# Sieve-based Coreference Resolution Enhances Semi-supervised Learning Model for Chemical-induced Disease Relations Extraction

Hoang-Quynh Le[1], Mai-Vu Tran[1], Thanh Hai Dang[1], Quang-Thuy Ha[1] and Nigel Collier[2]
[1]University of Engineering and Technology – VNUH, Hanoi, Vietnam
[2]University of Cambridge, Cambridge, United of Kingdom

## 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.

•BioCreative V proposed a challenge task for automatic extraction of CDRs with two sub-tasks:

- Disease Named Entity Recognition and Normalization (DNER)
- Chemical-induced diseases relation extraction (CID).

•The BioCreative V CDR corpus includes 1,000 annotated Pubmed abstracts for training and 500 for testing.

•The SilverCID corpus is built based on CTD database, contains 38,332 sentences, 1.25 millions tokens, 48,856 chemical entities, 44,744 disease entities and 48,199 CID relations.

## METHODS

**DNER: Named entity recognition and normalization module**

The ***Named entity recognition (NER)*** module uses structured perceptron method.

- •Trained on CDR dataset and silver CID corpus
- •Use standard lexicographic feature set: orthography features, context feature, POS tagging feature and dictionary (CTD) features.

The ***Named entity normalization*** (NEN) module is a sequential back-off model base on two word embedding (WE) methods, in which the second model receives negative results of first model as its input.

- •Semantic supervised indexing (SSI) - a supervised WE methods trained on CDR data to obtain correlation matrix $W$ between tokens in training data and MeSH.
- •Skip-gram – an unsupervised WE methods trained on large unlabeled data. Several techniques are used to convert skip-gram output into the correlation matrix form.

The ***DNER Joint-inference model*** boost performance and reduce noise [1]:

- •NER and NEN models are trained separately but decode simultaneously.
- •Propose a new scoring function for Beam search decoding:

$$argmax \sum_{i=1}^{n} \begin{pmatrix} w_{NER}\left(x_{t=i}, y_{t=i-1;NER}\right) + \\ w_{NEN}\left(x_{t=i}, x_{t=i-1}, y_{t=i-1;NER}, y_{t=i;NER}\right) \end{pmatrix}$$

## REFERENCES

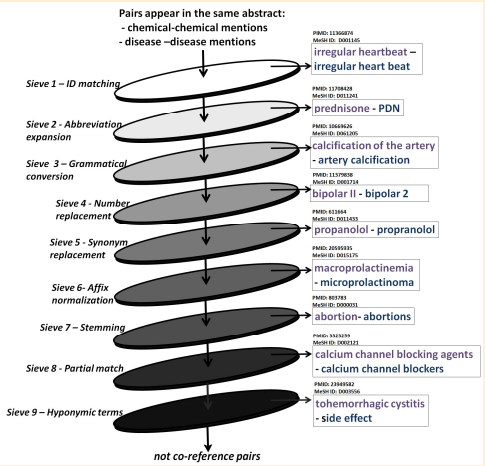
1. Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 402–412, Baltimore, Maryland, June.
2. D'Souza, J., & Ng, V (2015). Sieve-Based Entity Linking for the Biomedical Domain. In Proceedings of ACL-IJCNLP Volume 2: Short Papers, 297.
3. Miwa, M., Sætre, R., Kim, J. D., & Tsujii, J. I. (2010). Event extraction with complex event classification using rich features. Journal of bioinformatics and computational biology, 8(01), 131-146.

## METHODS (Cont.)

**CID relation extraction model**

CID relation extraction is based on a pipeline model of a co-reference resolution module and an intra-sentence relation extraction module.

- Co-reference resolution module find more mentions of chemicals and diseases in text. It is an improvement of multi-pass sieves method proposed by Souza and Vincent Ng (2015) [2] → Create a set of pairs (disease_mention, chemical_mention) appearing within a sentence.

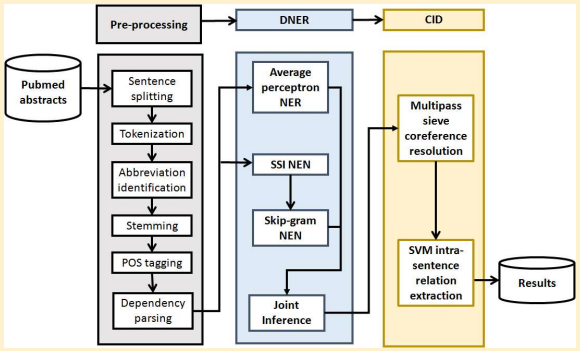



*Coreference resolution using nine-pass sieves*

| No | Feature type | Feature |
|---|---|---|
| 1 | Token features | Character types<br>Character n-grams (n=1-4)<br>Base form of the word<br>Part-of-speech |
| 2 | Neighboring word features | Features extracted by the token feature function for each word<br>Word and dependency n-grams (n=2-4)<br>Word n-grams (n=2; 3)<br>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 |

*Rich feature set of SVM model*

- Intra-sentence relation extraction module is a binary support vector machine classifier (L2- regularized L1-loss) - decides which pair has CID relations. SVM model is a trained on our silverCTD corpus set and CDR data set, uses a rich feature set proposed in [3]:




*Overall architecture of the proposed UET_CAM CID relation extraction system*

## RESULTS

| Task | Run | Precision | Recall | F-value |
|---|---|---|---|---|
| DNER | BM | 42.71 | 67.46 | 52.30 |
| | 1 | 79.90 | 85.16 | 82.44 |
| CID | BM | 16.43 | 76.45 | 27.05 |
| | 1 | 44.73 | 50.56 | 47.47 |
| | 2 | 53.41 | 49.91 | 51.60 |
| | 3 | 57.63 | 60.23 | 58.90 |

*Experimental result. BM: Benchmarking result.*

•Benchmark results of BioCreative organizer obtained using CTD names look-up method for DNER and co-occurrence method for CID.

•CID-1: Use SVM intra-sentence relation extraction model trained on CDR data

•CID-2: Use pipeline model of co-reference resolution and SVM intra-sentence relation extraction model trained on CDR data.

•CID-3: Add the silverCID corpus to train SVM model in CID-2

## DISCUSSION

•Experimental results demonstrated the strength of our proposed method compared to organizer's baseline methods.

• The multi-pass sieve co-reference (CID run 2) boosted performance by 4.13% F1.

•The SilverCID corpus (CID run 3) boosted performance by 7.3% F1.

•In a comparison between our multi-pass sieve method and the Expectation Maximization (EM) clustering method of Ng (2008): System using multi-pass sieves achieves 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) (trained on the CDR training dataset and tested on the CDR development set).

•The DNER back-off model can take advantage of both labeled CDR dataset and extremely large unlabeled data:

- The SSI model calculates the correlations matrix between tokens based on training data (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 (e.g. Skip-gram link '*disordered gastrointestinal motility*' to MeSH:D005767, '*hyperplastic marrow*' to MeSH:D001855).

•Joint inference is empirically demonstrated it power over traditional pipeline models in tackling errors propagation from NER to NEN and no feedback from NEN to NER.

- F-measure of Joint Inference model (82.44%) is better than of pipeline model (79.26%) (trained on CDR training data and tested on CDR development data).
- Joint Inference outperforms pipeline model in cases of long entities that belongs to MeSH, such as "*combined oral contraceptives*" and "*angiotensin-converting enzyme inhibitors*".

***Acknowledgment: EPSRC (grant number EP/M005089/1).***