Chemical-induced Disease Relation Extraction with Lexical Features

Jinghang Gu*, Longhua Qian, Guodong Zhou

School of Computer Science and Technology, Soochow University

*gujinghangnlp@gmail.com;
qianlonghua@suda.edu.cn;gdzhou@suda.edu.cn

Abstract. This paper briefly describes our basic work on the chemical-induced disease relation extraction task on BioCreative-V Track-3b. It is a machine learning based system which utilizes simple yet effective lexical features. Pairs of chemical and disease mentions are first constructed as relation instances for training and testing, then we merge classification results to acquire final relationships between chemicals and diseases. Our system achieves 57.7% precision and 53.2% recall on the development set with a maximum entropy model.

Keywords. Relation Extraction, Machine Learning, Maximum Entropy

1 Introduction

With the rapid accumulation of scientific literature, there is an increasing interest in extracting semantic relations between chemicals and diseases described in text repositories, as they play important roles in many areas of healthcare and biomedical research [1-3]. Despite some previous attempts [4-6] have been done on free-text dataset, automatic biomedical information extraction from identifying relevant biomedical concepts [7-9] to extracting relations [10], still remains challenging.

To address these issues, BioCreative-V community proposes a challenging task of automatic extraction of mechanistic and biomarker chemical-disease relations from the biomedical literature [11]. The task is aimed to provide practical benefits to biocuration, and contains two specific subtasks: (A) Disease Named Entity Recognition and Normalization (DNER); (B) Chemical-induced diseases relation extraction (CID).

In particular, the CID relation is determined between two entities, i.e. chemical and disease, which means the relation can not only derive from text stated in one sentence, but also from text spanning several sentences. Since chemical entities and disease entities may have several mentions in different sentences, the CID relations are interpreted in entity level. We take the case as "*Co-occurrence*" where entities of different types occur in the same sentence with separate mentions of each other; if the entities do not appear in the form of co-occurrence, we take it as "*Non Co-occurrence*". CID relation extraction task can be broken down from entity level to mention level, taking the following sentences into consideration:

(a) *After 2 individuals with psoriasis developed a* ***capillary leak syndrome*** *following treatment with oral* ***sirolimus*** *lesional skin cells and activated peripheral blood cells were analyzed for induction of apoptosis.*

(b) *OBSERVATIONS: A keratome skin specimen from 1 patient with* ***sirolimus****-induced* ***capillary leak syndrome*** *had a 2.3-fold increase in percentage of apoptotic cells (to 48%) compared with an unaffected sirolimus-treated patient with psoriasis (21%).*

(c) *Because patients with severe psoriasis may develop capillary leak from various systemic therapies, clinical monitoring is advisable for patients with* ***inflammatory*** *diseases who are treated with immune modulators.*

These example sentences are extracted from the same document (PMID: 10328196). Among them, the text in bold are mentions of chemical or disease entities, where "*sirolimus*" stands for a chemical entity whose concept identifier is D020123 (C1), "*capillary leak syndrome*" stands for a disease entity whose concept identifier is D019559 (D1), and "*inflammatory*" stands for another disease entity with concept identifier of D007249 (D2). The chemical of C1 has *Co-occurrence* with the disease of D1 in (a) and (b) respectively, while it has *Non Co-occurrence* with the disease of D2. Between C1 and D1, there is a true CID relation.

In this paper we report our approach to the CID relation extraction task of BioCreative IV Track-3b. Our primary goal is to develop a relation extraction system that can scale well over free text documents using machine learning methods. We first extract CID relations in mention level using a classifier with lexical features, then we merge these results to acquire final CID relations between entities. One

assumption is that a pair of entities may have multiple pairs of mentions with *Co-occurrence* or *Non Co-occurrence*, and if at least one pair of these mentions explicitly supports the CID relationship, we believe the two entities have the true CID relation.

2 Methodology

We adopt a supervised learning method to extract CID relations. The system takes raw text documents in Pubtator [12] format as input, then employs preprocessing techniques on it. It extracts CID relations in mention level using a maximum entropy model, and then merges the classification results to acquire final relations between entities. The whole process of our approach can be divided into sequential steps as follows:

1. Pre-processing: In this step, we initially employ several Natural Language Processing techniques in text analysis, including sentence splitting, tokenization, and part-of-speech tagging.

2. Relation instance construction: In this step, a pair of chemical mention and disease mention in the form of <*chemical mention*, *disease mention*>, is constructed as a relation instance through several filtering rules for both training and testing. The construction process discriminates between *Co-occurrence* and *Non Co-occurrence* by leveraging different filter strategies. The strategies for *Co-occurrence* and *Non Co-occurrence* are detailed in section 2.1 and section 2.2 respectively.

3. Feature extraction: The goal of this step is to find out relevant features that may help us reliably capture the CID relations between chemicals and diseases situated in text. The details of features used for *Co-occurrence* and *Non Co-occurrence* will be discussed in section 2.3 and section 2.4 respectively.

4. CID relation extraction: This step consists of training and testing on the dataset. First, we use training instances of *Co-occurrence* and *Non Co-occurrence* to train two classifiers respectively, then we classify the testing instances by applying these classifiers.

5. CID relation merging: After the extraction in mention level, we merge the results to acquire final relations between entities according to the assumption mentioned above.

2.1 Relation Instance Construction for Co-occurrence

Before the phase of relation extraction, instances for *Co-occurrence* for both training and testing should be constructed. We use some simple and effective rules as following to filter the instances:

a) The token distance between the two mentions in an instance should be less than *k* (here we set *k* to 10);
b) If there are multiple mentions in the sentence that refer to the same chemical or disease entity, we take the pair of the nearest chemical mention and disease mention as the instance;
c) Any mention that occurs in brackets should be omitted.

2.2 Relation Instance Construction for Non Co-occurrence

The relation instance construction for *Non Co-occurrence* complies with the following rules:

a) Only the entities that do not have any *Co-occurrence* instance with other entities are taken into consideration for *Non Co-occurrence*;
b) The sentence distance between the mentions in an instance should be less than *n* (here we set *n* to 3);
c) If there are multiple mentions that refer to the same chemical or disease entity, take the pair of the nearest chemical mention and disease mention as the relation instance.

2.3 Feature extraction for Co-occurrence

- CBOW: Bag-of-words of chemical mention
- CPOS: Part of speech of chemical mention
- DBOW: Bag-of-words of disease mention
- DPOS: Part of speech of disease mention
- WVNULL: when no verb in between
- WVFL: when only one verb in between
- WVF: first verb in between when at least two verbs in between
- WVL: last verb in between when at least two verbs in between
- WVO: other verbs in between except first and last verbs
- WBF: first word in between when at least two words in between
- WBL: last word in between when at least two words in between
- WBNULL: when no word in between
- WBFL: when only one word in between
- BM1F: first word before the first mention
- BM1L: second word before the first mention

• AM2F: first word after the second mention
• AM2L: second word after the second mention

2.4 Feature extraction for Non Co-occurrence

• CBOW: Bag-of-words of chemical mention
• CPOS: Part of speech of chemical mention
• DBOW: Bag-of-words of disease mention
• DPOS: Part of speech of disease mention
• SDIST: Sentence distance between chemical mention and disease mention
• CFRQ: Chemical frequency in document
• DFRQ: Disease frequency in document
• DBOW: Bag-of-words of sentence where mentions located
• NSE: Number of entity in the sentence where mentions located
• SMBLOCK: Whether the chemical and disease mentions occur in the same text block, e.g. "BACKGROUND" section, "CONCLUSIONS" section

3 Experimental Results

In this paper, a simple tokenizer is implemented which breaks tokens into either a contiguous block of letters and/or digits or a single punctuation mark. We also use Stanford NLP Tools [13] for sentence splitting, part-of-speech tagging and lemmatization. The gold annotation on chemical and disease entities from the original dataset is used as the input for our system. Relation instances are constructed from the training and development dataset respectively. The classification tool utilized is Mallet MaxEnt [14].

Table 1 shows the performance of instance classification for Co-occurrence and Non Co-occurrence in mention level and the final results for CID relation extraction after merging in entity level.

Table 1 System Performance

Results	P(%)	R(%)	F(%)
Co-occurrence	67.1	58.6	62.3
Non Co-occurrence	45.1	32.0	37.4
Final CID Relation	57.7	53.2	55.3

A close investigation to the results suggests that lexical features are simple but quite effective, especially in *Co-occurrence*. But when compared with the results of *Non Co-occurrence*, the usefulness of lexical features is limited. This is probably because the CID relations that stated in *Non Co-occurrence* span several sentences and have much

more complicated structure which the traditional lexical features cannot capture effectively.

4 Conclusion

In this paper we have described a chemical-induced disease relation extraction system. We found machine learning method with lexical features for CID task to be effective in overall. In the future, we plan to include richer features and incorporate more data from publicly available databases to achieve better results.

REFERENCES

1. Islamaj Dogan, R., Murray, G.C., Neveol, A., et al. Understanding PubMed user search behavior through log analysis. Database (Oxford), 2009, bap018.
2. Lu, Z. PubMed and beyond: a survey of web tools for searching biomedical literature. Database (Oxford), vol. 2011, baq036.
3. Neveol, A., Islamaj Dogan, R., Lu, Z. Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. J Biomed Inform, 44, 310-318.
4. Davis, A.P., Grondin, C.J., Lennon-Hopkins, K., et al. The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. Nucleic Acids Res, 2014 Oct 17,gku935.
5. Xu, R., Wang, Q. Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature. J Biomed Inform.
6. Kang, N., Singh, B., Bui, C., et al. Knowledge-based extraction of adverse drug events from biomedical text. BMC Bioinformatics, 15, 64.
7. Gurulingappa, H., Mateen-Rajput, A., Toldo, L. Extraction of potential adverse drug events from medical case reports. Journal of biomedical semantics, 3, 15.
8. Leaman, R., Wei, C.H., Lu, Z. tmChem: a high performance approach for chemical named entity recognition and normalization. Journal of Cheminformatics 2015, 7(Suppl 1):S3
9. Leaman, R., Islamaj Dogan, R., Lu, Z. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics, 29, 2909-2917.
10. Xu R, Wang Q. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC bioinformatics, 2013, 14(1): 181.
11. Wei CH, Peng Y, Leaman R, et al. (2015) Overview of the BioCreative V Chemical Disease Relation (CDR) Task, in Proceedings of the fifth BioCreative challenge evaluation workshop, Sevilla, Spain
12. Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Research 41:W518-W522 (2013).
13. Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and McClosky, David. The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.
14. McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.