease entities, this paper presents a Shortest Dependency Path Tree (SDPT), which could represent the most direct syntactic and semantic relationship between two entities. Based on SDPT, structured dependency features (SDF), structured phrase features (SPF) and flattened dependency features (FDF) are presented to represent syntactic information. These features are integrated by composite kernel [10]. Experiments on the CDR training and developing dataset show that our methods achieve 55.05% F1-score.

2 Discussion

In this section, we describe the methods of obtaining syntactic information, and present the experimental results on the CDR dataset.

Methods

To simplify the CDR problem, we ignore CDR over sentences and only identify CDR in a sentence. Each chemical and disease pair in a sentence is regarded as a candidate instance. In the following subsections, we describe SDPT, SDF, SPF and FDF based on SDPT. Besides, we also employ widely used basic features to further improve the performance of CDR extraction.

Shortest Dependency Path Tree (SDPT).
SDPT is the shortest path sub-tree linking two entities in dependency tree. Taking the sentence 1 as an example, there are a chemical entity denoted in wave line and four disease entities denoted in underline. The chemical entity "*fentanyl*" is associated with the four disease entities.

Sentence 1: *Various reported side effects of fentanyl administration include chest wall rigidity, hypotension, respiratory depression, and bradycardia.*

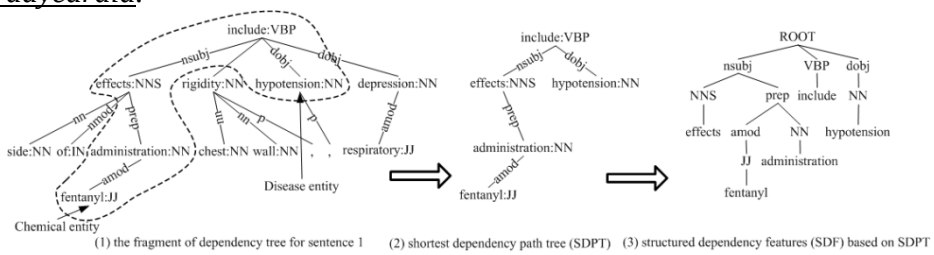

Fig. 1. Shortest dependency path tree (SDPT)

For the fragment of dependency tree for sentence 1 shown in Fig.1 (1). SDPT of the candidate "*fentanyl*" and "*hypotension*" is shown in Fig.1 (2). SDPT is the most direct syntactic representation connecting the two entities

Structured Dependency Features (SDF) based on SDPT.

For the SDPT shown in Fig.1 (2), tree kernel cannot capture dependency relation on the arcs (e.g., "dobj" relation between node "include" and "hypotension"). To capture dependency relation, we use the dependency relation labels to replace the corresponding word and PoS pairs on the nodes of original SDPT as shown in Fig.1 (3). And then, make the PoS tags as the children of the corresponding relation nodes, the fathers of their associated words.

Structured Phrase Features (SPF) based on SDPT.

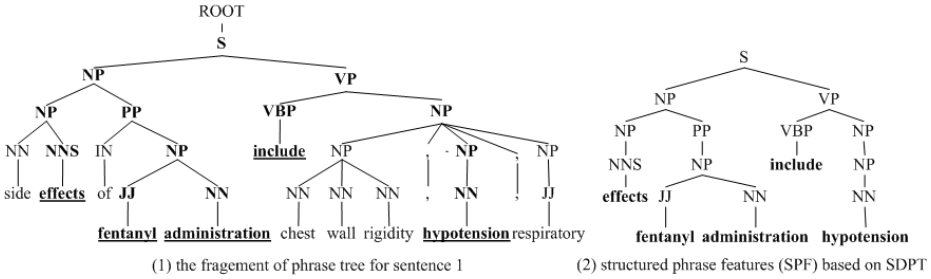

Fig. 2. Structured phrase features (SPF) based on SDPT

To capture constituents and exclude redundancy of two entities with long distance, we propose SPF based on SDPT. For the fragement of phrase tree for sentence 1 shown in Fig.2 (1), SPF of the candidate "*fentanyl*" and "*hypotension*" is shown in Fig.2 (2). SPF is a sub-tree

consisting of the words in SDPT (denoted in underline in Fig.2 (1)) and their ancestral constituents (denoted in bold).

Flattened Dependency Features (FDF) based on SDPT.
FDF based on SDPT contain keyword features and root features:

- Keyword features:

Trigger: whether SDPT contains any trigger word, e.g. *alter*, *effect*.
Negation: whether SDPT contains any negative word, such as *not*, *no*.
Trig&Neg: the combination features of Trigger and Negation features.

- Root features:

Position: the root word locates before, between, or after the two entities.
Context: word, PoS and chunk features in the window [-1, 1].

Basic Features.

- Entity: word, PoS and chunk of two entities in the window [-3, 3].
- Distance: the number of words between two entities.
- Number of Verbs: The number of verbs before, between and after the two entities.

Experimental Results

We use the CDR training and developing dataset [11-12] for training and testing respectively. Disease and chemical entity recognition are accomplished with tmChem [13] and Dnorm [14-15] toolkits. Berkeley Parser[1], Gdep Parser[2] and GENIA Tagger[3] are employed to get phrase tree, dependency tree and lexical information, respectively. SVM-LIGHT-TK 1.2 toolkit[4] is used, which provides polynomial kernel and tree kernel to capture flattened and structured information respectively.

Effects of syntactic representation based on SDPT.
Table 1 lists the performances of FDF, SDF, and SPF derived from SDPT. From the results, we can see that adding FDF to basic features, the F1-score is improved by 1.15%. The sole SDF with tree kernel performs better than complicated basic features. Combination of SDF and SPF can further improve the performance. These indicate that FDF,

[1] Available: http://code.google.com/p/berkeyparser/
[2] Available:http://pepple.ict.usc.edu/ sagae/parser/gdep/
[3] Available:http://www-tsujii.is.s.utokyo.ac.jp/GENIA/tagger/
[4] Available:http://disi.unitn.it/moschitti/Tree-Kernel.htm

SDF and SPF derived from SDPT are effective for CDR extraction. Combining flattened and structured features with composite kernel achieves significantly higher F1-scores compared to the sole flattened and sole structured features.

Table 1. Effects of syntactic representation based on SDPT

Features		P%	R%	F1 %
Flattened	Basic	56.22	48.13	51.86
	+ FDF	57.05	49.51	53.01
Structured	SDF	58.54	48.32	52.94
	+SPF	58.81	48.72	53.29
Combined	Basic+FDF+SDF+SPF	58.63	51.87	55.05

Comparision with other structured syntactic representation.
We compare our SDF with the other structured syntactic representation. The Path-enclosed Tree (PT) [7] is adopted for CDR, which performs worse than SDF as shown in Table 2. In addition, SDF are extended with the dependent nodes of all nodes in SDPT to enrich the context information. From Table 2, we can see that the extending SDPT is much worse than SDF. This indicates that SDF could provide the useful semantic and structured syntactic connecting the two entities.

Table 2. Comparision with other structured syntactic representation

Structured Features	P%	R%	F1%
SDF	58.54	48.32	52.94
PT	58.27	44.48	50.45 (-2.49)
Extending SDPT	56.22	41.42	47.70 (-5.24)

3 Acknowledgment

This research is supported by National Natural Science Foundation of China (Grant No. 61272375 and 61173100).

REFERENCES

1. Chowdhury, Md Faisal Mahbub, and Alberto Lavelli. "Combining tree structures, flat features and patterns for biomedical relation extraction." Paper presented at proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Stroudsburg, April, 2012.

2. Chowdhury, Md Faisal Mahbub, and Alberto Lavelli. "Impact of Less Skewed Distributions on Efficiency and Effectiveness of Biomedical Relation Extraction." Paper presented at COLING (Posters), December, 2012.
3. Chowdhury, Faisal Mahbub, Alberto Lavelli, and Alessandro Moschitti. "A study on dependency tree kernels for automatic extraction of protein-protein interaction." Paper presented at proceedings of BioNLP 2011 Workshop. June, 2011.
4. Chowdhury, Md Faisal Mahbub, and Alberto Lavelli. "Exploiting the Scope of Negations and Heterogeneous Features for Relation Extraction: A Case Study for Drug-Drug Interaction Extraction." Paper presented at proceedings of HLT-NAACL, Atlanta, Georgia, June 9-14, 2013.
5. Chowdhury, Md Faisal Mahbub, and Alberto Lavelli. "FBK-irst: A multi-phase kernel based approach for drug-drug interaction detection and classification that exploits linguistic information." Paper presented at Second Joint Conference on Lexical and Computational Semantics, Atlanta, Georgia, USA, June 14-15, 2013.
6. Moschitti, Alessandro. "A study on convolution kernels for shallow semantic parsing." Paper presented at Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004.
7. Zhang, Min, et al. "A composite kernel to extract relations between entities with both flat and structured features." Paper presented at Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Sydney, July 2006.
8. Zhou, GuoDong, et al. "Tree kernel-based relation extraction with context-sensitive structured parse tree information." Paper presented at Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, June 2007.
9. Qian, Longhua, et al. "Tree Kernel-Based Semantic Relation Extraction using Unified Dynamic Relation Tree." Paper presented at Advanced Language Processing and Web Information Technology, 2008.
10. Zhou, Huiwei, et al. "Hedge Scope Detection in Biomedical Texts: An Effective Dependency-Based Method." PloS ONE 10.7 (2015): e0133715.
11. Wei CH, Peng Y, Leaman R, et al. "Overview of the BioCreative V Chemical Disease Relation (CDR) Task", Paper presented at Proceedings of the fifth BioCreative challenge evaluation workshop, Sevilla, Spain, 2015.
12. Li J, Sun Y, Johnson R. et al. "Annotating chemicals, diseases, and their interactions in biomedical literature." Paper presented at Proceedings of the fifth BioCreative challenge evaluation workshop, Sevilla, Spain, 2015.
13. Leaman R, Wei C-H, Lu Z. "tmChem: a high performance tool for chemical named entity recognition and normalization." Journal of Cheminformatics, 7(Suppl 1): S3, 2015.
14. Robert Leaman, Rezarta Islamaj Doğan and Zhiyong Lu, "DNorm: Disease Name Normalization with Pairwise Learning to Rank." Bioinformatics (2013) 29 (22): 2909-2917, doi:10.1093/bioinformatics/btt474.
15. Robert Leaman and Zhiyong Lu. "Automated Disease Normalization with Low Rank Approximations. " Paper presented at Proceedings of BioNLP 2014.