Sieve-based coreference resolution enhances semi-supervised learning model for chemical-induced disease relations extraction

Journal:	*DATABASE*
Manuscript ID	DATABASE-2015-0217
Manuscript Type:	Original Article
Date Submitted by the Author:	04-Dec-2015
Complete List of Authors:	Le, Hoang-Quynh; University of Engineering and Technology, Vietnam National University, Information Technology Tran, Mai-Vu; University of Engineering and Technology, Vietnam National University, Information Technology Dang, Thanh Hai; University of Engineering and Technology, Vietnam National University, Information Technology Ha, Quang-Thuy; University of Engineering and Technology, Vietnam National University Collier, Nigel; University of Cambridge
Keywords:	Relation extraction, Chemical-induced Disease relation, Named entity recognition, Named entity normalization

Original Article

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, Vietnam National University, Hanoi, Vietnam

[2]Department of Theoretical and Applied Linguistics, University of Cambridge, Cambridge, UK

*Corresponding author: Tel: +84 (0)945885500; Email: lhquynh@vnu.edu.vn

Abstract

The BioCreative V chemical-disease relation (CDR) track was proposed to accelerate progress of text mining in facilitating integrative understanding of chemical substances, diseases and their relations. In this article, we describe an extension of the UET-CAM system for mining chemical-disease relations from text data, of which performance was ranked 4th among 18 participating corresponding systems by the BioCreative CDR track committee. In Disease Named Entity Recognition and Normalization (DNER) phase, our system employs joint learning with a perceptron-based named entity recognizer (NER) and a back-off model with Semantic Supervised Indexing (SSI) and Skip-gram for named entity normalization (NEN). Crucially, for solving the chemical-induced disease (CID) sub-task, we propose a pipeline that includes a coreference resolution module and a SVM intra-sentence relations extraction model. The former module utilizes a multi-pass sieve to identify inter-sentence references for entities while the latter is trained on both the CDR data and our silverCID corpus with a rich feature set. SilverCID is the silver standard corpus contains more than 50 thousands sentences which are automatically built based on the CTD database in order to provide evidence for the CID relation extraction. We critically evaluated our method on the CDR test set in order to clarify the contribution of our system components. Results show an F1 of 82.44 for the DNER task, and a best performance of F1 58.90 on the CID task. The comparisons also demonstrate the significant contribution of the multi-pass sieve coreference resolution method and the silverCID corpus.

Availability: SilverCID- The silver standard corpus for Chemical-induced Diseases relation extraction is available from: https://zenodo.org/record/34530 (doi:10.5281/zenodo.34530)

1 Introduction

A survey of PubMed user search behavior (1) found that diseases and chemicals were two of the most frequently requested entities by PubMed users worldwide: diseases appear in 20% of queries and and chemicals in 11%. These two entities are central to several topics such as developing drugs for therapeutics, discovering adverse drug reactions (ADRs) as well as chemical safety/toxicity among patient groups and facilitating hypothesis discovery for new pharmaceutical substances. As a consequence, extracting the chemical-disease relations from unstructured free text into structured knowledge has become an important field in biomedical text mining.

Compared with other biomedical entities such as genes and proteins, comparatively less research has been done on capturing disease and chemical entities and their relations. In recent years, capturing the critical significance of diseases and chemicals as well as drug-side-effect relations, has led to an expansion in this field. The Comparative Toxicogenomics Database (CTD) (2) is a manually curated database that promotes understanding about the effects of environmental chemicals (e.g., arsenic, heavy metals and dioxins) on human health. The CTD database had 1,842,746 chemical–disease associations as of June, 2015. Due to the high cost of manual curation and the rapid growth of the biomedical literature, several researches have attempted to extract chemical – disease relations or drug side effects automatically. The simplest approach is based on the co-occurrence statistics of chemical and disease entities (3), i.e. if two entities are mentioned together in the same sentence or abstract, they are probably related. This approach achieves high recall, but low precision and fails to distinguish the CID relation from other relations that commonly occur between chemical and disease. Rule-based techniques such as pattern-based approaches are also used for ADR extraction (4). This approach demands a large accumulation of rules, which caused by the large number of contexts and ambiguities. Such therefore often lead rule-based systems to having low recalls. Kang et al. (5) developed a knowledge-based relation extraction system that requires minimal training data, and applied the system for the extraction of ADRs from biomedical text. Other approaches are based on sophisticatedly advanced machine learning techniques, such as (6, 7) but still have gotten limited successes. The most important factor caused these limitations is the lack of a comprehensive dataset for training, moreover the abundant expression of ADRs in context as well as inter-sentence expression also make this problem more difficult. This research field, therefore, is still potential and challenged.

To accelerate progress, BioCreative V proposed a challenge task for automatic extraction of Chemical-induced disease (CDR) (8, 9) that has two sub-tasks:

(A) Disease Named Entity Recognition (DNER). This task includes automatic recognition of disease mentions (named entity recognition, NER) in PubMed abstracts and assignment of Medical Subject Heading (MeSH) (10) identifiers to these mentions (named entity normalization, NEN). They are initial steps for automatic CDR extraction.

(B) Chemical-induced disease relation extraction (CID). Participating systems were provided with raw text from PubMed articles as input and asked to return a list of <*chemical, disease*> pairs with normalized concept identifiers for which drug-induced diseases are associated in the abstract.

In these challenge tasks, disease are annotated following the '*Diseases*' [C] branch of MeSH 2015, including disease, disorder, signs and symptoms; chemical terminologies are annotated following the '*Drugs and Chemicals*' [D] branch of MeSH 2015; the CID relation refers to relationship between a chemical and a disease which are marked as 'marker/mechanism' in the CTD database. There are two types of such relationships (i) biomarker relations between a chemical and disease indicating that the chemical correlates with the disease and (ii) putative mechanistic relationships between a chemical and disease indicating that the chemical may play a role in the etiology of the disease (see figure 1 for examples).

As a team participating in the challenge for the first time, we proposed a modular system which solved the DNER and CID tasks in separated. For the DNER phrase, we proposed a reasonable manner for combining several state-of-the-art word-embedding techniques in NEN module in order to take advantages of both golden standard annotated corpus and the large scale unlabeled data. The NEN and NER modules are combined in a joint inference model to boost performance and reduce noise. The CID sub-task has to face many challenges such as (i) complex grammatical structures, (ii) the participation of entities in a CID relation occurring in both single sentences or spanning multiple sentences, additionally, (iii) the problem of expressing entities in MeSH IDs instead of mentions also make this problem more complex. To overcome these challenges, a traditional machine learning model for relation extraction – based only on explicit mentions of entities in a single sentence - is definitely not adequate. Using a SVM intra-sentences relation extraction module as a central core, we use a coreference module to find more disease/chemical mention in text then help to convert inter-sentence relations to intra-sentence relations. We also believe that the bigger training set is, the more useful information we can obtain, thus, we build a silver-standard annotation corpus (namely '*silverCID*' corpus) based on a carefully selected sub-set of the CTD database that does not occur in the testing set. This corpus is used for training the SVM model. In additional, we explore a rich feature set, which used successfully for event extraction, to adapt with complexities of CID relations appearances.

The novel contributions of this paper are as follows: (i) we proposed an ensemble model for structured-perceptron NER model, SSI and skip-gram NEN model in a rational manner of DNER joint inference model, (ii) we automatically build the SilverCID corpus- a sentence-level corpus to serve CID relation extraction as well as evaluate its influence on the system, (iii) we present evidence for the efficacy of using the multi-pass sieve in a biomedical relation extraction task and (iv) we demonstrate the strength of the rich feature set for CID relation extraction.

2 Materials and methods

2.1 Data set

Our experiments were conducted on the BioCreative V CDR data. In order to take advantage of the CTD database, we also built the SilverCID corpus –based on PubMed articles which are cited in the CTD database but not appearing in the BioCreative CDR track data set.

2.1.1 BioCreative CDR track data set

To assist CDR system development and assessment, the BioCreative V workshop organizers created an annotated text corpus that consists of human annotations for all chemicals, diseases and their chemical-induced disease relations. This corpus contains a total of 1,500 PubMed articles which are separated into three sub-sets each of: 500 for the training, development and test set (the details are shown on Table 1). Following the data survey of BioCreative (9), of these 1,500 articles, 1,400 were selected from an existing CTD-Pfizer collaboration related dataset which was generated via a previous collaboration curation between CTD and Pfizer (11), the remaining 100 articles contain newly curated data and are incorporated into the test set.

Table 1. Summary of the CDR track dataset

Data set	Articles	Chemical		Disease		CID
		Men	ID	Men	ID	
Training	500	5,203	1,467	4,182	1,965	1,038
Development	500	5,347	1,507	4,244	1,865	1,012
Test	500	5,385	1,435	4,424	1,988	1,066

Men: Mention, CID: CID relations

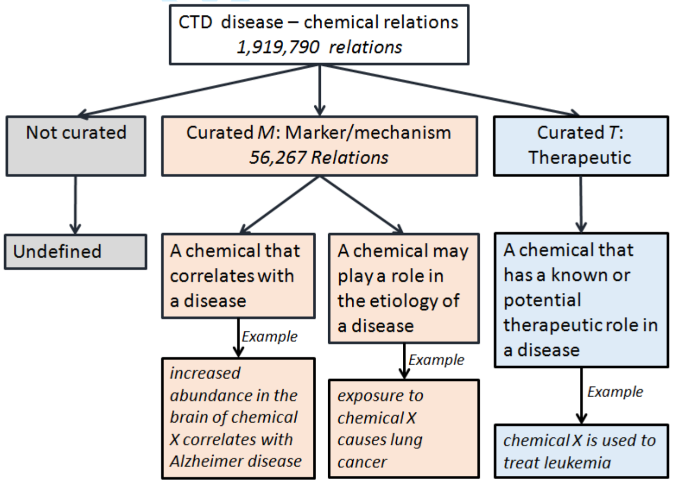
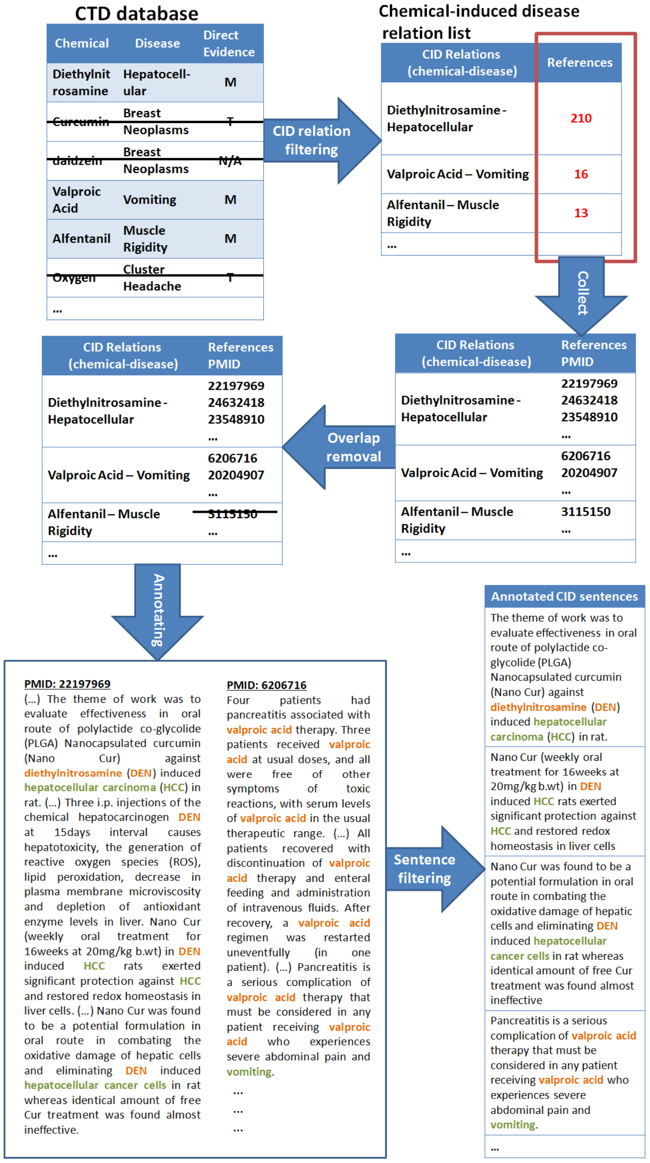

CTD disease – chemical relations
1,919,790 relations
Not curated
Curated M: Marker/mechanism
56,267 Relations
Curated T: Therapeutic
Undefined
A chemical that correlates with a disease
A chemical may play a role in the etiology of a disease
A chemical that has a known or potential therapeutic role in a disease
Example
Example
Example
increased abundance in the brain of chemical X correlates with Alzheimer disease
exposure to chemical X causes lung cancer
chemical X is used to treat leukemia

Fig. 1. Analysis of the Direct Evidence field in the CTD database

2.1.2 SilverCID corpus

The Comparative Toxicogenomics Database (2) (CTD) is a robust, publicly available database that aims to advance understanding about how environmental exposures affect human health. Chemicals in CTD come from the chemical subset of MeSH. CTD's disease vocabulary is a modified subset of descriptors from the "Diseases" category of MeSH, combined with genetic disorders from the Online Mendelian Inheritance in Man (OMIM) (12) database.

In more than 28 million CTD toxicogenomic relationships, there are 1,919,790 disease-chemical relations (curated or inferred via CTD-curated chemical–gene interaction) (October, 2015). There are several types of relations between diseases and chemicals, which may be described within the *Direct Evidence* field of the CTD database. This field has two labels *M* and *T*, in which the label *M* indicates relations seem to be very similar to the chemical–induced disease relations we are interested in (figure 1).



CTD database

Chemical	Disease	Direct Evidence
Diethylnitrosamine	Hepatocellular	M
~~Curcumin~~	~~Breast Neoplasms~~	~~T~~
~~daidzein~~	~~Breast Neoplasms~~	~~N/A~~
Valproic Acid	Vomiting	M
Alfentanil	Muscle Rigidity	M
~~Oxygen~~	~~Cluster Headache~~	~~T~~
...		

Chemical-induced disease relation list

CID Relations (chemical-disease)	References
Diethylnitrosamine - Hepatocellular	210
Valproic Acid – Vomiting	16
Alfentanil – Muscle Rigidity	13
...	

CID Relations (chemical-disease)	References PMID
Diethylnitrosamine - Hepatocellular	22197969 24632418 23548910 ...
Valproic Acid – Vomiting	6206716 20204907 ...
Alfentanil – Muscle Rigidity	3115150 ...
...	

CID Relations (chemical-disease)	References PMID
Diethylnitrosamine - Hepatocellular	22197969 24632418 23548910 ...
Valproic Acid – Vomiting	6206716 20204907 ...
Alfentanil – Muscle Rigidity	~~3115150~~ ...
...	

PMID: 22197969

(...) The theme of work was to evaluate effectiveness in oral route of polylactide co-glycolide (PLGA) Nanocapsulated curcumin (Nano Cur) against diethylnitrosamine (DEN) induced hepatocellular carcinoma (HCC) in rat. (...) Three i.p. injections of the chemical hepatocarcinogen DEN at 15days interval causes hepatotoxicity, the generation of reactive oxygen species (ROS), lipid peroxidation, decrease in plasma membrane microviscosity and depletion of antioxidant enzyme levels in liver. Nano Cur (weekly oral treatment for 16weeks at 20mg/kg b.wt) in DEN induced HCC rats exerted significant protection against HCC and restored redox homeostasis in liver cells. (...) Nano Cur was found to be a potential formulation in oral route in combating the oxidative damage of hepatic cells and eliminating DEN induced hepatocellular cancer cells in rat whereas identical amount of free Cur treatment was found almost ineffective.

PMID: 6206716

Four patients had pancreatitis associated with valproic acid therapy. Three patients received valproic acid at usual doses, and all were free of other symptoms of toxic reactions, with serum levels of valproic acid in the usual therapeutic range. (...) All patients recovered with discontinuation of valproic acid therapy and enteral feeding and administration of intravenous fluids. After recovery, a valproic acid regimen was restarted uneventfully (in one patient). (...) Pancreatitis is a serious complication of valproic acid therapy that must be considered in any patient receiving valproic acid who experiences severe abdominal pain and vomiting.

...

...

...

Annotated CID sentences
The theme of work was to evaluate effectiveness in oral route of polylactide co-glycolide (PLGA) Nanocapsulated curcumin (Nano Cur) against diethylnitrosamine (DEN) induced hepatocellular carcinoma (HCC) in rat.
Nano Cur (weekly oral treatment for 16weeks at 20mg/kg b.wt) in DEN induced HCC rats exerted significant protection against HCC and restored redox homeostasis in liver cells
Nano Cur was found to be a potential formulation in oral route in combating the oxidative damage of hepatic cells and eliminating DEN induced hepatocellular cancer cells in rat whereas identical amount of free Cur treatment was found almost ineffective
Pancreatitis is a serious complication of valproic acid therapy that must be considered in any patient receiving valproic acid who experiences severe abdominal pain and vomiting.
...

Fig. 2. An example of constructing silverCID corpus

This research is based on two assumptions: (i) Relations curated as *M* are CID relations and (ii) If two entities of a relation appear in the same sentence, it is highly probable that this sentence contains a grammatical relation. We do not know that these assumptions are correct in each case, therefore we consider this dataset to be a silver standard corpus.

The SilverCID corpus is constructed according to five steps as followed:

(i) *Relation filtering*: Filter chemical-induced disease relations in CTD database. This step use information from 'Direct evidence' field in CTD database, only relations marked as ''marker/mechanism' are chosen.

(ii) *Collecting*: Search for and collect all PubMed abstracts in the reference list of relations chosen in (i).

(iii) *Overlap removal*: To avoid overlap between the SilverCID corpus and the test set, we remove all the PubMeb abstracts which appeared in the CDR track data set to ensure that the use of this silver set results in a fair evaluation of its contribution.

(iv) *Annotating*: For each relation, automatically annotate all disease and chemical mentions of this relation in its referring PubMed articles.

(v) *Sentence filtering*: Filter sentences in abstract from (iv) which contain both chemical and disease entities of a relation. i.e., sentences that do not contain any entity or contain only one entity are removed.

Figure 2 illustrates the SilverCID corpus's construction steps.

Two novel aspects that makes the SilverCID corpus different from other sources are (i) it is built automatically and (ii) it is a sentence-level corpus, i.e, a set of sentences, in which each sentence contains at least one intra-sentence CID relation with its chemical and disease entities.

This data set contains 38,332 sentences, 1.25 millions tokens, 48,856 chemical entities (1,196 unique chemical entities), 44,744 disease entities (2,098 unique disease entities) and 48,199 CID relations (12,776 unique CID relations). It is available at URL: https://zenodo.org/record/34530 (doi:10.5281/zenodo.34530).

2.2 Proposed model

The overall architecture of the system is described in figure 3. Compared to our participation in the BioCreative CDR track, the improved system uses the SilverCID corpus for training in both DNER and CID phase, the impact on the results due to this improvement will be analyzed in the following sections. Pre-processing steps include sentence splitting, tokenization, abbreviation identification, stemming, POS tagging and dependency parsing (Stanford[1]). The main system presented here is based on the integration of several state-of-the-art machine learning techniques in order to maximize their strengths and overcome the weaknesses.

2.2.1 Named entity recognition and normalization

This module corresponds to the CDR sub-task DNER. It is a joint-inference model consisting of NER and NEN module to boost performance and reduce noises (13). In which, the NER and NEN modules are trained separately and then decoded simultaneously.

Following reports of high level performance in joint-inference model by Li and Ji, 2014 (13) and Zhang and Clark, 2008 (14), we decided to apply a structured perceptron model for NER. Its output has weighted-form, the same with that of the NEN model that is therefore suitable for joint-inference in the decoding phase. The structured perceptron is an extension of the standard perceptron for structured prediction by applying inexact search with violation-fixing update methods (15). It is trained on the CDR training, development set and SilverCID corpus with

[1] Stanford Dependencies: http://nlp.stanford.edu/software/stanford-dependencies.shtml

a standard lexicographic feature set: orthography features, context features, POS tagging features and dictionary (CTD) features.

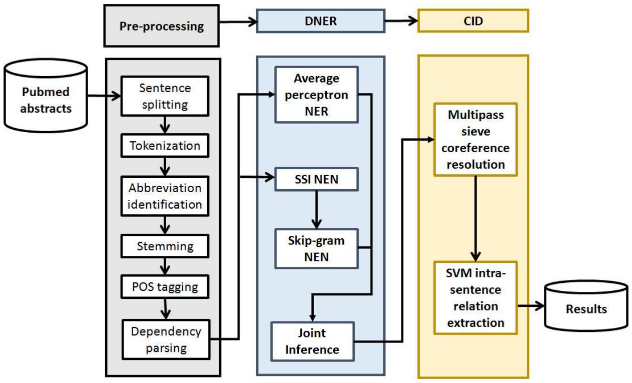


Fig. 3. Architecture of the proposed CDR extraction system. Model includes the pipeline of processing modules and material resources used, boxes with dotted lines indicate sub-modules.

The NEN module is a sequential back-off model based on two word embedding (WE) methods: semantic supervised indexing (SSI) (16) - a supervised WE methods, and skip-grams (17)– an unsupervised WE methods. The SSI model is trained on the CDR training and development set to obtain a correlation matrix *W* between tokens in the training data as well as MeSH. Skip-gram is a state-of-the-art word-to-vector method that takes advantage of large unlabeled data. We use an open source skip-gram model provided by NLPLab[2], which is trained on all PubMed abstracts and PMC full texts (4.08 million distinct words) with 200 dimensions. Several techniques are then applied to convert its output into 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.

The CID subtask requires system to extract chemical-disease relations at the abstract level. In simple cases, CID relations can be expressed in a single sentence (intra-sentence relation), its corresponding entities appear in the same sentences. Unfortunately, the CID relation may be expressed in multiple sentences (inter-sentence relation). Our system is based on a strategy that first converts inter-sentence relations to intra-sentence relations by using a coreference resolution method and then applies a machine learning model to extract them.

Our DNER system is based on joint inference using a modified beam search for decoding (13, 18), with which we train 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 (see formula 1).

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

The scoring function for NEN is:

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

If $W_{NEN} < W_{NEN}$ *(NONE)* = *threshold*, re-write formula 1 to formula 3:

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

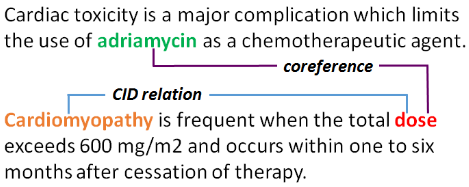
In which, W_{NER} is returned from the structured perceptron model.

2.2.2 Coreference resolution

Formally, coreference consists of two linguistic expressions - antecedent and anaphor (19). Figure 4 is an example of coreference, in which the anaphor 'dose' is the expression whose interpretation depends on that of the other expression, and the antecedent '*adriamycin*' is the linguistic expression on which an anaphor 'dose' depends.

Although the traditional coreference resolution task is to discover the antecedent for each anaphor in a document, from the perspective of this study, it is not necessary to always make clear which is the antecedent or anaphor. We consider both the antecedents and the anaphors as mentions of entities, and our system strives to recognize as much as possible mentions of an entity.

Studies on coreference resolution in the general English domain dates back to 1960s and 1970s and often focus on person, location and organization. In biomedicine, because that the types of entities that a coreference resolution system resolves are atypical to the general domain (i.e. protein, gene, disease, chemical, etc.), coreference research in this domain has received comparatively less attention (19). Previous approaches apply several methods, from heuristics-based (20, 21) to machine learning (22, 23).


Cardiac toxicity is a major complication which limits the use of adriamycin as a chemotherapeutic agent.

Cardiomyopathy is frequent when the total dose exceeds 600 mg/m2 and occurs within one to six months after cessation of therapy.

Fig. 4. An example of coreference between chemical entities. Two sequential sentences are extracted from PubMed abstract PMID: 7449470

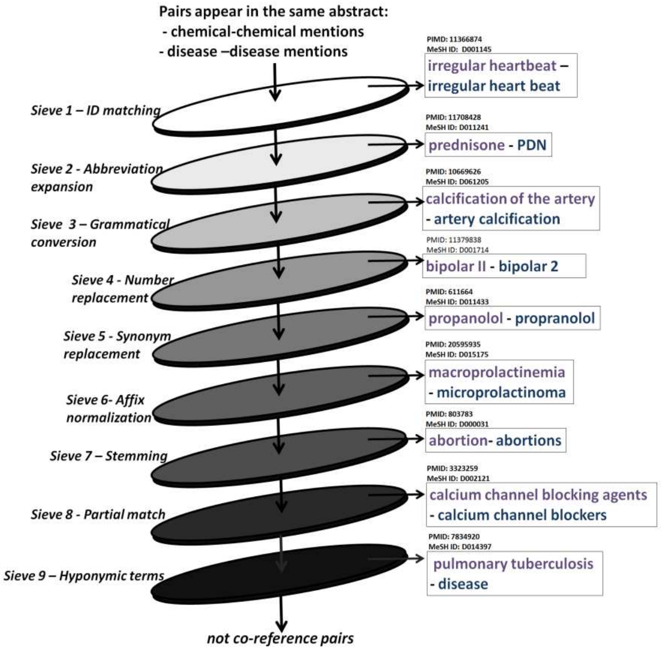

Pairs appear in the same abstract:
- chemical-chemical mentions
- disease –disease mentions
Sieve 1 – ID matching
Sieve 2 - Abbreviation expansion
Sieve 3 – Grammatical conversion
Sieve 4 - Number replacement
Sieve 5 - Synonym replacement
Sieve 6- Affix normalization
Sieve 7 – Stemming
Sieve 8 - Partial match
Sieve 9 – Hyponymic terms
not co-reference pairs
PIMD: 11366874
MeSH ID: D001145
irregular heartbeat – irregular heart beat
PMID: 11708428
MeSH ID: D011241
prednisone - PDN
PMID: 10669626
MeSH ID: D061205
calcification of the artery - artery calcification
PMID: 11379838
MeSH ID: D001714
bipolar II - bipolar 2
PMID: 611664
MeSH ID: D011433
propanolol - propranolol
PMID: 20595935
MeSH ID: D015175
macroprolactinemia - microprolactinoma
PMID: 803783
MeSH ID: D000031
abortion- abortions
PMID: 3323259
MeSH ID: D002121
calcium channel blocking agents - calcium channel blockers
PMID: 7834920
MeSH ID: D014397
pulmonary tuberculosis - disease

Fig. 5. Coreference resolution using nine-pass sieve

In our approach, the coreference module is based on the strategy called a multi-pass sieve by Souza and Ng, 2015 which was evaluated as a simple yet effective means for disorder mention normalization (21). Firstly, we process each abstract by noun phrase (NP) chunking (Genia

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

tagger[3]) and then create a set of NPs pairs for each abstract. These pairs of NPs are passed through the sieves, pairs which are kept by any sieve are considered as coreference pairs, pairs are not kept in each sieve continue pass through to the next sieve to the end. There are nine sieves, each corresponds to a set of rules, figure 5 is an illustration of the sieve-based coreference resolution module with example pairs that are kept by each sieve.

- *Sieve 1 – ID matching*: Two mentions of a chemical or a diseases having the same MeSH ID are coreferent. This sieve uses information coming from the previous NEN step. E.g., *'irregular heartbeat'* and *'irregular heart beat'* are both normalized to MeSH ID: D001145, thus, they are coreference.
- *Sieve 2 - Abbreviation expansion*: In this sieve we use the BioText Abbreviation recognition software[4] to identify abbreviations and their full forms (e.g, full form of 'PND' in abstract PMID:11708428 is *'prednisone'*) . We then check the MeSH ID of the full form and apply it to the abbreviation in order to unify mentions.
- *Sieve 3 – Grammatical conversion*: Generate similar forms of a mention by changing grammatical elements in mentions including subject, object, preposition, etc and then check the ID match criterion. New forms are obtained by applying rules proposed by D'Souza and Ng (21), which are (i) replacing the preposition in the name with other prepositions, (ii) dropping the preposition from the name and swapping the substrings surrounding it, (iii) bringing the last token to the front, inserting a preposition as the second token, and shifting the remaining tokens to right by two and (iv) moving the first token to the end, inserting a preposition as the second to last token, and shifting the remaining tokens to the left by two. Examples include '*calcification of the artery*' and '*artery calcification*', '*mental status alteration*' and *'alteration in mental status'*.
- *Sieve 4 - Number replacement*: Generate similar forms of a mention by replacing numbers with other forms and then check the ID match criterion. In this research, we consider the numeral, roman numeral, cardinal, and multiplicative forms of a number for generating new mention forms, i.e., '*two'* can be converted to *'2', 'ii'* and *'double'*.
- *Sieve 5 - Synonym replacement*: Check the ID match criterion for synonyms of mentions. This sieve uses a synonym dictionary constructed from MeSH, which contains 780,982 entries.
- *Sieve 6- Affix normalization*: Generate new forms of a mention by changing affixes (includes prefixes and suffixes) then check the ID match criterion. For examples, *'macroprolactinemia*' and '*microprolactinoma'* (PMID: 20595935), *'nephrotoxicity'* and *'nephrotoxic'* (PMID: 19642243) are coreference.
- *Sieve 7 – Stemming*: Mentions are stemmed using the Porter stemmer[5], and then check the ID match criterion.
- *Sieve 8 - Partial match*: This sieve uses the output information from the abbreviation expansion sieve and applies the criterion for partial matching as proposed by D'Souza,J. and Ng,V. (21). It is said that "a mention can be partially matched with another mention for which it shares the most tokens". To give an example, '*calcium channel blocking agents'* and *'calcium channel blockers'* in abstract PMID:3323259 are marked as coreference.
- *Sieve 9 – Hyponymic terms*: We create two dictionaries for chemical and disease includes hyponymic nouns which often referred to chemical/disease. For example, chemical hyponymic dictionary includes '*drug*', '*dose*', etc.; disease hyponymic dictionary includes '*disease*', 'case', *'infection', 'side effect',* etc. In this sieve, NER information is used to find chemical and disease entities, and if in its context window of two sentences before/after there is any term in dictionary, we can determine a coreference.

2.2.3 SVM intra-sentence relation extraction

In this research, we accept the statement that if a noun phrase and an entity are coreferent, the noun phrase can be considered as an entity of that type, too. The intra-sentence relation extraction module receives sentences that contain a disease -chemical pair as input and classifies whether this pair have CID relation or not.

The example in Figure 4 also shows how to combine the coreference resolution module and the intra-sentence relation extraction module for handling inter-sentence relation. The strategy is that if the intra-sentence relation extraction module can recognize the relation between '*cardimyopathy*' and 'dose', we can also determine the relation between '*cardimyopathy*' and '*adriamycin*' because '*dose*' and '*adriamycin*' is coreference.

The intra-sentence relation extraction module is based on a Support Vector Machine (SVM) (24) – one of the most popular machine learning methods which has been successfully applied for biomedical relation extraction (25, 26). We use the Liblinear tool[6] to train a supervision binary SVM classifier (L2- regularized L1-loss) on CDR track training/development data set and our SilverCID corpus. In this study, we observe that the complexities of CID relations (several structural forms, abundant related vocabulary set, difficult to determine the distance between the two entities, etc.) are similar to the event extraction problem. As a consequence, the feature set that is specially constructed for event extraction may work better than that commonly used for normal relation extraction (they are words, entity type, mention level, overlap, dependency, parse tree and dictionary (27-29)). Following reports of high performance in event extraction (30), we decided to use a rich features set including four types of features: Token features, neighboring token features, token features n-gram, pair features n-gram and shortest features path, the feature details are shown in table 2.

Table 2. Rich feature set used in intra-sentence relation extraction module

Feature types	Description	Features
Token features	Token itself information	- Token orthography (capitalization, first letter of sentence, number, etc.) - Base form of token - N-grams (*n*=1-4) of token - Part-of-speech tagging
Neighboring token features	Extracts all 2-step dependency paths from the target token, which then are used to extract n-grams	- Features extracted by the token feature function for each token - Token and dependency n-grams (*n*=2-4) - Token n-grams (*n*=2; 3) - Dependency n-grams (*n*=2)
Token n-gram features	Extract token n-grams (*n*=1-4) within a window of three tokens before and three tokens after the target token	- N-grams of word

[3] http://www.nactem.ac.uk/GENIA/tagger/

[4] http://biotext.berkeley.edu/software.html

[5] http://tartarus.org/martin/PorterStemmer/

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

Pair n-gram features	Extracts word n-grams (*n*=1-4) within a window from three tokens before the first tokens to three tokens after the last token in target chemical-disease pair.	- Dependency n-grams (*n*=2) - Token n-grams (*n* = 2, 3) - N-grams (*n* = 2-4) of dependencies and tokens
Shortest path features	Shortest dependency paths between two words (in which, each word belongs to a disease or chemical entity)	- Length of path - Word n-grams (*n*=2- 4) - Dependency n-grams (*n*=2-4) - Consecutive word n-grams (*n*=1-3) representing governor-dependent relationships - Edge walks (word-dependency-word) and their sub-structures - Vertex walks (dependency-word-dependency) and their sub-structures

3 Experimental results

3.1 Evaluation metrics

For evaluation, diseases entities and relations (chemical- disease pairs) are compared to golden standard annotated CDR test data set using standard metrics: precision (P), recall (R) and F1. P indicates the percentage of system positives that are true instances, and recall indicates the percentage of true instances that the system has retrieved. More formally this is shown by the equations 4, 5 and table 3.

$$Precision = \frac{TP}{TP+FP} \qquad (4)$$

$$Recall = \frac{TP}{TP+FN} \qquad (5)$$

F1 is the harmonic means of R and P, calculated as in equation 6.

$$F1 = \frac{2PR}{P+R} \qquad (6)$$

Table 3. Defining the evaluation metrics

		Gold standard annotation	
		Positive	Negative
System annotation	**Positive**	True positive (FP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

BioCreative V also evaluates participated systems' ability to return real-time results in a timely manner. It is calculated by response time via teams respective web services.

3.2 DNER results

The experimental results of DNER on the CDR track test data set are shown on table 4, note that only disease entities are evaluated. We compare our results with benchmarks and task results provided by BioCreative organizer, including:

- DNER benchmark 1: A straightforward dictionary look-up method that relied on disease names from CTD database.
- DNER benchmark 2: Retrained models using the out-of-box DNorm (16) which is a competitive system which achieved the highest performance in a previous disease challenge. DNorm combines an approach based on rich features and conditional random fields for NER with a pairwise learning to rank for NEN.
- DNER average results: Average results of 16 teams participated in the DNER task (best run for each participating team).
- DNER best team results: Results from team that is ranked number 1 (in term of F1) in DNER task (31). This system uses linear chain conditional random fields (CRF) with rich features for NER, they use three lexicons resources to generate CRF dictionary features and multiple post processing steps to optimize results. In NEN step, they use a dictionary-lookup method based on the collection of MEDIC, NCBI disease corpus and CDR task data set.

In this paper, we improve our system that participated DNER task by training NER perceptron model using silverCID corpus, new results are also listed on table 4.

Table 4. DNER results

	P (%)	R (%)	F (%)
Dictionary look-up*	42.71	67.46	52.30
DNorm*	81.15	80.13	80.64
DNER AVG results*	78.99	74.81	76.03
DNER No1 team*	89.63	83.50	86.46
Our system in BioCreative V*	**73.20**	**79.98**	**76.44**
Our improved system**	**79.90**	**85.16**	**82.44**
NER-NEN pipeline	78.26	83.17	80.64

*Results provided by the BioCreative 2015 organizer. **silverCID corpus is used for training NER module.

In the BioCreative V evaluation, our system performs outstanding compared to dictionary look up method, but worse than DNorm which is considered as a very strong benchmark (note that there are only 7 participating teams achieved performances better than DNorm). Using the silverCID corpus for training NER model can boost performance by 6% F1 and become better than the DNorm's result.

To demonstrate the benefit of joint inference model, we also build a baseline system that is based on the traditional pipeline model: NER is exerted first and its result is then used for NEN. In this manner, NER and NEN module is totally similar with our joint inference model and also trained on the silver-CID corpus. The results show that joint inference model boosts the performance by 1.8% of the F1 score.

Following the results reported by BioCreative (8), the average response time in the DNER task is 5.6 seconds and our system is among participating systems that have smallest response time (276 ms, ranked no. 2).

3.3 CID results

Table 5 shows the results of our system on the CID task. It serves for two purposes, i.e. the first for comparing our results with the BioCreative benchmark results, and the second for evaluating the contribution of coreference resolution approach and silver-CID corpus as well as finding the best combination of them. The BioCreative benchmarks include:

- The CID benchmark results from the co-occurrence baseline method with two variants: abstract-level and sentence-level.
- The CID average results of 18 teams participated in the CID task (best run for each team).
- The CID best team results from team that is ranked number 1 (in term of F1) in the CID task (32). This system combines two SVM classifiers trained on sentence- and document-level, its novel as-

pect is at using rich features coming from CID relations in other biomedical resources.

Table 5. CID results

	P (%)	R (%)	F (%)
Co-occurrence*	16.43	76.45	27.05
CID AVG result*	47.09	42.61	43.37
CID best team*	55.67	58.44	57.03
Our system in BioCreative V (SVM+ CR MPS)*	**53.41**	**49.91**	**51.60**
Our improved system (SVM+ CR MPS+ silverCID corpus)	**57.63**	**60.23**	**58.90**
SVM	44.73	50.56	47.47
SVM+ silverCID corpus	51.42	52.81	52.11
SVM+ CR EMC	47.64	50.28	48.93

SVM: SVM intra-sentence relation extraction. CR: Coreference resolution. EMC: expectation maximization clustering. MPS: Multi-pass sieve. *Results provided by the BioCreative 2015 organizer.

The configuration of our system which participated in the CID task is the pipeline of multi-pass sieve coreference resolution module and the SVM intra-sentences relation extraction module, achieving 51.60% F1. This is much better than the co-occurrence benchmark method. Further, using the SilverCID corpus for training SVM module can boost performance by 7.3% of F1. It can be noted that this result exceeds the highest ranking system in the CID task. However evaluation results for biomedical relation extraction methods vary greatly and are largely incomparable across different studies – particularly in this case because the use of the SilverCID corpus would not have been allowed under the original rules of the task because it was unknown which subset of the database was used in the test evaluation. Thus, we should be cautious in reading too much into such a direct comparison. Note that since DNER is the initial step of CID, DNER results greatly influenced the CID results. Therefore, the comparison hereby requires further validations because we use NER and NEN information provided by our DNER phase while other systems use theirs.

The contribution coreference resolution and silverCID corpus are evaluated by comparing results of SVM based intra-sentence relation extraction module with and without adding coreference resolution module/silverCID corpus. A comparative evaluation between systems with different combination strategies shows that an original SVM approach (only trained on CDR training and development set) achieved F1 of 47.47%, whilst adding the SilverCID corpus boosts F1 by 4.64 % (51.60%) and adding multi-pass sieve coreference resolution module boosts F1 by 4.13% more (58.90%).

We also make a comparison between our heuristic-based multi-pass sieve method and another state-of-the-art machine learning based method for coreference resolution. In this regard, we re-implement a method proposed by Ng (22), it is an expectation maximization (EM) clustering co-reference resolution - an unsupervised machine learning method. This system uses intra-sentence relation extraction SVM model trained on the CDR training and development set. The results demonstrate the strength of our multi-pass sieves method. We achieve 53.41% in precision (5.77% better than that of the EM clustering-based), 49.91 % in recall (0.37% worse) and 51.60% in F1 (2.67% better).

The feature set that used in SVM model contains 332,570 features-this is a clearly a non-trivially large feature space to compute. In our experiments, the SVM model takes more than an hour for training. According to the results reported by BioCreative (8), the average response time in CID task is 9.3 seconds and our system response time is 8.993 second.

4 Discussion

Firstly, we emphasize that our silverCID corpus is built automatically whilst other resources which are based on the CTD database, such as CDR track data set (section 2.1.1) and the CTD-Pfizer collaboration data set (11) are a results of manual curation. A further novel aspect that makes the SilverCID corpus different from other sources is that it is a sentence-level corpus, which is especially built in order to serve the purpose of CID relation extraction.

Traditionally, NER and NEN is exerted as two separate tasks, in which, NEN takes the output of NER as its input. Following Liu et al. (33), one big limitation of this pipeline approach is that errors propagate from NER to NEN and there is no feedback from NEN to NER, Khalid et al. (34) also demonstrated that most NEN errors are caused by recognition errors. Joint inference is expected to overcome these disadvantages of such a traditional pipeline model. The results in table 4 show that join inference model boots performance by 1.8% in term of the F1 score. 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*".

In the DNER phase, the NEN back-off model can take advantage of both labeled CDR dataset and extremely large unlabeled data. SSI calculates the correlation 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 the recall (e.g. Skip-gram link '*disordered gastrointestinal motility*' to MeSH:D005767, '*hyperplastic marrow*' to MeSH:D001855, they are false negative of SSI).

In the CID phase, we compared the true positive results of three comparative systems (Table 6). This includes (i) SVM model which is only trained on CDR training and development data, (ii) pipeline model of above SVM and multi-pass sieve coreference resolution and (iii) the same model as in (ii), but with the silver-CID corpus used for training the SVM model. The disagreements between these three systems clarify contribution of method and data set used in our model. SVM intra-sentence relation extraction model plays the central core of our system, it works in cases of intra-sentence CID relations (example 1 and 2), thus, if SVM fails on an intra-sentence relation, adding multi-pass sieve coreference resolution module is not helpful (example 3 and 4). Since the silverCID corpus enriches the training data of SVM, using it may help to find more relations than only SVM model does (example 3). It, however, also may bring some noises lead to the small adverse effects for the system, i.e., adding silverCID lead to missing the results (example 2). It is certain that only SVM model, even trained on silverCID corpus or not, cannot catch the inter-sentences relation (example 5-8). Therefore, coreference resolution is completely necessary for handling inter-sentence relation (example 5 and 7). Similar to intra-sentence relation cases, adding the silverCID corpus may help (example 6) or reject a very small amount true positive result classified by SVM+ coreference model (example 7). Example 4 and 8 are failed by all systems.

Table 6. Analysis of the contribution of methods and resources used in the proposed system for capturing CID relationships

	Chemical - Disease relation example	PMID	Type of relation		SVM	SVM +CR	SVM +SS	SVM +CR +SC
			Intra-	Inter-				
1	maleate (C030272) - nephrotoxicity (D007674)	25119790	✓		✓	✓	✓	✓
2	quinacrine hydrochloride (D011796) - atrial thrombosis (D003328)	6517710	✓		✓	✓		
3	metolachlor (C051786) -liver cancer (D008113)	26033014	✓				✓	✓
4	galantamine (D005702) – headaches (D006261)	17069550	✓					
5	methoxamine (D008729)- headache (D006261)	11135381		✓		✓		✓
6	gemfibrozil (D015248) - myositis (D009220)	1615846		✓				✓
7	oxidized and reduced glutathione (D019803) - reperfusion injury (D015427)	1943082		✓		✓		
8	metolachlor (C051786)- follicular cell lymphoma (D008224)	26033014		✓				

SVM: SVM intra-sentence relation extraction. CR: Multi-pass sieve coreference resolution. SC: silverCID corpus. Intra-: Intra-sentence CID relation. Inter-: Inter-sentence CID relation. ✓: Chemical-disease pair is classified as CID relation correctly.

Table 7 shows examples of where our system disagreed with the annotation standard. There are two types of errors: missing results (FN) and wrong results (FP). Some errors are caused by the previous DNER phase: in example 4, NER module did not recognize '*calcium channel blockers*' as chemical; in example 5, FP result '*acute insulin*' of NER module leads to the FP error of the whole system; in example 6, NER determined the wrong boundary of entity '*acute hepatitis'* and in example 7, NEN module matched '*heart hypertrophy*' to the wrong MeSH ID (it should be D006332). However, for errors caused by the system, since entities in relations are expressed by MeSH ID and evaluation is made in abstract-level, it is very hard to clarify cause of errors. Our comments for these cases are empirical, based on heuristic surveying the system output. Inter-sentences relations often have very complex structures, two entities may belong to two sentences that are not adjacent, in some cases, one entity even is hidden, which causes many FN errors (example no. 1- 2). The SVM module depends on the training data set, thus, it may lead to several limitation of finding new relations (which are not similar with relations in the training set) and classifying confusion (example no. 9). Coreference resolution is not a trivial problem, it has several types of errors by itself, FP in example 10 seems to be brought about by the coreference resolution module, i.e. linking the term 'dose' to the wrong entity. One more type of errors caused by the silverCID corpus: We know that this corpus may bring much more valuable information, but it also may bring some noise, leading to the FN errors (example no. 3) and the FP errors (example no. 8) which would disappear if we remove silverCID corpus from our system.

Table 7. Sources of errors by our system system on the CDR test set

	Relation	PMID	Type of error		Cause of error
			FP	FN	
1	corticosteroid (D000305) - systemic sclerosis (D012595)	22836123		✓	Complex inter-sentence structure
2	cyclophosphamide (D003520) - edema (D004487)	23666265		✓	Complex inter-sentence structure
3	chlorhexidine diphosphanilate (C048279) - pain (D010146)	2383364		✓	Noise from silverCID corpus
4	theophylline (D013806) - tremors (D014202)	3074291		✓	Error from NER
5	scopolamine (D012601) - retention deficit (D012153)	3088653	✓		Error from NER
6	clopidogrel (C055162) - acute hepatitis (D017114)	23846525	✓		Error from NER
7	isoproterenol (D007545) - heart hypertrophy (D006984)	2974281	✓		Error from NEN
8	nicotine (D009538) - anxiety (D001008)	15991002	✓		Noise from silverCID corpus
9	oxitropium bromide (C017590) - nausea (D009325)	3074291	✓		Error from SVM model
10	gamma-vinyl-GABA (D020888) - status epilepticus (D013226)	3708328	✓		Error from coreference resolution module

Intra-: Intra-sentence CID relation. Inter-: Inter-sentence CID relation. FP: False positive. FN: False negative.

5 Conclusions

In this article we have presented our systematic study in an attempt to participate the BioCreative V CDR task: (i) A joint inference approach to NER and NEN based on several state-of-the-art machine learning methods for the DNER sub-task and (ii) Improving a SVM based model by using a rich feature set, silverCID corpus and crucially, a multi-pass sieve coreference resolution module for the CID sub-task. Our top performing configuration achieved an F1 of 81.93 for DNER and 58.90 for CID. This result is better than the performance of DNorm – a very strong DNER benchmark and exceeds the highest ranking system in the CID task.

Based on the CTD database, we built a silver standard data set (called silverCID corpus), including 51,719 sentences that contains CID relations with silver annotations of NER, NEN and CID relation. This silverCID corpus has proven its effectiveness by boosting the system performance by 7.3% in term of the F1 (note that there are no overlap between CTD-silver set and test set).

Several comparisons were made to compare our results with other system and analyze the system errors. The evidence points towards complementarities between the SVM model, the use of the SilverCID corpus and the co-reference resolution module. We observed the advantage of using multi-pass sieve coreference resolution to handle inter-sentence relations.

Our system is extensible in several ways. Firstly from improvements to the coreference resolution module: although the coreference resolution module plays a central role in extracting inter-relation, at this time, it only boosted performance by 4.13% in term of the F1. We are going to use the SilverCID corpus for training a multi-pass sieve coreference module, and the more results coreference resolution module find, the more inter-relations can be found. The second approach comes from several useful biomedical resources that we did not utilize. In the report of best team in the CID sub-task of BioCreative V (32), they exploited many databases such as CTD, MEDI (35), SIDER (36), etc., they can be used as knowledgebase features for machine learning or dictionary for matching. The third approach is application of several post-processing steps, such as abbreviation resolution and consistency improvement, which was applied by the best team in DNER sub-task of the BioCreative V and showed its effectiveness (31).

Acknowledgements

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

THD gratefully acknowledges partly funding support from Vietnam National University, Hanoi (VNU), under Project No. QG.15.21

References

1. Dogan,R.I., Murray,G.C., Névéol,A. et al. (2009) Understanding PubMed user search behavior through log analysis. *Database, 2009, bap018.*
2. Davis,A.P., Murphy,C.G., Saraceni-Richards,C.A. et al. (2009) Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical–gene–disease networks. *Nucleic acids research, 37(suppl 1), D786-D792.*
3. Leaman,R., Wojtulewicz,L., Sullivan,R. et al. (2010) Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. *Proceedings of the 2010 workshop on biomedical natural language processing (pp. 117-125). Association for Computational Linguistics.*
4. Liu,J., Li,A. and Seneff, S. (2011) Automatic drug side effect discovery from online patient-submitted reviews: Focus on statin drugs. *Proceedings of First International Conference on Advances in Information Mining and Management (IMMM), Barcelona, Spain (pp. 23-29).*
5. Kang,N., Singh,B., Bui,C. et al. (2014) Knowledge-based extraction of adverse drug events from biomedical text. *BMC bioinformatics, 15(1), 64.*
6. Sugihara,D., Masuichi,H. and Ohe, K. (2010) Adverse–Effect Relations Extraction from Massive Clinical Records. *23rd International Conference on Computational Linguistics (p. 75).*
7. Hammann,F., Gutmann H., Vogt N. et al. (2010) Prediction of adverse drug reactions using decision tree modeling. *Clinical Pharmacology & Therapeutics, 88(1), 52-59.*
8. Wei,C.H., Peng,Y., Leaman,R. *et al.* (2015) Overview of the BioCreative V Chemical Disease Relation (CDR) Task. *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, 154-166.*
9. Li,J., Sun,Y., Johnson,R. *et al.* (2015) Annotating chemicals, diseases, and their interactions in biomedical literature. *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, 154-166.*
10. Lipscomb CE (2000) Medical subject headings (MeSH). *Bulletin of the Medical Library Association 88: 265.*
11. Davis,A.P., Wiegers,T.C., Roberts,P.M. et al. (2013) A CTD–Pfizer collaboration: manual curation of 88 000 scientific articles text mined for drug–disease and drug–phenotype interactions. *Database, 2013, bat080.*
12. Hamosh,A., Scott,A.F., Amberger,J.S. et al. (2005) Online mendelian inheritance of man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res., 33(Suppl 1), D514–D517.*
13. Li,Q. and Ji,H. (2014) Incremental joint extraction of entity mentions and relations. *Proceedings of the Association for Computational Linguistics.*
14. Zhang, Y. and Clark, S. (2008) Joint Word Segmentation and POS Tagging Using a Single Perceptron. *Proceedings of the Association for Computational Linguistics (pp. 888-896).*
15. Huang,L., Fayong,S. and Guo,Y. (2012) Structured perceptron with inexact search. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 142-151). Association for Computational Linguistics.*
16. Leaman,R., Doğan,R.I. and Lu,Z. (2013) DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics, btt474.*
17. Mikolov,T., Sutskever,I., Chen,K. *et al.* (2013) Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems (pp. 3111-3119).*
18. Miwa,M. and Sasaki,Y. (2014) 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 (pp. 1858-69).*
19. Zheng,J., Chapman,W.W., Crowley,R.S. et al. (2011) Coreference resolution: A review of general methodologies and applications in the clinical domain. Journal of biomedical informatics, 44(6), 1113-1122.
20. Lee,H., Peirsman,Y., Chang,A. et al. (2011) Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task (pp. 28-34). Association for Computational Linguistics.
21. D'Souza,J. and Ng,V. (2015) Sieve-Based Entity Linking for the Biomedical Domain. *Proceedings of ACL-IJCNLP Volume 2: Short Papers, 297.*
22. Ng,V. (2008) Unsupervised models for coreference resolution. *Proceedings ofthe Conference on Empirical Methods in Natural Language Processing (pp. 640-649). Association for Computational Linguistics.*
23. Bejan,C.A. and Harabagiu,S. (2010) Unsupervised event coreference resolution with rich linguistic features. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 1412-1422). Association for Computational Linguistics.
24. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20(3), 273-297.*
25. Song,S.J., Heo,G.E., Kim,H.J. et al. (2014) Grounded Feature Selection for Biomedical Relation Extraction by the Combinative Approach. *Proceedings of the ACM 8th International Workshop on Data and Text Mining in Bioinformatics (pp. 29-32).*
26. Kim,S., Liu,H., Yeganova,L. et al.(2015) Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach. *Journal of biomedical informatics, 55, 23-30.*
27. Kambhatla,N. (2004) Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions (p. 22).*
28. GuoDong,Z., Jian,S., Jie,Z. et al. (2005) Exploring various knowledge in relation extraction. *Proceedings of the 43rd annual meeting on association for computational linguistics (pp. 427-434).*
29. Jiang,J. and Zhai,C. (2007) A Systematic Exploration of the Feature Space for Relation Extraction. *In HLT-NAACL (pp. 113-120).*
30. Miwa,M., Sætre,R., Kim,J.D. et al. (2010) Event extraction with complex event classification using rich features. *Journal of bioinformatics and computational biology, 8(01), 131-146.*
31. Lee,H.C., Hsu,Y.Y. and Kao, H Y. (2015) An enhanced CRF-based system for disease name entity recognition and normalization on BioCreative V DNER Task. *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, paper no 34.*
32. Xu,J., Wu,Y., Zhang,Y. et al. (2015) UTH-CCB@ BioCreative V CDR Task: Identifying Chemical-induced Disease Relations in Biomedical Text. *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, paper no 38.*
33. Liu,X., Zhou,M., Wei,F. et al. (2012) 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 (pp. 526-535). Association for Computational Linguistics.*
34. Khalid,M.A., Jijkoun,V. and De Rijke, M. (2008) The impact of named entity normalization on information retrieval for question answering. *Advances in Information Retrieval (pp. 705-710). Springer Berlin Heidelberg.*
35. Wei,W.Q., Cronin,R.M., Xu,H., et al. (2013) Development and evaluation of an ensemble resource linking medications to their indications. *Journal of the American Medical Informatics Association: JAMIA 20, 954-961.*
36. Kuhn,M., Campillos,M., Letunic,I., et al. (2010) A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology 6, 343.*