Recognition and normalization of disease mentions in PubMed abstracts

Jitendra Jonnagaddala[1, 2], Nai-Wen Chang[3, 4], Toni Rose Jue[2], Hong-Jie Dai*[5]

[1]School of Public Health and Community Medicine, UNSW Australia
[2]Prince of Wales Clinical School, UNSW Australia
[3]Institution of Information Science, Academia Sinica, Taiwan
[4]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taiwan
[5]Department of Computer Science and Information Engineering, National Taitung University, Taiwan

{z3339253, t.jue}@unsw.edu.au
d00945020@ntu.edu.tw
hjdai@nttu.edu.tw

Abstract.
The rapidly increasing number of available PubMed documents calls the need for an automatic approach in the identification and normalization of disease mentions in order to increase the precision and effectivity of information retrieval. We herein describe our team's participation for the Disease Named Entity Recognition and Normalization subtask under the chemical-disease relations track of the BioCreative V shared task. We developed a CRF-based model using BIESO tagging format to allow automated recognition of disease entities in PubMed abstracts. Recognized disease entities were normalized to MeSH concepts using a dictionary look-up method based on Lucene. Performance is reported using precision, recall and F-measure on three separate runs. Our best run achieved F-measure of 80.74% on disease mention recognition and 67.85 % on disease normalization.

Keywords: Disease normalization; Disease recognition; Dictionary lookup; CRF; Disorder identification; Information extraction

1 Introduction

The importance of recognizing disease mentions and normalizing these to a standardized vocabulary is increasing with the yearly increase of published biomedical literature [1]. Keywords relating to diseases are the second most common user search query in PubMed, one of the most popularly used biomedical literature database [2]. Due to the doubling increase of biomedical literature available, researchers are now

* Corresponding Author

faced with the challenge of identifying relevant biomedical documents that would address their needs [3, 4]. Medical subject headings (MeSH) was developed by the National Library of Medicine (NLM) to speed up and increase the precision of information retrieval [5]. Where possible, documents in PubMed are indexed with relevant disease specific keywords using MeSH terminology[2]. Manually assigning disease specific MeSH terms to documents is a labor- and time-intensive process which would require monetary investment as well. Information extraction, text mining and information retrieval techniques can be employed on biomedical literature to assist in overcoming these challenges. Thus, automatic identification and normalization of disease mentions has become a crucial step in biomedical information extraction. It has various applications such as document indexing, document triage, biocuration, drug discovery, and drug safety surveillance [6]. Bringing an automated approach to this task has been the goal of previous studies but still calls for further improvements [7, 8]. The BioCreative[3] and BioNLP[4] shared tasks provide a unique opportunity for researchers in the biomedical text mining field to develop novel tools and methods to support this. The chemical-disease relations (CDR) track of the 2015 BioCreative V include two sub-tasks (i) Disease Named Entity Recognition and Normalization (DNER) and (ii) Chemical-induced diseases relation extraction (CID) [15]. As part of DNER sub-task, participants were provided with PubMed abstracts and was requested to return disease mentions normalized specifically to MeSH concepts. In this paper, we present our participation in CDR track. We present methods to automatically recognize disease mentions from PubMed abstracts using Conditional Random Fields (CRF) followed by normalization to MeSH concepts using Lucene[5] based dictionary look-up.

2 Methods

2.1 Dataset and Baseline System

The BioCreative V CDR Track organizers have provided participants with development and training sets to develop DNER systems. The development and training sets included 500 PubMed abstracts each from articles in the Comparative Toxicogenomics Database (CTD) [9]. Table 1 illustrates the distribution and characteristics of both training and development sets. The abstracts were manually annotated using PubTator tool [10]. Annotators used MeSH 2015 version to assign diseases with MeSH concepts [16]. The abstracts were processed following the basic workflow illustrated in Figure 1.

The organizers implemented a baseline system for benchmarking purposes. The baseline system used dictionary look-up method using CTD disease terms.

[2] https://www.nlm.nih.gov/mesh/
[3] http://www.biocreative.org/
[4] http://www.bionlp-st.org/
[5] https://lucene.apache.org/core/

	Training Set	Development Set
No. of documents	500	500
No. of disease mentions	4595	4601
No. of unique disease mentions	1464	1311
No. of unique MeSH concepts	665	605
No. of disease mentions without MeSH concepts (MeSH ID = -1)	32	16

Table 1. Overview of the BioCreative 4 CDR Track dataset

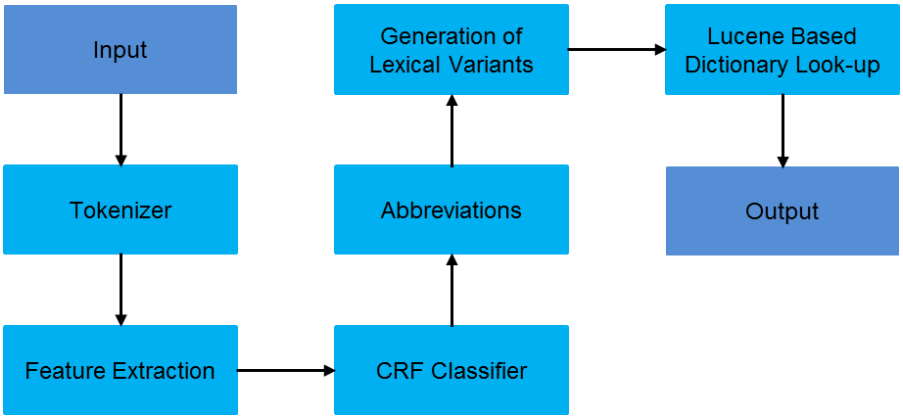

Fig. 1. High level components of the developed system

2.2 Disease mentions recognition

We developed a CRF-based model using BIESO tagging format to recognize disease entities in PubMed abstracts [11]. Abstracts were tokenized using Stanford's PTBTokenizer[6]. We used Stanford's CRF-NER package[7] to train our model. Training and development sets were merged together to extract over 50 different features and train the model to recognize disease mentions in a held out test set. The features were extracted by mainly using Stanford CoreNLP[8] and OpenNLP[9] packages. UMLS Metathesaurus was used to extract semantic features such as semantic groups and types.

[6] http://nlp.stanford.edu/software/tokenizer.shtml
[7] http://nlp.stanford.edu/software/CRF-NER.shtml
[8] http://nlp.stanford.edu/software/corenlp.shtml
[9] https://opennlp.apache.org/index.html

2.3 Disease normalization

CTD Medic disease vocabulary was used to map recognized disease mentions to MeSH concepts [12]. The CTD Medic disease vocabulary includes both MeSH and OMIM terminologies. However, for this study OMIM concepts have been excluded. Lucene was used for a dictionary look-up and for indexing MeSH disease concepts. A custom lexicon of abbreviations was developed from the training and development sets. If our CRF classifier identified any abbreviations from this lexicon, the abbreviations were then mapped to MeSH concepts in the custom lexicon. We have also generated lexical variants for the recognized disease entities. Initially, the dictionary look-up is configured to identify exact matches. In cases where no exact matches were found, the dictionary look-up tried to map partial matches to MeSH concepts using three different techniques. We have submitted these three different techniques as three different runs. Run1 uses cosine similarity where the lexical variant and concept pair with the highest cosine similarity score was selected. Run2 uses Lucene's phrase match and Run 3 uses Lucene's phrase and term match. In cases where there is more than one match in either run 2 or run 3, we considered the number of potential MeSH concepts retrieved and weights to choose the best candidate. In situations where the dictionary look-up failed to find a disease concept then '-1' was assigned automatically as a MeSH ID.

3 Results and Discussion

Abstracts from both the development and test set were processed using the system developed based on the CRF model [11]. The test set was not directly provided to the participants. Instead, the organizers requested the participants to provide access to the developed systems through web services. The organizers then processed the test set through the systems provided by the participants and supplied the participants with results. The breakdown of true positives (TP), false positives (FP) and false negatives (FN) were also provided and was calculated as shown in Figure 2.

$$P = TP \div (TP + FP)$$
$$R = TP \div (TP + FN)$$
$$F = (2 \times P \times R) \div (P + R)$$

Fig. 2. Standard metrics used for performance analysis. P – Precision, R – Recall and F – F-measure.

The baseline system implemented by organizers achieved 42.71% of P, 67.46% of R and 52.30% of F. Our system's best performing run is run 3. Run 3 achieved 66.73% of P, 69.01 % of R and 67.85 % of F. Table 2 illustrates the official results achieved for individual runs. Based on official results, our system ranked 9 and 12 out of 16 participating teams for disease named entity recognition (NER) and normalization, respectively. Additionally, our system's best performing run ranked 11 out of 40 runs from all the participating teams based on the average response time (msec). The system

had higher performance for DNER than DNorm based of the overall weighted average mean. The disease mention recognition achieved a higher precision compared to its recall. On the other hand, disease normalization achieved a higher recall than precision.

Due to lack of gold standard annotations, we are not able to discuss the system's performance on the test set in detail. However, during our system development stage, we noticed several issues with abbreviations. For example, our CRF based classifier repeatedly identified 'APC' (adenomatous polyposis coli) as a disease entity. This resulted to our dictionary look-up techniques assigning wrong concept codes. These kind of issues can be handled by employing dismbiguation techniques [7, 13]. The current dictionary look-up heavily relies on the common terms in indentified diseases mentions and MeSH concepts. Sophisticated machine learning techniques like learning to rank approaches should be employed to identify semantically related terms [7, 14].

Run	Average response time (msec)	Normalization (Concept-level)			NER (Mention-level)		
		P	R	F	P	R	F
1	2,602.3	64.08	67.76	65.87	84.58	77.24	80.74
2	1,069.0	66.43	66.40	66.42	84.58	77.10	80.67
3	1,552.4	66.73	69.01	67.85	84.58	77.24	80.74

Table 2. BioCreative V CDR Track official results

4 Conclusion

We developed a CRF-based model using the BIESO tagging format that successfully recognized disease entities in PubMed abstracts. The recognized disease entity were normalized to MeSH concepts. Overall, our system performed better compared to the baseline system. However, further improvements must be made to increase the system's performance, especially develop more effective methods to handle abbreviations and disambiguation. The generation of lexical variants assisted in normalizing entities to MeSH concepts. However, they also resulted in many false positives. In future, we would like to perform detailed error analysis on the test set to devise better methods. We also would like to train our CRF classifier by combining multiple corpuses.

5 Acknowledgment

The authors would like to thank the organizers of BioCreative V CDR Track. This study was conducted as part of the electronic Practice Based Research Network (ePBRN) and Translational Cancer research network (TCRN) research programs. ePBRN is funded in part by the School of Public Health & Community Medicine, Ingham Institute for Applied Medical Research, UNSW Medicine and South West Sydney Local Health District. TCRN is funded by Cancer Institute of New South Wales and Prince of Wales Clinical School, UNSW Medicine. The content of this publication is solely the

responsibility of the authors and does not necessarily reflect the official views of the funding bodies.

REFERENCES

1. Lu, Z., *PubMed and beyond: a survey of web tools for searching biomedical literature.* Database (Oxford), 2011. **2011**: p. baq036.
2. Islamaj Dogan, R., et al., *Understanding PubMed user search behavior through log analysis.* Database (Oxford), 2009. **2009**: p. bap018.
3. Rebholz-Schuhmann, D., A. Oellrich, and R. Hoehndorf, *Text-mining solutions for biomedical research: enabling integrative biology.* Nature Reviews Genetics, 2012. **13**(12): p. 829-839.
4. Zhu, F., et al., *Biomedical text mining and its applications in cancer research.* Journal of biomedical informatics, 2013. **46**(2): p. 200-211.
5. Lipscomb, C.E., *Medical Subject Headings (MeSH).* Bulletin of the Medical Library Association, 2000. **88**(3): p. 265-266.
6. Arighi, C.N., et al., *An overview of the BioCreative 2012 Workshop Track III: interactive text mining task.* Database, 2013. **2013**: p. bas056.
7. Leaman, R., R.I. Doğan, and Z. Lu, *DNorm: disease name normalization with pairwise learning to rank.* Bioinformatics, 2013: p. btt474.
8. Arighi, C., et al., *Proceedings of the fourth BioCreative challenge evaluation workshop.* 2013.
9. Davis, A.P., et al., *The comparative toxicogenomics database: update 2013.* Nucleic acids research, 2012: p. gks994.
10. Wei, C.-H., H.-Y. Kao, and Z. Lu, *PubTator: a web-based text mining tool for assisting biocuration.* Nucleic acids research, 2013: p. gkt441.
11. Lafferty, J.D., A. McCallum, and F.C.N. Pereira, *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*, in *Proceedings of the Eighteenth International Conference on Machine Learning*. 2001, Morgan Kaufmann Publishers Inc. p. 282-289.
12. Davis, A.P., et al., *MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database.* Database, 2012. **2012**: p. bar065.
13. Leaman, R., R. Khare, and Z. Lu, *Challenges in clinical natural language processing for automated disorder normalization.* Journal of biomedical informatics, 2015. **57**: p. 28-37.
14. Huang, M., J. Liu, and X. Zhu, *GeneTUKit: a software for document-level gene normalization.* Bioinformatics, 2011. **27**(7): p. 1032-1033.
15. 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
16. 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