HITSZ_CDR System for Disease and Chemical Named Entity Recognition and Relation Extraction

Haodi Li, Qingcai Chen, Kai Chen, Buzhou Tang*

Intelligent Computing Research Center,
Harbin Institute of Technology Shenzhen School, China

haodili.hit@gmail.com; qingcai.chen@gmail.com;
ck_ican@163.com; tangbuzhou@gmail.com

Abstract. In this paper, an end-to-end machine learning-based system was proposed for the challenge task of chemical and disease named entity recognition (DNER) and chemical-induced diseases (CID) relation extraction in BioCreative V, where DNER includes chemical and disease mention recognition (CDMR) and normalization (CDN). The system consists of six components: a preprocessing module, two individual sequence labeling module, an ensemble module, a normalization module and a CDR extraction module. The two sequence labeling modules and the ensemble module were designed for CDMR. Evaluation using the challenge corpus showed that our system achieved the highest F1-scores of 86.76% on CDMR, 67.82% on CDN and 41.26% on CID relation extraction, respectively.

Keywords. Chemical and disease named entity recognition, chemical-induced diseases relaiton extraction, sequence labeling, ensemble learning;

1 Introduction

As chemicals (or drugs), diseases, and their relations play important roles in many areas of biomedical research and healthcare such as drug discovery and safety surveillance, they have attracted considerable attention in recent years. Automatic chemical and disease recognition and chemical-disease relation (CDR) extraction has become the main direction on this topic. Despite some attempts, few automatic tools are freely available. CDR extraction remains challenging.

Through BioCreative V, a challenge task of automatic extraction of mechanistic and biomarker CDRs from the biomedical literature in

* Corresponding author

support of biocuration, new drug discovery and drug safety surveillance was proposed to advance text-mining research on relationship extraction and provide practical benefits to biocuration [1]. This task included two subtasks: chemical and disease mention recognition and normalization (DNER) and chemical-induced diseases (CID) relation extraction.
We developed an end-to-end machine learning-based system for the challenge of automatic extraction of mechanistic and biomarker CDRs, including a stacked ensemble subsystem for chemical and disease mention recognition (CDMR), a re-ranking subsystem for chemical and disease normalization (CDN) and a ranking subsystem for CID relation extraction. Evaluation on the corpus of the challenge shows that our system achieves the highest F1-scores of 86.76 on CDMR, 67.82% on CDN and 41.26% on CID relation extraction, respectively.

2 Methods

Dataset

The CDR task organizers of BioCreative V manually annotated 1500 PubMed records, of which 1000 records were used as training and development sets, and the remaining 500 records were used as a test set. Not only disease and chemical mentions with MeSH identifiers (IDs), but also CID pairs with relations were marked up [2]. Figure 1 shows an example of annotation records.

Title: Cardiovascular complications associated with terbutaline treatment for preterm labor.
Abstract: Severe cardiovascular complications occurred in eight of 160 patients treated with terbutaline for preterm labor. Associated corticosteroid therapy and twin gestations appear to be predisposing factors. Potential mechanisms of the pathophysiology are briefly discussed.

NER:

Position		Mention	Label	MeSH identifier
Start	End			
0	28	Cardiovascular complications	Disease	D002318
45	56	terbutaline	Chemical	D013726
71	84	preterm labor	Disease	D007752
93	121	cardiovascular complications	Disease	D002318
169	180	terbutaline	Chemical	D013726
185	198	preterm labor	Disease	D007752

CID pairs:

Chemical MeSH identifier	Disease MeSH identifier
D013726	D002318

Fig.1 Example of annotation records.

Overview of system

Our system, as shown in Figure 2, consists of six components: a preprocessing module, two individual sequence labeling module, an ensemble module, a normalization module and a CID relation extraction module. Given a record with title and abstract, the preprocessing module first finished sentence boundary detection and tokenization. Then the two individual sequence labeling modules extracted chemical and disease mentions. Subsequently, the ensemble module used a stacked ensemble learning method to combine the predictions of the previous two sequence labeling modules. After that, the normalization module linked each extracted mention to a MeSH ID. Finally, the CID relation extraction module found out between which chemicals and diseases there have CID relations.

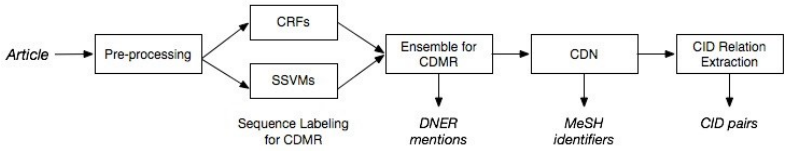

Fig 2. Overview architecture of our CDR system.

Sequence Labeling for CDMR

In our study, CDMR was recognized as a sequence labeling problem. Two individual sequence labeling modules: CRF [3] and SSVM [4] were employed for CDMR. They used the same features as shown in Table 1.

Table 1. Features used in the two sequence labeling modules

Feature	Description
Bag-of-words	Unigrams: w_0, w_{-1}, w_1,w_{-2}, w_2; Bigrams: $w_{-2}w_{-1}, w_{-1}w_0, w_{-1}w_0$, w_0w_1, w_0w_1, w_1w_2; Trigrams: $w_{-2}w_{-1}w_0, w_{-1}w_0w_1, w_0w_1w_2$
POS tags	Unigrams: $p_0, p_{-1}, p_1, p_{-2}, p_2$ Bigrams: $p_{-2}p_{-1}, p_{-1}p_0, p_{-1}w_0$, w_0w_1, w_0w_1, w_1w_2; Trigrams: $w_{-2}w_{-1}w_0, w_{-1}w_0w_1, w_0w_1w_2$
Combinations of tokens and POS	$p_{-2}w_{-1}, p_{-1}w_0, p_0w_1, p_1w_2; w_{-2}p_{-1}, w_{-1}p_0, w_0p_1, w_1p_2$
Sentence information	Length of the current sentence; Whether there is any bracket unmatched in the current sentence.

Affixes	Prefixes and suffixes of the length from 1 to 5.
Orthographical features	Whether the current word is an upper Caps word, contains a digit or not, has uppercase characters inside, etc.
Word shapes	Any or consecutive uppercase character(s), lowercase character(s), digit (s) and other character(s) in the current word is/are repaced by 'A', 'a', '#' and '-' respectively.
Section information	Which section the current word belongs to, title or abstract?
Word representation features[5]	Brown clustering (https: //github.com/percyliang/brown-cluster); Word2vec (https://code. google.com/p/word2vec/)
Dictionary features	Chemical dictionary: The comparative toxicogenomics database (CTD) [6], drugbank [7], Medical subject headings (MeSH) [8] Pharmacogenetics Knowledge Base (PharmGKB) [9], The unified medical language system (UMLS) [10], and Wikipedia; Disease dictionary: CTD, MeSH, UMLS, disease ontology [11], National Drug File Reference Terminology (NDF-RT) [12] and Wikipedia.
Frequency features	If the frequency of the current word is higher than a given value (4 in our system) and the inverse document frequency of it is less than another given value (0.1 in our system)?
Character N-grams	Character N-grams (N=1s, …, 4) within the current word.

Stacked Ensemble for CDMR

A meta-classifier based on Support Vector Machines (SVMs) with a linear kernel was used to ensemble the mentions predicted by the previous sequence labeling modules by checking whether any predicted mention was correct. A variety of features were used to describe the agreement and consistency between the previous modules. Each concept predicted by a sequence labeling module was compared with all other concepts predicted in the same sentence. For each pair of concepts, we extracted eight features from the text spans, such as "if the text spans match", "if the text spans partially match (any word overlap)", "if the text spans have the same start position" and "if one text span subsumes the other". Furthermore, given a mention, how many modules predicted it and which module predicted it were also taken into account.

Normalization Module

A re-ranking subsystem was used for normalization. Firstly, we normalized mentions, especially abbreviations, according to the context of documents, and then employed the MeSH normalization module and Wikipedia to generate candidates for ranking. Finally, we ranked the candidates using SVM-rank and regarded the first-ranked item as the normalized name with a Mesh ID. The features used for candidate ranking includes: Bag-of-words; Similarity between a candidate and its mention; Similarity between a candidate and the normalized mention according to the context of documents; Whether a candidate generated by MeSH, Wikipedia or both of them; Place of a candidate in the rank list of MeSH; Place of a candidate in the rank list of Wikipedia;Type of a candidate generated by MeSH.

CID relation extraction module

We used a linear SVM classifier for CID relation extraction according to the context between CID pairs. The features used in this classifier included bag-of-words, the number of other mentions between a CID pair and chemical-disease relation based on CTD.

3 Results

In the challenge, we were allowed to submit three runs for DNER and CID relation extraction, respectively. The best results of our system on CDMR, CDN and CID relation extraction were shown in Table 2. The highest F1-scores of our systems were 86.76% on CDMR (ranking first), 67.82% on CDN and 41.26% on CID relation extraction.

Table 2. Best results of our system for the challenge (%).

Task		Precision	Recall	F1-score
DNER	CDMR	89.21	84.45	86.76
	CDN	71.76	64.29	67.82
CID relation extraction		54.46	33.21	41.26

Acknowledgment

This paper is supported in part by grants: NSFCs (National Natural Science Foundation of China) (61402128, 61173075 and 61272383), Strategic Emerging Industry Development Special Funds of Shenzhen (ZDSY20120613125401420 and JCYJ20120613151940045).

REFERENCES

[1] 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
[2] 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
[3] Okazaki, Naoaki. CRFsuite: A Fast Implementation of Conditional Random Fields (CRFs), 2007.
[4] Bertelli L, Yu T, Vu D, et al. Kernelized structural SVM learning for supervised object segmentation[C],Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011: 2153-2160.
[5] Tang B, Cao H, Wang X, et al. Evaluating word representation features in biomedical named entity recognition tasks[J]. BioMed research international, 2014, 2014.
[6] Davis, Allan Peter, Cynthia Grondin Murphy, Robin Johnson, Jean M. Lay, Kelley Lennon-Hopkins, Cynthia Saraceni-Richards, Daniela Sciaky, et al. "The Comparative Toxicogenomics Database: Update 2013." Nucleic Acids Research, 2012.
[7] Law, Vivian, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam Maciejewski, et al. "DrugBank 4.0: Shedding New Light on Drug Metabolism." Nucleic Acids Research 42, no. D1 (2014): D1091–97.
[8] Lipscomb, Carolyn E. "Medical Subject Headings (MeSH)." Bulletin of the Medical Library Association 88, no. 3 (2000): 265.
Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space." arXiv Preprint arXiv:1301.3781, 2013. http://arxiv.org/abs/1301.3781.
[9] Hewett, Micheal, Diane E. Oliver, Daniel L. Rubin, Katrina L. Easton, Jshua M. Stuart, Russ B. Altman, and Teri E. Klein. "PharmGKB: The Pharmacogenetics Knowledge Base." Nucleic Acids Research 30, no. 1 (January 1, 2002): 163–65.
[10] Bodenreider, Olivier. "The Unified Medical Language System (UMLS): Integrating Biomedical Terminology." Nucleic Acids Research 32, no. suppl 1 (2004): D267–70.
[11] Schriml L M, Arze C, Nadendla S, et al. Disease Ontology: a backbone for disease semantic integration[J]. Nucleic acids research, 2012, 40(D1): D940-D946.
[12] Pathak, Jyotishman, and Christopher G. Chute. "Analyzing Categorical Information in Two Publicly Available Drug Terminologies: RxNorm and NDF-RT." Journal of the American Medical Informatics Association 17, no. 4 (2010): 432–39.