UTH-CCB@BioCreative V CDR Task: Identifying Chemical-induced Disease Relations in Biomedical Text

Jun Xu, Yonghui Wu, Yaoyun Zhang, Jingqi Wang, Ruiling Liu,
Qiang Wei, and Hua Xu

SBMI, The University of Texas Health Science Center at Houston
7000 Fannin St, Houston, TX, USA 77030
{Jun.Xu,Yonghui.Wu,Yaoyun.Zhang,Jingqi.Wang,
Ruiling.Liu,Qiang.Wei,Hua.Xu}@uth.tmc.edu

Abstract. This paper describes the system developed by the UTH-CCB team from the University of Texas Health Science Center at Houston (UTHealth), for the 2015 BioCreative V shared tasks of Track 3 on extraction of chemical disease relation (CDR). We participated in both tasks: Task A for "Disease Named Entity Recognition and Normalization (DNER)" and Task B for "Chemical-induced Diseases Relation Extraction (CID)". For Task A, we developed a Conditional Random Fields based named entity recognition system and used a general Vector Space Model-based approach for entity normalization. To extract the chemical-induced disease relation, we combined two Support Vector Machines-based classifiers, which were trained on sentence- and document- level, respectively. Our system achieved a F1 score of 83.53 for Task A and 57.03 for Task B, demonstrating the effectiveness of machine learning-based approaches for automatic extraction of entities and their relations from biomedical literature.

Keywords. Named Entity Recognition, Relation Extraction, Chemical, Disease

1 Introduction

Understanding the relations between chemicals and diseases is crucial in various tasks such as developing new drugs and preventing adverse drug reactions. Biomedical researchers have studied a great amount of associations between chemicals and drugs and published their findings in the biomedical literature. However, there is no comprehensive database containing all the relations between chemicals and diseases. While manually extracting these relations from the biomedical literature is possible, such a procedure is often time-consuming and difficult to keep up-to-date. Text mining methods could automatically detect the chemical and disease concepts as well as their relations from the biomedical literature, and help in the curation of chemical-disease relation knowledgebase from literature.

The BioCreative V track 3 CDR task[1] aims to examine the current text mining methods on chemical-disease relation extraction from PubMed abstracts. This challenge consists of two specific tasks: (A) Disease Named Entity Recognition and Normalization (DNER); (B) Chemical-induced Diseases Relation Extraction (CID). The chemical named entity recognition and normalization are required but not evaluated. In this paper, we describe our approaches and results for both tasks.

2 Methods

2.1 Datasets

In this challenge, the organizers developed a chemical-disease relation corpus composed of 1,500 PubMed abstracts [2], which were divided into a training set (500 abstracts), a development set (500 abstracts) and a test set (500 abstracts). We developed our machine learning models using the training set and optimized the parameters according to the performance on the development set. In the final submissions, we combined the training and the development data to build the final models.

2.2 Task A – Disease Named Entity Recognition/Normalization

Task A consist of two subtasks: 1) recognize disease entities, and 2) encode the recognized entities to Medical Subject Headings concept identifiers (MeSH® IDs).

Disease recognition is a typical named entity recognition (NER) task. We developed a machine learning model based on Conditional Random Fields (CRFs) [3]. We used the CRFs implementation in CRFsuite package[1]. The features used by our system can be categorized into the following groups:

(1) Word Level Features: Bag-of-word, Part of Speech (PoS) tags, orthographic information, such as case patterns, char n-gram, prefixes and suffixes of words;

(2) Dictionary Lookup Features: We built a dictionary based semantic tagger by leveraging the vocabularies and corresponding semantic tags (e.g. disorder, problem, drug, etc.) from UMLS.

(3) Contextual Features: Bi- and tri- grams of tokens, including word, word stem, PoS and semantic tags extracted by our semantic tagger.

(4) Chemical/Disease Related Features: We adopted the features representing characteristics specific to chemicals from tmChem[4]. We also defined several binary features for diseases, including suffixes (e.g. "-algia", "-emia", etc.) and prefixes (e.g. "ab-", "hemo-", etc.).

(5) Distributed Word Representation Features: In this challenge, we explored the deep neural network based word embeddings. We developed a deep neural network [5] to train word embeddings from all PubMed abstracts published in 2013.

We adapted a previously developed Vector Space Model (VSM) based approach[6] to find the correct MeSH® ID for a given entity. We calculated the cosine similarity

[1] http://www.chokkan.org/software/crfsuite/

score between the target entity and all of the candidate terms and returned the top ranked MeSH® ID. The encoding system assigned '-1' if the top-ranked MeSH® ID was not a disease, or there were no retrieved candidates.

For chemical recognition and normalization, we used the same NER system and encoding approach as described above for disease recognition.

2.3 Task B – Chemical-induced Disease Relation Extraction

We treated the CID as a classification task and developed a sentence-level classifier (C_S) and a document-level classifier (C_D). We used the LIBSVM [7] module for SVMs implementation.

C_S aims to identify the relation of the CID pair located in the same sentence. We systematically extracted the following features to train the classifier C_S, including:

(1) Context words with position. The unigram and bigram of words before, between and after the target chemical and disease entities. Other entities were replaced with their entity type if they were found between the target entities.

(2) Knowledgebase features. We extracted all relations of the chemical and disease pair in the CTD[8], MEDI[9] and SIDER[10] database as features. Meanwhile, the MeSH® tree structures of the chemical and disease are also extracted as features.

(3) Others. We extracted both of the mentions and normalized values of the chemical and disease as features. For each document, we also extracted a set of core chemicals. Core chemicals are those chemicals, which either have the highest frequency or occurred in the title. Whether the target chemical was a core chemical or not was also extracted as a feature.

As the relations in the challenge corpus were annotated at the document level, we tried two different strategies to construct the training corpus for the sentence-level classifier. We extracted all sentences that had a candidate CID pair from the document. The first strategy automatically generated the sentence level annotation according to the document-level annotations. The CID pair was annotated as "TRUE" if this pair was in the document level annotations. Otherwise, the CID pair was annotated as "FALSE". We named this auto-labeled data set as CID-S_A. By manual review of the sentence-level candidate pairs, we obtained another data set that we designated CID-S_M. Here, we manually reviewed the candidate pairs and annotated the "TRUE/FALSE" label according to the sentence.

The classifier C_D utilized document level information to classify the relations between the chemical and disease. C_D used the feature set (2) and (3) from the classifier C_S. In addition, we extracted the number of sentences between the two entities and presence of trigger words in the context as features.

We applied C_D for all candidate CID pairs. The union set of the predictions of both classifiers was used as our system predictions. If the union set was empty, we added all candidate pairs in which the chemical was a core chemical into the empty set as a final submission.

2.4 Experiments and Evaluation

We combined the training and the development datasets to build the final models. Since each task allows for three submissions, we tried different strategies for the three runs, as shown in Figure 1. For Task A, in Run 1 and Run 2, we trained two separate CRFs models (one for disease and the other for chemical), while in Run 3, we trained a unified CRFs model for both the disease and chemical entities. Run 2 also used the BioCreative IV CHEMDNER corpus[11] as it improved the performance of chemical recognition in our experiments.

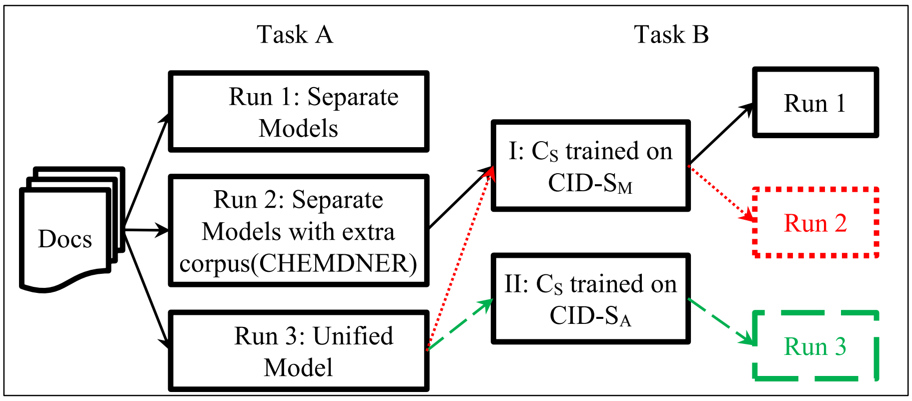
Task A
Task B
Docs
Run 1: Separate Models
Run 2: Separate Models with extra corpus(CHEMDNER)
Run 3: Unified Model
I: C_S trained on CID-S_M
II: C_S trained on CID-S_A
Run 1
Run 2
Run 3

Figure 1. The differences between the 3 runs for Task A and Task B

For Task B, Run 1 and Run 2 used the same approach where the C_S was trained on the data set CID-S_M. The C_S of Run 3 was trained on the CID-S_A.

The evaluation metrics for this task include F1 score, precision, and recall. Task A used a 2-tuple of document ID and disease concept ID as a data point while Task B used a 3-tuple containing document ID, chemical, and disease concept ID as a data point. For more details, please refer to the task description[1].

3 Results and Discussion

Table 1 shows the overall performance of our system in Task A as reported by the organizer, where 'P', 'R', 'F' denotes precision, recall, and F1 score, respectively. The best run for Task A were both Run 1 and Run 2, where we trained two independent CRFs models for the disease and chemical. The separate training CRFS models outperformed the unified CRFs model that combined the chemical and disease entities recognition.

Run	P	R	F
1	0.8312	0.8395	**0.8353**
2	0.8312	0.8395	**0.8353**
3	0.8254	0.8395	0.8324

Table 1. The performances of our system in Task A.

Table 2 shows the overall performance of our systems in Task B. Run 3 achieved the best F1-score. We are surprised to see that Run 3, which utilized automatically generated sentence level annotations outperformed Run 2, which utilized manually annotated sentence level annotations.

Run	P	R	F
1	0.5660	0.5591	0.5625
2	0.5665	0.5713	0.5689
3	0.5567	0.5844	**0.5703**

Table 2. The performances of the 3 runs of our system on Task B.

4 Conclusion

In this paper, we describe our participation in the 2015 BioCreative V CDR challenge. Our system participated all tasks. Our results show that it’s feasible to detect the chemical and disease entities from biomedical literature using machine learning methods. The domain specific features and distributed word representation features are useful for named entity recognition. However, the chemical disease relation extraction is still a challenging task.

5 Acknowledgment

This study is supported in part by grants from NLM 2R01LM010681-05, NIGMS 1R01GM103859 and 1R01GM102282, and CPRIT R1307. The first author (JX) is partially supported by NSFC 61203378.

REFERENCES

1. Wei, C.H., Peng, Y., Leaman, R., et al.: Overview of the BioCreative V Chemical Disease Relation (CDR) Task. In: the fifth BioCreative challenge evaluation workshop. (2015)
2. Li, J., Sun, Y., Johnson, R., et al.: Annotating chemicals, diseases, and their interactions in biomedical literature. In: the fifth BioCreative challenge evaluation workshop. (2015)
3. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282-289. Morgan Kaufmann Publishers Inc. (2001)
4. Leaman, R., Wei, C.H., Lu, Z.: tmChem: a high performance approach for chemical named entity recognition and normalization. Journal of cheminformatics 7, S3 (2015)

5. Collobert, R., Weston, J., Bottou, L., et al.: Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research 12, 2493-2537 (2011)
6. Zhang, Y., Wang, J., Tang, B., et al.: UTH_CCB: A Report for SemEval 2014 – Task 7 Analysis of Clinical Text. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 802-806. ACL, Dublin, Ireland (2014)
7. Chang, C.-C., Lin, C.-J.: LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 1-27 (2011)
8. Davis, A.P., Grondin, C.J., Lennon-Hopkins, K., et al.: The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. Nucleic acids research 43, D914-920 (2015)
9. Wei, W.Q., Cronin, R.M., Xu, H., et al.: Development and evaluation of an ensemble resource linking medications to their indications. Journal of the American Medical Informatics Association : JAMIA 20, 954-961 (2013)
10. Kuhn, M., Campillos, M., Letunic, I., et al.: A side effect resource to capture phenotypic effects of drugs. Molecular systems biology 6, 343 (2010)
11. Krallinger, M., Leitner, F., Rabal, O., et al.: CHEMDNER: The drugs and chemical names extraction challenge. Journal of cheminformatics 7, S1 (2015)