An enhanced CRF-based system for disease name entity recognition and normalization on BioCreative V DNER Task

Hsin-Chun Lee [1], Yi-Yu Hsu [2], Hung-Yu Kao[3,*]

[1]Institute of Medical Informatics, National Cheng Kung University, Tainan, Taiwan, R.O.C. 701
[2,3] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan,Taiwan, , R.O.C. 701

[1] hclee@ikmlab.csie.ncku.edu.tw;
[2] alan.hsu@ikmlab.csie.ncku.edu.tw;
[3,*] hykao@mail.ncku.edu.tw

Abstract. Disease plays a central role in many areas of biomedical research and healthcare. However, the rapid growth of disease and treatment research creates barriers to the knowledge aggregation of PubMed database. Thus, a framework of disease mention recognition and normalization has become increasingly important for biomedical text mining. In this work, we utilize conditional random fields (CRFs) to develop a recognition system and optimize the results by customizing several post-processing steps, such as abbreviation resolution and consistency improvement. At the DNER subtask of BioCreative V CDR task, the system performance of disease normalization is 0.8646 of F-measure, especially a high precision (0.8963) on the normalization task.

Keywords. Disease name entity recognition and normalization, Conditional random fields, Biomedical text mining

1 Introduction

In the biomedical field, it has a rapid and exponential growth of producing large-scale biomedical literature [1]. To facilitate the integration of biomedical articles, building an automatic annotation system will effectively help biocurators to curate the relations between bioconcepts. For example, chemicals, diseases, and their relations play central roles in many areas of biomedical research and healthcare. Before extracting their relations, the system has to retrieve the mentions of bioconcepts from unstructured free texts and assign the mention a relative database identifier.

In past few years, the recognition and normalization of bioconcepts in biomedical literature have attracted attention, such as chemicals [2], diseases [3], genes [4-7], species [8, 9], and variations [10, 11], respectively. When biocurators investigate biomedical articles, disease mentions are very important in many lines of inquiry involving disease, including etiology (e.g. gene-variation-disease relationships) and clinical aspects (e.g. diagnosis, prevention and treatment) [12]. In the biomedical texts (e.g., literature and medical records), diseases recognition faces four major variation challenges. 1) Disease terminology: disease names naturally exhibit a complicate and inconsistent terminology problem. Such like 'cancer', 'carcinoma', and 'malignant tumour' share a similar meaning [3]. 2) Combination word: as principles of disease word formation, they are mostly composed of prefix, suffix and root. For instance, the word 'hyper-' represents overactive, therefore, 'hypertension' means high blood pressure. For another example, 'nephritis' and 'nephropathy' are the root word 'nephro' combined with '-itis' and '-pathy', which mean inflammation and disease. 3) Abbreviation: diseases abbreviation are frequently used in text (i.e. 'HD' presents 'Huntington disease') which may be ambiguous with other concepts (e.g., gene). 4) Composite disease mention: a coordination ellipsis which refers to two or more diseases. For example, 'ovarian and peritoneal cancer' indicates that two individual diseases are MeSH: D010051 (Ovarian Neoplasms) and MeSH: D010534 (Peritoneal Neoplasms), respectively.

Continuing the previous BioCreative IV CTD task, the BioCreative V CDR Track focuses on the two topics: Chemical-induced-disease relation (CID) and disease recognition (DNER). Note that the disease recognition is difficult and may affect the performance in downstream information extraction (e.g., relation extraction) according to the four variation challenges. In our participation of DNER subtask in the CDR track, we developed a machine learning-based disease recognition system to deal with the four major variation challenges. The details of our method are descripted in the second session. The experiments and results are shown in the third session. We also recommend the reader studying the overview paper [13] to have a complete description for CDR Task.

2 Methods

To handle the DNER task, we defined a semantic based recognition method which contains two individual modules. 1) Disease name recognition. We defined recognition model based on the linear chain conditional random fields (CRF) [14] with rich features and used three lexicons (NCBI Disease corpus [15], MEDIC [16], CDR Training corpus [17]) to generate our CRF dictionary features. Moreover, we also employed multiple post processing steps to further optimize the recognition results. 2) Disease name normalization. To normalize disease mentions to specific concepts in existed repository, we developed a dictionary-lookup method based on the collection of MEDIC, NCBI disease corpus and CDR task released corpus released on June 4, 2015. MEDIC is a well-known resource which uses MeSH and OMIM identifiers to organize disease concepts.

2.1 Disease Recognition

In this model, we leveraged the CRF++ toolkit (https://taku910.github.io/crfpp/) to train a disease named entity recognition model. The model also utilizes BIEO states (B: begin, I: insides, E: end and O: outside) and a second order template of CRF which assists our system to recognize a disease named entity. In the preprocessing stage, we break tokens not only at white spaces and punctuations but also between letters and digits. We also divide tokens between lower case letters followed by an uppercase letter. The mention recognition issue has been addressed for many years. To minimize the development effort, we adapted the feature extraction from three recent recognition tools (i.e., CoINNER [18], tmChem[2], tmVar [11]). The significant features we utilized in this model are described as below:

1. Morphology: We contain the general features including the original tokens, stemmed tokens (extracted by Snowball library), and its prefixes/suffixes (length 1 to 5).
2. Terminology: We defined three significant types, disease terminologies, body part and human ability, to determine whether each token matches the condition we set. The detailed conditions are described as follows:

Table 2. The groups of three disease terminologies

groups	Conditions
disease terminologies	impairment, nausea, vomiting, disease, cancer, toxicity, insufficiency, effusion, deficit, dysfunction, injury, pain, neurotoxicity, infect, syndrome, symptom, hyperplasia, retinoblastoma, defect, disorder, failure, hamartoma, hepatitis, disease, tumors, tumor, cancer, damage, illness, illnesses, abnormality, tumour, abnormalities, abortion, abortions
body part	pulmonary, neuronocular, orbital, breast, renal, hepatic, liver, hart, eye, pulmonary, ureter, bladder, pleural, pericardial, colorectal, head, neck, pancreaticobiliary, cardiac, leg, back, cardiovascular, gastrointestinal, myocardial, kidney, bile, intrahepatic, extrahepatic, memorygastric
human ability	visual, auditory, learning, opisthotonu, sensory, motor, memory, social, emotion

3. Part Of Speech: A series of binary features for each part of speech.
4. Vowel: We defined a frame to represent the token. Continued vowels change to "-". For example, the words "tumor" and "tumour" are turned into the same frame "t-m-r".
5. Dictionary-lookup: We used the CTD disease vocabulary (MEDIC) as features. Furthermore, we set the length parameter which is greater than 3 to avoid false positives (i.e., an abbreviation of other concepts).
6. Abbreviation: We also annotated abbreviations and full names which are detected by BIOADI [19].

2.2 Post-Processing for Disease Recognition

We employed a post-processing to improve the recognition results. The first step, we improved the consistency by tagging all instances of a disease mention if that mention was tagged by the CRF model at least 25% mentions within an abstract. For example, a disease term "toxicity" appears four times in an abstract, but it missed three instances. Then, the missed instances would be added in the recognition results. Next, we employed an abbreviation tool – BIOADI [19] to deal with the abbreviation challenge. We defined two rules in recognizing

the correct disease abbreviation pairs. Once the long form of the pair is recognized as a disease mention, all instances of the abbreviation pairs including both long and short forms are recognized as diseases. Besides, when the short form is recognized as a disease mention, the long form would be also recognized as a disease mention if the long form contains a disease terminology, for example, 'disease', 'cancer', 'syndrome', 'symptom', 'tumor', 'deficiency' and 'disorder'. The final step of our post-processing is stopwords filtering which can be found on MySQL website (https://dev.mysql.com/doc/refman/5.1/en/fulltext-stopwords.html).

2.3 Disease Normalization

To identify relevant MeSH identifiers for recognized disease mentions, we developed a dictionary-lookup approach. The disease name lexicon is collected from the MEDIC (Comparative Toxicogenomics Database), the NCBI disease corpus, and the CDR training/development sets [17] which are adapted from a subset of the BioCreative IV CTD training corpus. All the disease names and their synonyms are utilized for normalization. Note that all the punctuations and white spaces are removed.

Since the term variation is a very critical issue in the disease recognition, we especially extend the MEDIC lexicon. In our observation, 'disease' and 'failure' are highly variant in the texts. For instance, 'Infection', 'damage', 'abnormalities', 'disorder', 'impairment', 'loss', 'complication', 'injury', 'deficit', 'anomaly' and 'symptom' are regarded as the synonyms of 'disease'; likewise, 'deterioration', 'diminished', 'reduced', 'subnormal', 'dysfunction', 'degeneration', 'decrease', 'impairment', 'insufficiency', 'weakness', 'lesion' are regarded as the synonym of 'failure'. Therefore, we extend MEDIC lexicon to solve this issue. Further, all plurals of those synonyms are appended as well.

Extracted mentions in text and disease synonyms in lexicon are translated to lowercase. The disease identifier is assigned to the exact matched mention. The dictionary-lookup approach is following the priority as follow disease lexicons: CDR development sets > CDR training sets > MEDIC > NCBI disease corpus > MEDIC extension lexicon. Once a mention indicates to more than two or more identifiers, the identifier in higher priority lexicon would be admitted.

Besides, if an abbreviation is not in the lexicon, our system would assign the identifier which matched by the long form. Finally, we defined two heuristic mention modifications to handle the British and American English issue (e.g., turn 'ae' into 'e') and suffix issue (e.g., turn '-mia' into '-mic'). By this step, few mentions can match exactly if the original mentions cannot match correctly. Otherwise, our system will assign '-1' if the mentions cannot be matched in lexicons.

3 Experiment and Result

In our evaluation, we developed three individual runs which are trained by the DNER subtask training/development sets (totally 1000 abstracts) and evaluated on the Testing sets (500 abstracts). The evaluation result is reported in Table 3.

Table 3. The performance of disease normalization on the CDR Testing set

Run	Training set for CRF	Precision	Recall	F-score
1	Train	0.8942	0.8244	0.8579
2	Train+Dev	0.8963	0.8350	0.**8646**
3	Train+Dev+NCBI disease corpus	0.8832	0.8365	0.8592

4 Conclusion

Also, we found several features which can assist the system in developing robust CRF models for disease recognition. By using these features, our model presents a superior result in the recognition step. Besides, our defined post-processing could raise the F-score about 3%. In the normalization step, we present a dictionary-lookup approach which obtains 0.8646 of F-score. However, the recall of our methods is relatively lower. Besides, we have further improved the processing speed of our system after the official evaluation due date. It is currently able to process 500 abstracts within 50 secs. In our future work, we will focus on raising the performance of normalization by a robust machine learning approach.

Acknowledgement

Authors would like to thank Chieh Fang for the discussion of the system improvement and the installation of the RESTful API.

REFERENCES

1. Z. Lu, "PubMed and beyond: a survey of web tools for searching biomedical literature," Database, vol. 2011, p. baq036, 2011.
2. R. Leaman, C.-H. Wei, and Z. Lu, "tmChem: a high performance approach for chemical named entity recognition and normalization," Journal of cheminformatics, vol. 7, 2015.
3. R. Leaman, R. I. Doğan, and Z. Lu, "DNorm: disease name normalization with pairwise learning to rank," Bioinformatics, p. btt474, 2013.
4. J. Hakenberg, M. Gerner, M. Haeussler, I. Solt, C. Plake, M. Schroeder, G. Gonzalez, G. Nenadic, and C. M. Bergman, "The GNAT library for local and remote gene mention normalization," Bioinformatics, vol. 27, pp. 2769-2771, 2011.
5. C.-N. Hsu, Y.-M. Chang, C.-J. Kuo, Y.-S. Lin, H.-S. Huang, and I.-F. Chung, "Integrating high dimensional bi-directional parsing models for gene mention tagging," Bioinformatics, vol. 24, pp. i286-i294, 2008.
6. R. Leaman and G. Gonzalez, "BANNER: an executable survey of advances in biomedical named entity recognition," in Pacific Symposium on Biocomputing, 2008, pp. 652-663.
7. C.-H. Wei, H.-Y. Kao, and Z. Lu, "GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains," BioMed Research International, vol. 2015, 2015.
8. M. Gerner, G. Nenadic, and C. M. Bergman, "LINNAEUS: a species name identification system for biomedical literature," BMC bioinformatics, vol. 11, p. 85, 2010.
9. C.-H. Wei, H.-Y. Kao, and Z. Lu, "SR4GN: a species recognition software tool for gene normalization," Plos one, vol. 7, p. e38460, 2012.
10. J. G. Caporaso, W. A. Baumgartner, D. A. Randolph, K. B. Cohen, and L. Hunter, "MutationFinder: a high-performance system for extracting point mutation mentions from text," Bioinformatics, vol. 23, pp. 1862-1865, 2007.
11. C.-H. Wei, B. R. Harris, H.-Y. Kao, and Z. Lu, "tmVar: a text mining approach for extracting sequence variants in biomedical literature," Bioinformatics, p. btt156, 2013.
12. N. Limsopatham, C. Macdonald, and I. Ounis, "Inferring conceptual relationships to improve medical records search," in Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, 2013, pp. 1-8.
13. C.-H. Wei, Y. Peng, R. Leaman, A. P. Davis, C. J. Mattingly, J. Li, T. C. Wiegers, and Z. Lu, "Overview of the BioCreative V Chemical Disease Relation (CDR) Task," presented at the Proceedings of the fifth BioCreative challenge evaluation workshop, 2015.
14. J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

15. R. I. Doğan, R. Leaman, and Z. Lu, "NCBI disease corpus: a resource for disease name recognition and concept normalization," Journal of biomedical informatics, vol. 47, pp. 1-10, 2014.
16. A. P. Davis, T. C. Wiegers, M. C. Rosenstein, and C. J. Mattingly, "MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database," Database, vol. 2012, p. bar065, 2012.
17. J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu, "Annotating chemicals, diseases and their interactions in biomedical literature," presented at the Proceedings of the fifth BioCreative challenge evaluation workshop, 2015.
18. Y.-Y. Hsu and H.-Y. Kao, "Curatable Named-entity Recognition using Semantic Relations," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 12, pp. 785-792, 2014.
19. C.-J. Kuo, M. H. Ling, K.-T. Lin, and C.-N. Hsu, "BIOADI: a machine learning approach to identifying abbreviations and definitions in biological literature," BMC bioinformatics, vol. 10, p. S7, 2009.