Disease Named Entity Recognition and Normalization using Conditional Random Fields and Levenshtein Distance

Qikang Wei, Ruifeng Xu*, and Lin Gui

Shenzhen Engineering Laboratory of Performance Robots at Digital Stage,
Harbin Institute of Technology Shenzhen Graduate School,
Shenzhen, China
weiqikang@hotmail.com, xuruifeng@hitsz.edu.cn, guilin.nlp@gmail.com

Abstract. This presents a machine learning-based approach for disease named entity recognition and normalization (DNER) subtask of Chemical Disease Relation (CDR) task in BioCreative V. This approach employs a Conditional Random Fields (CRF) based model with domain specific features in biomedical area in disease named entity recognition. In order to improve the performance of entity normalization, the method based on Levenshtein distance is applied. Furthermore, some post-processing techniques were adopted. Experiments on the development set using provided training dataset shows that the proposed approach increases the performance of disease named entity recognition and normalization effectively. In the final evaluation, using the 1,000 PubMed abstracts released as training dataset, this approach achieves F1 performance of 0.7924 on mention level and 0.6050 on the concept level.

Key words: Disease named entity recognition, Conditional random fields, Levenshtein distance

1 Introduction

Extraction of chemical, disease entities and their relations has a potential to facilitate knowledge acquirement of drug discovery and safety surveillance. The Chemical Disease Relation (CDR) task[1] in BioCreative V aims to encourage the further development of techniques for recognizing chemical and disease entities and detecting the chemical-disease relations.

Many systems and techniques for disease entity recognition from texts were developed[2]. However, there are still challenges to high performance disease named entity recognition attributes to the fact that diversity in naming conventions and the lack of appropriate training corpora. Meanwhile, each disease name entity has many types of naming conventions, including common, trivial

* Corresponding Author

and abbreviated version etc. This affects the recognition and normalization of disease entities.

BioCreative V organized Chemical Disease Relation (CDR) task in which the disease named entity recognition (DNER) and normalization is an intermediate step before CDR extraction. It was found highly difficult in previous BioCreative CTD tasks[3]. In this subtask, the participator system is required to return recognized and normalized disease concept identifiers for the given PubMed[4] abstract.

We developed a pipeline machine learning based system for this sub-task. First, a Conditional Random Fields (CRF)[5] based model is developed with domain specific features in biomedical area for disease named entity recognition. Second, the method based on Levenshtein distance is applied to normalize the recognized disease named entity and align the named entity to concept. In the evaluation, using the 1,000 PubMed abstracts released as training dataset, this approach achieves F1 performance of 0.7924 on mention level and 0.6050 on the concept level which is higher than the baseline based on dictionary lookup.

2 Methods

2.1 Dataset

For the CDR task, sample (50 abstracts), training (500 abstracts) and development (500 abstracts) datasets have been released by the task organizers[6]. Each dataset contains PubMed abstracts and disease annotations with aligned concept ID from MeSH database.

2.2 Architecture

Our system consists of six basic components, which is showed in Figure 1: (1) a rule-based method which splits the PubMed abstracts to sentences; (2) the component determines the boundaries of multiword disease named entity using BIO notation, where B, I, and O denote the beginning, intermediate and outside tokens of an entity, respectively; (3) the feature generation component for extracting the features including chunking, part-of-speech and morphological feature. The complete feature set as shown in Table 1, for disease named entity recognition; (4) a machine learning method based on Conditional Random Fields for recognizing disease named entity; (5) post-processing component which a rule-based method to identify some missed name entities, especially the abbreviations; and (6) named entity normalization component which uses dictionary matching and Levenshtein distance[7] .

3 Discussion

We consider *(-n, n)* words as the context window of the observing word, where n is set from 1 to 3 based on the feature. As for the machine learning algorithm,

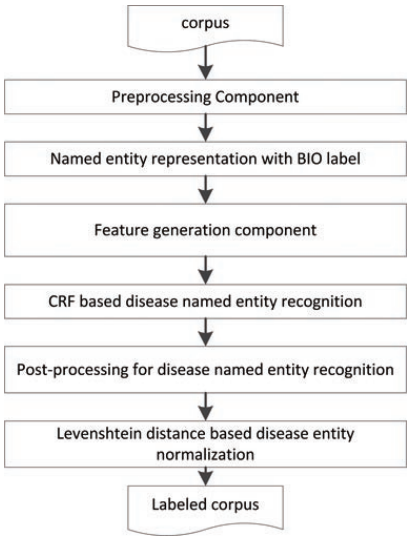

Fig. 1. System Architecture

Table 1. Features for Disease Named Entity Recognition

Feature	Description
Word	the word itself
POS	part-of-speech generated by GENIA tagger[8]
Chunk	chunk information generated by GENIA tagger
Word Shape[9]	match alphanumeric and punctuation patterns to simple representation
Orthographical	all upper, all digit, all symbolall upper digit etc.
Dictionary	dictionary look up against disease collection in MeSH and other dictionary

CRFsuite[10] is adopted to implement the CRFs model based disease named entity recognition. In the post-processing component, we tag all instances of a specific entity as a disease name mention if that entity was tagged by the CRF model at least twice within an abstract[11]. we also make effort on abbreviation resolution[12]. These post-processing techniques are expected to improve the recall of the system. In the normalization module, we firstly establish a mapping between MeSH disease entity and its ID while many kinds of morphological transformation are considered. After a simple match between entity and MeSH ID, we employ a Levenshtein distance method to rank the pairs of tagged entity and MeSH items. The most appropriate MeSH ID is assigned to the tagged entity in PubMed abstract.

4 Results

In this task, performance was reported as precision, recall and F-measure.

Table 2 gives the performance achieved by our system on the development set in DNER subtask.

Table 2. The performance of our system on development set after training on the training set for the DNER subtask.

Method	Concept-level			Mention-level		
	P	R	F	P	R	F
Our System	60.25%	64.13%	62.13%	84.33%	78.52%	81.32%

Table 3 gives the performance achieved by our system on the final evaluation in DNER subtask.

Table 3. The best performance of our system on test set after training on the training and development set for the DNER subtask.

Method	Concept-level			Mention-level		
	P	R	F	P	R	F
Our System	58.34%	62.83%	60.50%	82.45%	76.27%	79.24%
Baseline	42.71%	67.46%	52.30%	40.88%	59.95%	48.61%

The baseline offered by the organizer employs a dictionary look up method.

It is observed that our system achieves F1 performance of 0.7924 on mention level and 0.6050 on the concept level which is higher than the baseline based on dictionary look up. Meanwhile, the performance achieved on the final dataset is slightly lower than the performance achieved on the development dataset which shows the stability of our approach.

5 Conclusions

This paper presented a system for recognizing and normalizing the disease named entity. Firstly, a CRF-based method is developed for recognizing disease named entities. It employs rich features including orthographic, morphological and domain knowledge. Meanwhile, some post-processing techniques are developed to identify the missed entities. Secondly, dictionary look and Levenshtein distance based ranking are employed to improve the performance of disease named entity normalization. This system achieved F1 performance of 60.5% on concept-level and 79.24% on mention-level, respectively. Further improvements may include using more domain knowledge and optimized feature set, and employing better strategies such as information retrieval approach in normalization.

Acknowledgments. This work is supported by the National Natural Science Foundation of China No. 61370165, National 863 Program of China 2015AA015405, the Natural Science Foundation of Guangdong Province No. S2013010014475, Shenzhen Development and Reform Commission Grant No.[2014]1507 and Shenzhen Peacock Plan Research Grant KQCX20140521144507925.

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. Leaman, R., Islamaj Dogan, R., Lu, Z. (2013) DNorm: disease name normalization with pairwise learning to rank. Bioinformatics, 29, 2909-2917.
3. Davis, A.P., Grondin, C.J., Lennon-Hopkins, K., et al. (2014) The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. Nucleic Acids Res, 2014 Oct 17,gku935.
4. Lu, Z. (2010) PubMed and beyond: a survey of web tools for searching biomedical literature. Database (Oxford), vol. 2011, baq036.
5. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. ICML 2001, pp. 282C289 (2001)
6. Li J, Sun Y, Johnson RJ, Sciaky D, et al., (2015) Annotating chemicals, diseases, and their interactions in biomedical literature, in Proceedings of the fifth BioCreative challenge evaluation workshop, Sevilla, Spain
7. Levenshtein, Vladimir I. "Binary codes capable of correcting deletions, insertions, and reversals." Soviet physics doklady. Vol. 10. No. 8. 1966.
8. Tsuruoka, Yoshimasa, et al. "Developing a robust part-of-speech tagger for biomedical text." Advances in informatics (2005): 382-392.
9. B. Settles, Biomedical named entity recognition using conditional random fields and rich feature sets, in Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications (JNLPBA 04), pp. 104C107, 2004.
10. Okazaki, Naoaki. "CRFsuite: a fast implementation of conditional random fields (CRFs)." URL http://www. chokkan. org/software/crfsuite (2007).

11. Leaman, Robert, Chih-Hsuan Wei, and Zhiyong Lu. "tmChem: a high performance approach for chemical named entity recognition and normalization." Journal of cheminformatics 7.supplement 1 (2015).
12. Schwartz, Ariel S., and Marti A. Hearst. "A simple algorithm for identifying abbreviation definitions in biomedical text." Pacific Symposium on Biocomputing. Vol. 8. 2003.