Resolution of Chemical Disease Relations with Diverse Features and Rules

Dingcheng Li*, Naveed Afzal*, Majid Rastegar Mojarad, Ravikumar Komandur Elayavilli, Sijia Liu, Yanshan Wang, Feichen Shen, Hongfang Liu

Biomedical Statistics & Informatics, Mayo, Clinic
200 First Street, Rochester, MN, USA, 55901

*li.dingcheng@mayo.edu; afzal.naveed@mayo.edu; mojarad.majid@mayo.edu; komandurelayavilli.ravikumar@mayo.edu; liu.sijia@mayo.edu; wang.yanshan@mayo.edu; shen.feichen@mayo.edu; liu.hongfang@mayo.edu

Abstract. This paper describes the system developed by Mayo Clinic team for the extraction of Chemical-Disease relations. We employed two approaches: a rule-based approach to extract relations within a single sentence and a machine learning approach that uses diverse set of features to extract relations from a single sentence as well as multiple sentences. We trained the machine learning approach and designed rules based on the 750 PubMed abstracts (500 from training and 250 development dataset) and used the remaining 250 PubMed abstracts from the development dataset for blind evaluation. The rule-based approach was able to achieve an F-score of 41% when used with gold-standard named entity annotations on testing data and 31% when used with organizer provided named entity annotation tools while machine learning approach attain F-score of 65% for gold-standard named entity annotations and 42% when used with organizer provided named entity annotation tools.

Keywords. Chemical-induced diseases; Rule-based; Machine Learning; UIMA;

1 Introduction

Chemicals, diseases, and their relations (CDR) are among the most studied topics by biomedical researchers and healthcare worldwide for drug discovery or safety surveillance [2, 5 and 6]. Manual annotation of CDR from unstructured free text into structured knowledge has become an important theme for several bioinformatics databases such as the Comparative Toxicogenomics Database (CTD) [1, 4]. However, manual curation of chemical diseases relations from the literature is laborious, costly and insufficient due to the rapid literature growth.

Despite these previous attempts such as [3], free text-based automatic biomedical relation detection remains challenging. In order to address this challenge, we constructed a UIMA-based (unstructured information management application) pipeline to detect CDR from PubMed abstracts with both machine learning and rule-based approaches.

2 Methods

A. Rule-based Approach

To begin with, we split the abstract into individual sentences, tagged the individual tokens with part of speech using NLTK[1] and identified the lemma of the individual words using BioLemmatizer[2]. Subsequently, we identified the chemical and disease mentions in both title and abstract sections (i.e. disease or chemical) and normalized them to their MeSH ID's. We replaced the entity mentions with the normalized one.

In this work, we limited our rule-based CDR extraction to only a single sentence. Our rule-based system considers only those sentences that contain at least one disease and chemical.

Table 1 lists some of the patterns that we developed to extract CDR from a single sentence. For example, Rule 1 (Table 1) matches the following sentence: "*Two patients had a cholestatic hepatitis induced by carbimazole.*" to extract the relation between "*cholestatic hepatitis*" and "*carbimazole*". However the design of the rule is generic enough to allows selective intervening words between the disease and the verb. Similarly, while the pattern allows for any verb phrase it is further constrained by the immediate prepositional phrase ("by").

Table 1: Rule-based patterns

Rule ID	Rule-based Patterns
1	DISEASE [VERB]+ by CHEMICAL
2	CHEMICAL [VERB]+ [ADJECTIVE]+ DISEASE
3	DISEASE (after|during|with) [ADJECTIVE/NOUN]+ CHEMICAL
4	CHEMICAL [NOUN]+ [PREPOSITION]+ [ADJECTIVE]+ DISEASE

We applied certain post-processing filters to the patterns in order to improve the precision of the system. Filters include features like, negation, disallowing overlapping relations; ontology based filtering to

[1] http://www.nltk.org/

[2] http://biolemmatizer.sourceforge.net/

eliminate non-specific/generic relations. We encountered certain instances where two chemicals and a disease were mentioned in a single sentence. For example in an abstract (PMID-20003049) we have the following phrase "**argatroban** in a cardiac transplant patient … of **heparin**-induced **thrombocytopenia**" where the system initially extracts two CDR pairs (argatroban-thrombocytopenia and heparin-thrombocytopenia). However the system filters CDR pair (argatroban-thrombocytopenia) that overlaps with another pair (heparin-thrombocytopenia).

In biomedical abstracts, authors quite often mention multiple diseases, one being more specific than the other. In the CDR documentation, it is mentioned that a chemical is paired with the most specific disease mentioned in the abstract. In our system, we used MeSH ontology to identify the most specific CDR pairs from abstracts. If two or more diseases are mentioned in the abstract and they belong to the same hierarchy in MeSH ontology then we retain that pair which has the most specific disease. For example, in an abstract (PMID-3070035) the system extracted two CDR pairs namely, "captopril-renal failure" and "captopril-sudden deterioration of renal function". However "renal failure" and "sudden deterioration of renal function" share hierarchical relationship and the latter being the most specific disease, the system filters the former relation.

B. *Machine Learning Approach*

We employed a supervised machine-learning model as an alternative to our rule-based approach. Figure 1, shows the overall architecture of our model. The collection reader is designed to read BioC format data. The aggregate analysis engines generate diverse rules and relevant annotations, including tokens, semantic roles, dependency trees, syntactic constituents and candidate chemical-disease relation pairs. All annotations are saved in UIMA data structure allowing feature extractions to be accomplished in an on-line fashion. We employed MALLET data structure to pass and deliver feature vectors. We evaluated several machine learning models (such as MaxEnt, NaïveBayes, LibSVM etc.) during the training phase, and decided to use LibLinear as it produced the best results. Parameters of all classifiers were optimized by 10-fold cross-validation on the training dataset.

In this approach, we extracted relations from a single sentence, successive sentences and across multiple sentences. For this approach, besides the annotated instances, which refer to related CDR pairs, it is important to add negative instances. In this work, negative instances refer to “no relation” between chemicals and diseases. However, if all pairs of chemical and diseases were used as candidates, the number would be quite large so that it is impossible to train an effective predictive model. Further, it is also found that a CDR pair may appear multiple times in one abstract but pairs, which are closer in token distance, are more relevant. Therefore, we resort to a priority assignment strategy (the closer the pair is, the higher the priority is given) to reduce the number of instances. After candidate CDR pairs were generated and selected, the next step is a binary classification task. We assign candidates that appear in the gold standard of training data with *TRUE* and assign *FALSE* to rest of the candidate pairs. We use two sources for feature generation. First, relations determined by rule-based approaches are employed as one of the features. Secondly, features listed in Table 2 are found to be discriminative for detecting CDR relations.

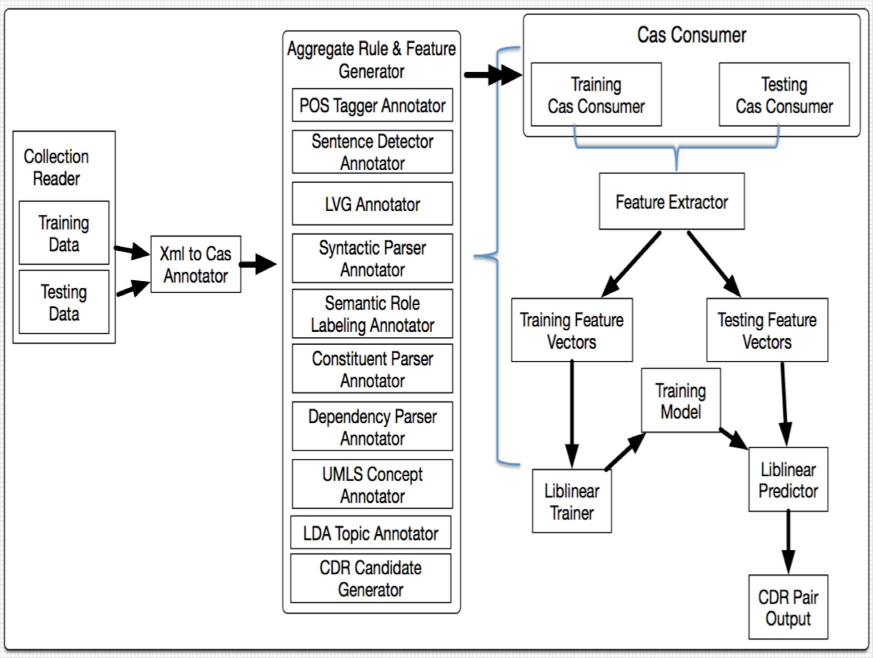

Figure 1: Architecture of UIMA pipeline for CDR Detection

Table 2: Machine learning features

Bag-of-words	Treating each entity as a word, the unigram, bigram and trigram of context within a window of [-2, 2] were extracted.
Part-of-speech (POS)	The unigram, bigrams, trigrams of POS within a window of [-2, 2].
Dependency features	The path of words and the path of POS in related CDR pairs.
Semantic role features	Semantic roles come from shallow semantic parsing, consisting of the detection of the semantic arguments associated with the predicate or verb of a sentence and their classification into their specific roles.
Topic distribution	Topic distributions of each word in the context of chemical and disease pair, generated by Latent Dirichlet Allocation (LDA).
Distance	The distance between two candidate entities in number of words, and the count of the number of entities between two candidates.
Overlapping chemical distance	The number of chemicals between chemicals and disease pairs. (see rule-based approach)
Conjunction	The conjunction between two entities of the CDR pair.
MeSH filtration	MeSH ontology employed aiming at filtering out generic CDR and retain only the more specific CDR.

3. Results

We randomly divided the development dataset into two equal parts: one part for final testing while the other part was used for training performance adjustment along with the training dataset. Overall, our training data contained 750 PubMed abstracts (500 abstracts from training dataset and 250 abstracts from development dataset) while testing data contained 250 unseen PubMed abstracts from the development dataset. Table 3 shows the results of both systems with gold-standard named entity annotations and results with organizer-provided NER tools [7, 8].

Table 3: Results of rule-based and machine learning systems

Using gold standard named entity annotations			
System	**Precision**	**Recall**	**F-score**
Rule-based system	0.59	0.31	0.41
ML system	0.77	0.56	0.65
Using named entity annotations from DNorm and tmChem			

Rule-based System	0.47	0.23	0.31
ML system	0.40	0.46	0.42

4. Discussion

One of the main reasons for low performance of rule-based approach was due to its fundamental design of having token level constraint while matching the sentences. For machine learning approach, the right combinations of features play essential roles in boosting the prediction performance. Predetermined CDR pairs by rules are strong discriminative features. Through division of classifiers, diverse feature engineering, we have obtained promising results. In future work, we will explore more discriminative features and kernel methods for machine learning and dependency-based relations for rule-based system so that a more practical end-to-end hybrid system will be constructed with high performance.

Acknowledgment

We would like to thank our IT team for their support, Kaggal Vinod and Matthew Bockol in particular. This work was made possible by joint funding from National Institute of Health 1K99LM012021-01A1, R01GM102282, R01LM11369, R01LM11829, and R01 LM011934.

REFERENCES

1. Davis, A. P., et al. (2011). "The comparative toxicogenomics database: update 2011." Nucleic acids research 39(suppl 1): D1067-D1072.
2. Dogan, R. I., et al. (2009). "Understanding PubMed® user search behavior through log analysis." Database 2009: bap018.
3. Gurulingappa, H., et al. (2012). "Extraction of potential adverse drug events from medical case reports." J Biomed Semantics 3(1): 15.
4. Kang, N., et al. (2014). "Knowledge-based extraction of adverse drug events from biomedical text." BMC bioinformatics 15(1): 64.
5. Lu, Z. (2011). "PubMed and beyond: a survey of web tools for searching biomedical literature." Database 2011: baq036.
6. Névéol, A., et al. (2011). "Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction." Journal of Biomedical Informatics 44(2): 310-318.
7. Leaman R., Doğan R.I. and Lu Z. (2013)." DNorm: Disease Name Normalization with Pairwise Learning to Rank", Bioinformatics (2013) 29 (22): 2909-2917.
8. Leaman R, Wei C-H, Lu Z (2015)." tmChem: a high performance tool for chemical named entity recognition and normalization", Journal of Cheminformatics, 7(Suppl 1): S3.