The UET-CAM System in the BioCreAtIvE V CDR Task

Hoang-Quynh Le[1], Mai-Vu Tran[1], Thanh Hai Dang[1] and Nigel Collier[2]

[1]University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam
[2]Department of Theoretical and Applied Linguistics, University of Cambridge, Cambridge, UK
{lhquynh,vutm,hai.dang}@vnu.edu.vn, nhc30@cam.ac.uk

Abstract. In this paper, we describe a system developed for the BioCreative V chemical-disease relation (CDR) task. The Disease Named Entity Recognition and Normalization (DNER) model employs joint learning using a perceptron-based named entity recognizer (NER) and a back-off model for named entity normalization (NEN). In order to maximize both precision and recall, our NEN adopts a sequential back-off ensemble approach based on Semantic Supervised Indexing (SSI) - a supervised Word Embedding (WE) method- giving results by inference from training data, and Skip-gram - an unsupervised WE method- taking advantage of large unlabeled data. In the Chemical-induced diseases relation extraction (CID) model, we firstly resolve co-references by using a multi-pass sieve to identify cross-sentence references for entities, thus enabling intra-sentence relations to be discovered more easily. Following this we extract CID relations using a support vector machine model trained on supervised sentence data from the CDR training and development dataset. We evaluated our method on both the DNER test set and the CID test set. Results show an F1 of 76.44 for the DNER task, and a best performance of 51.6 on the CID task using the multi-pass sieve.

Keywords. Named entity recognition, named entity normalization, relation extraction, joint inference, word embedding, semantic supervised indexing, skip-gram, support vector machine, perceptron, multi-pass sieves.

1 Introduction

Mining disease and chemical information from scientific texts is important to support an integrated understanding of chemical safety among patient groups and to facilitate hypothesis discovery for new pharmaceutical substances. As a consequence, extracting chemical-disease relations (CDR) from unstructured free text into structured knowledge has become an important sub-research field in biomedicine. To accelerate the progress of this important field of study, BioCreative V propose a challenge task for automatic extraction of CDRs [1, 2] with two sub-tasks: (A) Disease Named Entity Recognition and Normalization (DNER) and (B) Chemical-induced diseases relation extrac-

tion (CID). In the development phase, the organizer released a golden corpus containing training and development dataset, each of 500 Pubmed abstracts annotated with chemicals, diseases, their MeSH[1] IDs and their relations.

2 Materials and methods

As a team participating the challenge, we proposed a CDRs extraction system based on several state-of-the-art machine learning techniques. The overall architecture of the system is described in Figure 1. In which, pre-processing steps include sentence splitting, tokenization, abbreviation identification, stemming, POS tagging and dependency parsing (Stanford[2]).

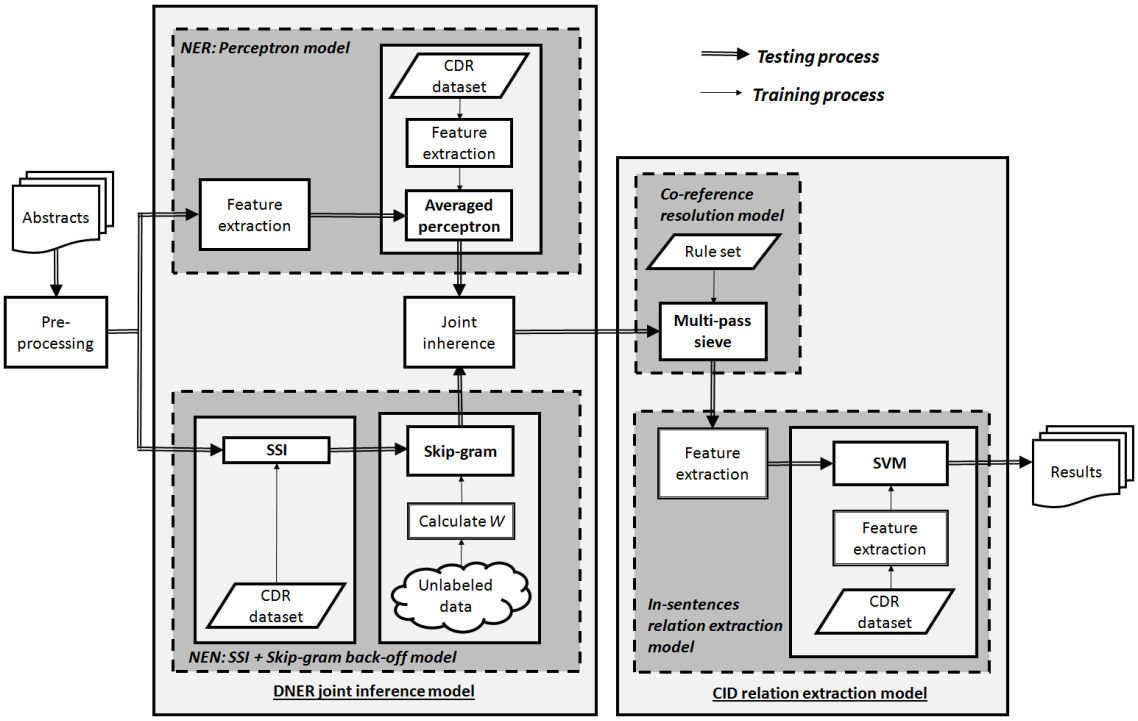

Fig. 1. Overall architecture of the proposed CDR extraction system, which includes the pipeline of processing modules and material resources used. Boxes with dotted lines indicate sub-models.

[1] Medical Subject Headings: www.ncbi.nlm.nih.gov/mesh
[2] Stanford Dependencies: http://nlp.stanford.edu/software/stanford-dependencies.shtml

2.1 Named Entity Recognition and Normalization

In our system, we built an NER model using an averaged perceptron method [3]. The perceptron is trained on the CDR data set with a standard lexicographic feature set: orthography features, context feature, POS tagging feature and dictionary (CTD[3]) features.

The NEN module is a sequential back-off model based on two word embedding (WE) methods: semantic supervised indexing (SSI) - a supervised WE methods, and skip-grams – an unsupervised WE methods. The SSI model is trained on the CDR dataset to obtain correlation matrix *W* between tokens in training data as well as MeSH. This approach was used by Leaman et al. [4] for disease NEN. Skip-gram [5] is a state-of-the-art word-to-vector method which takes advantage of large unlabeled data. Our system uses open source skip-gram model provided by NLPLab[4], which is trained on all PubMed abstracts and PMC full texts (4.08 million distinct words) with 200 dimensions. We used several techniques to convert skip-gram output into the correlation matrix form. In a sequential back-off manner, firstly, we implement the SSI model to find which pairs are linked, and then not-linked pairs are processed once again by the skip-gram model.

DNER is a joint-inference model to boost performance and reduce noise. Based on joint inference research using a modified beam search for decoding [6, 7], we trained two separate models for NER and NEN and then decode them simultaneously. We also propose a new scoring function for Beam search decoding as followed (formula 1).

$$argmax \sum_{i=1}^{n} (w_{NER}(x_{t=i}, y_{t=i-1;NER}) + w_{NEN}(x_{t=i}, x_{t=i-1}, y_{t=i-1;NER}, y_{t=i;NER})) \quad (1)$$

The scoring function for named entity normalization is:

$$w_{NEN}(x_{t=i}, x_{t=i-1}, y_{t=i-1;NER}, y_{t=i;NER}) = \begin{cases} 0, if\ y_{t=i;O} \\ w_{NEN}(x_{t=i}), if \begin{cases} y_{t=i-1;B-DS|I-DS|O}\ and\ y_{t=i;B-CD} \\ y_{t=i-1;B-CD|I-CD|O}\ and\ y_{t=i;B-DS} \end{cases} \\ w_{NEN}(x_{t=i}, x_{t=i-1}), if \begin{cases} y_{t=i-1;B-DS|I-DS}\ and\ y_{t=i;I-DS} \\ y_{t=i-1;B-CD|I-CD}\ and\ y_{t=i;I-CD} \end{cases} \end{cases} \quad (2)$$

if $w_{NEN} < w_{NEN}(NONE) = threshold$; re-write formula (1) to (3)

$$argmax \sum_{i=1}^{n} (w_{NER}(x_{t=i}, y_{t=i-1;NER}) + w_{NEN}(NONE)) \quad (3)$$

In which, W_{NER} is returned from the averaged perceptron model.

[3] Comparative Toxicogenomics Database: http://ctdbase.org

[4] http://evexdb.org/pmresources/vec-space-models/wikipedia-pubmed-and-PMC-w2v.bin

2.2 Chemical-induced-disease Relation Extraction

CID relation extraction is based on a pipeline model of a co-reference resolution module and an intra-sentence relation extraction module. The co-reference resolution module helps us to find more mentions of diseases and chemicals by improving on the multi-pass sieve method proposed by Souza and Ng (2015) [8].

Table 1. CID relation extraction feature set used in the proposed model

No	Type	Feature
1	Token features	Character types Character n-grams (n=1-4) Base form of the word Part-of-speech
2	Neighboring word features	Features extracted by the token feature function for each word Word and dependency n-grams (n=2-4) Word n-grams (n=2; 3) Dependency n-grams (n=2)
3	Word n-gram features	Word n-grams (n=1-4) within a window of three words before or three words after the target word
4	Pair n-gram features	Word n-grams (n=1-4) within a window of three words before the first word in the target pair and three words after the last word.
5	Shortest path features	Shortest dependency paths between a word pair

After that, we create a set of pairs (disease_mention, chemical_mention) appearing within a sentence. And then the intra-sentence relation extraction module classifies which pair has CID relations. Our intra-sentence relation extraction module is a supervision binary support vector machine classifier (L2- regularized L1-loss). The Liblinear tool[5] is then used to train it on the CDR dataset with the rich feature set proposed in [9] (listed in table 1).

3 Results and discussion

The official results of our system on the test dataset (500 documents) are shown on Table 2. We compare our results with the benchmark results of BioCreative organizer obtained using CTD names look-up method for DNER and co-occurrence method for CID. Comparative evaluation between CID runs shows that a SVM approach for CID by

[5] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

itself (CID run 1) achieved 47.47% F1, whilst the multi-pass sieve co-reference (CID run 2) boosted performance by 4.13% F1.

Table 2. CDR official results of UET-CAM system on the test dataset of 500 documents. Time is response time (msec). BM: Benchmarking results.

Task	Run	Time	Precision (%)	Recall (%)	F1 (%)
DNER	BM	N/A	42.71	67.46	52.30
	1	276.0	73.20	79.98	76.44
CID	BM	N/A	16.43	76.45	27.05
	1	8,906.0	44.73	50.56	47.47
	2	8,993.2	53.41	49.91	51.60

We also make a comparison between our multi-pass sieve method and the Expectation Maximization (EM) clustering method of Ng (2008) [10] for co-reference resolution. In this regard, systems are trained on the CDR training dataset and tested on the CDR development set. The results demonstrates the strength of the our multi-pass sieves method, achieved 63.46% in Precision (7.09% better than EM clustering-based), 73.62 % in Recall (0.99% better than EM clustering-based) and 68.16% in F1 (4.69% better than EM clustering-based).

The DNER back-off model can take advantage of both labeled CDR dataset and extremely large unlabeled data. SSI calculates the correlations matrix between tokens, it works better than Skip-gram in case that token appeared in training data or MeSH (e.g. SSI links '*arrhythmias*' to MeSH:D001145, '*peripheral neurotoxicity*' to MeSH:D010523). The skip-gram model calculates similarity between tokens by taking advantage of large unlabeled data, and helps improve recall (e.g. Skip-gram link '*disordered gastrointestinal motility*' to MeSH:D005767, '*hyperplastic marrow*' to MeSH:D001855, which are false negative of SSI).

Traditionally, NER and NEN is processed as two separated tasks, in which, NEN takes the output of NER as its input. One limitation of this pipeline approach is that errors propagate from NER to NEN and there is no feedback from NEN to NER [11]. As demonstrated by Khalid et al. [12], most NEN errors are caused by recognition errors. Joint inference is expected to overcome the disadvantage of a traditional pipeline model. When trained on the CDR training data and tested on the CDR development data, joint inference has an F1 of 82.34%, significantly better than that of the pipeline model (79.26%). Joint inference outperforms the pipeline model in cases of long entities that belongs to

MeSH, such as *"combined oral contraceptives"* and "*angiotensin-converting enzyme inhibitors*.

4 Acknowledgment

NC gratefully acknowledges funding support from the EPSRC (grant number EP/M005089/1).

REFERENCES

1. Wei CH, Peng Y, Leaman R, et al. "Overview of the BioCreative V Chemical Disease Relation (CDR) Task". Proceedings of the fifth BioCreative challenge evaluation workshop, Sevilla, Spain, 2015.
2. Li J, Sun Y, Johnson R, et al. "Annotating chemicals, diseases, and their interactions in biomedical literature". Proceedings of the fifth BioCreative challenge evaluation workshop, Sevilla, Spain, 2015.
3. Huang, Liang, Suphan Fayong, and Yang Guo. "Structured perceptron with inexact search." Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2012.
4. Leaman, Robert, Rezarta Islamaj Doğan, and Zhiyong Lu. "DNorm: disease name normalization with pairwise learning to rank." Bioinformatics (2013): btt474.
5. Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.
6. Li, Qi, and Heng Ji. "Incremental joint extraction of entity mentions and relations." Proceedings of the Association for Computational Linguistics. 2014.
7. Miwa, Makoto, and Yutaka Sasaki. "Modeling joint entity and relation extraction with table representation." Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. EMNLP. Stroudsburg, PA, USA: Association for Computational Linguistics. 2014.
8. D'Souza, Jennifer, and Vincent Ng. "Sieve-Based Entity Linking for the Biomedical Domain." Volume 2: Short Papers: 297. Proceedings of ACL-IJCNLP Volume 2: Short Papers, 297, 2015.
9. Miwa, Makoto, et al. "Event extraction with complex event classification using rich features." Journal of bioinformatics and computational biology 8.01 (2010): 131-146.
10. Ng, Vincent. "Unsupervised models for coreference resolution." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008.
11. Liu, Xiaohua, et al. "Joint inference of named entity recognition and normalization for tweets." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012.
12. Khalid, Mahboob Alam, Valentin Jijkoun, and Maarten De Rijke. "The impact of named entity normalization on information retrieval for question answering." Advances in Information Retrieval. Springer Berlin Heidelberg, 2008. 705-710.