A Hybrid System for Extracting Chemical-Disease Relationships from Scientific Literature

Halil Kilicoglu and Willie J. Rogers

Lister Hill National Center for Biomedical Communications
National Library of Medicine
National Institutes of Health
Bethesda, MD, 20894
{kilicogluh,wjrogers}@mail.nih.gov

Abstract. We propose a hybrid system for extracting chemical-disease relationships from Medline abstracts. At the core of our approach is a general, rule-based system that extracts causal relations from text, using a combination of trigger lists and syntactic dependencies. We augmented this system with supervised learning. We trained two binary classifiers: one extracts intra-sentential relationships between chemical-disease mention pairs, and the other attempts to extract relationships across sentences. Our hybrid system yielded an F_1 score of 36.49. Our results on the development corpus reveal that chemical and disease named entity recognition are still major problems, and that improvements made in this area are likely to have a significant impact in chemical-disease relationship extraction.

Key words: rule-based relation extraction, binary classification, discourse-level relations

1 Introduction

With the exponential growth in biomedical literature and user-generated health content, it is critical to develop automated systems that can assist pharmacovigilance and curation tasks. Task 3 of the BioCreative 5 challenge [12] aims to stimulate research in this area and consists of two sub-tasks: disease named entity recognition (DNER) and chemical-induced disease relation extraction (CID), the former task an intermediate step for the latter. A corpus of 1,000 Medline abstracts annotated with chemical and disease mentions and the relationships between them is provided for training and validation [7]. The participating systems are evaluated on 500 abstracts, through web services, in order to evaluate not only the accuracy of the systems but also their response time and scalability.

Of the two sub-tasks, our interest mainly lies in the relationship extraction (CID) task. For this task, we propose a hybrid approach. We adapted a general, rule-based relation extraction system that deals with causal relations to the CID task and expanded it with supervised machine learning models to recover more challenging sentence-bound relationships as well as relationships that cross sentence boundaries. Our results show that the hybrid approach is more effective than both rule-based and machine learning approaches, even though there is much room for improvement. Our experiments with the development data also showed that accurate identification of chemical-disease mentions in text is critical for relation extraction; using gold standard named entity annotations yielded an approximately 62% increase in F_1 score.

2 Methods

We used a pipeline architecture; named entity recognition/ normalization was followed by rule-based relation extraction and, if necessary, sentence-bound and discourse-level extraction using supervised models. We used DNorm [5] for disease and tmChem [6] for chemical name recognition/normalization. DNorm was retrained on the training corpus while tmChem was used with pre-built models.

2.1 Rule-based relation extraction

Similarly to the CID task, the biological event extraction task [4] considers causal relationships, such as POSITIVE_REGULATION. To take advantage of the similarity, we adapted a linguistically grounded, rule-based system that was originally developed for the BioNLP shared task on event extraction [3] and enhanced since. The system adopts a two-phase approach. In the first phase (Composition), a general, underspecified semantic interpretation is composed from syntactic dependency relations in a bottom-up manner. This phase presupposes named entities, and relies on a trigger dictionary and argument identification rules to extract predicate-argument structures (PAS) consisting of the trigger and its logical subject, object, and adjunct arguments. The second phase is meant to be task-specific and, for the CID task, is concerned with tailoring the resulting semantic interpretation to the CID task requirements.

Disease and chemical mentions were recognized and normalized using DNorm and tmChem, respectively. We used a existing dictionary of causal indicators, previously compiled from several corpora, as relation triggers. For the CID task, the trigger list consists of 201 triggers and mainly includes triggers for REGULATION and POSITIVE_REGULATION events from the the BioNLP shared task corpora (e.g., *induce*, *effect*, *role*) as well as discourse connectives that describe causal (e.g. *as a result*) or temporal relations (e.g., *before*, *after*) compiled from the Penn Discourse TreeBank [10]. For each trigger, the dictionary encodes part-of-speech, lemma, and the dependency patterns that can be used by argument identification rules, described below. Dependency relations (collapsed format) were extracted using the Stanford CoreNLP toolkit [8].

After recognizing the named entity and trigger mentions in text, the Composition phase proceeds with transforming the syntactic dependency graph for each sentence into an intermediate semantic graph through a series of transformation rules. These rules serve several purposes: a) integrating semantic information into the dependency graph, b) making the semantic dependencies between textual units explicit, c) correcting potential errors in dependency relations, and

d) handling syntactic phenomena, such as coordination (for more details about these transformations, see [2]). Once a semantic graph of a sentence is formed, a bottom-up traversal of the graph, guided by *argument identification rules*, is performed to determine the logical arguments of the triggers (logical subject, logical object, and adjuncts). The argument identification rules define a mapping from a lexical category and a dependency pattern to a logical argument type. Inclusion and exclusion constraints for these rules can also be defined. Two such rules are given in Table 1.

Category	**Pattern**	**Include**	**Exclude**	**Argument**
NN	*prep_on*	*influence, impact, effect*	-	Object
VB	*agent*	-	-	Subject

Table 1. Argument identification rules

The first rule indicates, for example, that the object argument of *effect* can be found along the dependency path that begins with the outgoing *prep_on* arc. Composition phase yields a list of PASs. An example sentence and the PAS extracted from it are given in Example (1).

(1) (a) *A 2-year-old child with known neurologic impairment developed a dyskinesia soon* ***after*** *starting phenobarbital therapy for seizures.*
(b) Trigger=*after(*CAUSE*)*
Subject=*phenobarbital(*CHEMICAL*)*, Object=*dyskinesia(*DISEASE*)*

For the CID task, in the second phase, we simply keep the PASs in which the trigger is causal, logical subject argument corresponds to a chemical term and the object to a disease, and prune the rest of the PASs (for instance, those where both the subject and object are diseases). Adjunct arguments of the relevant PASs, if any, are also pruned. Since the CID relations are document-level relationships, we also prune duplicate instances of the same relationship.

2.2 Augmenting rule-based extraction with supervised learning

Causal relationships can be notoriously difficult to extract, since they can be expressed implicitly, not only at the sentence level but also at the discourse level and may require inference [9]. Therefore, we augmented the rule-based extraction with two supervised models. The first of these models (ML_SENTENCE) addresses sentence-bound relationships and the second addresses discourse-level relationships (ML_DISCOURSE). No explicit triggers are required for the relationships extracted with these models. We formulate the implicit relation extraction task as a binary classification task, where training examples consist of chemical-disease mention pairs and we predict whether a relationship holds between them or not. For ML_SENTENCE, training examples consisted only of mention pairs that appear in the same sentence and, for ML_DISCOURSE, of those that do not appear in the same sentence in a given abstract. We used linear SVM [1] to train the classifiers. Both classifiers use n-gram features as well as trigger-related and discourse-level features, provided in Table 2 and illustrated on the *phenobarbital:dyskinesia* mention pair from the sentence in Example (1a). Feature extraction presupposes the pre-processing performed for the rule-based extraction

Feature	**Description**	**Sentence**	**Discourse**
F_1	Uncased tokens of the mention sentence(s)	Y	Y
F_2	Uncased bigrams of the mention sentence(s)	Y	Y
F_3	5 uncased tokens preceding the chemical mention	Y	Y
	{*starting, after, soon, dyskinesia, a*}		
F_4	5 uncased bigrams preceding the chemical mention	Y	Y
	{*after_starting,soon_after,dyskinesia_soon,a_dyskinesia,developed_a*}		
F_5	5 uncased tokens following the chemical mention	Y	Y
	{*therapy,for,seizures*}		
F_6	5 uncased bigrams following the chemical mention	Y	Y
	{*therapy_for,for_seizures*}		
$F_7 - F_{10}$	same as $F_3 - F_6$, for the disease mention	Y	Y
F_{11}	Uncased causal triggers preceding the chemical mention (*after*)	Y	Y
F_{12}	Uncased causal triggers following the chemical mention $(\emptyset)$	Y	Y
$F_{13}-F_{14}$	same as $F_{11} - F_{12}$, for the disease mention	Y	Y
F_{15}	chemical in focus (*true*)	Y	Y
F_{16}	disease in focus (*true*)	Y	Y
F_{17}	number of textual units between mentions (*3*)	Y	N
F_{18}	normalized section name of the mention (*ABSTRACT*)	Y	N
F_{19}	whether the other concept in the mention pair exists in the sentence	N	Y
F_{20}	whether the other semantic type in the mention pair exists in the sentence	N	Y
F_{21}	distance between the sentences of the mentions	N	Y

Table 2. The features using by the linear SVM models

component (i.e., mention recognition and coordination detection). Features F_{15} and F_{16} check whether the mentions are in the focus of the article. For this, we assume that if the concept associated with the mention occurs in the title of the article, it is in focus. F_{18} takes the value of the normalized section name (e.g. *OBJECTIVE, RESULTS*) if the abstract is structured [11] or *TITLE*/*ABSTRACT* depending on the mention's position, otherwise.

If a pair A-B is predicted to be related by a classifier and one or both of the mentions in the pair (for instance, A) are coordinated with other terms of the same type (C), we add new relationships to our predictions (in this case, C-B is added). In one of our official runs (Run 1), the sentence-based classification is only performed if the number of relationships extracted at the rule-based extraction step is fewer than two. Similarly, discourse-based classification is performed only when the number of relationships extracted in previous steps is fewer than two. In our other run (Run 2), we only used the machine learning-based components. In this run, the discourse-based classification was performed only when the sentence-based classifier generated fewer than two relationships.

3 Results and Discussion

The official results for our two runs on the test dataset are provided in Table 3. Using rule-based extraction yields better performance at the expense of slightly longer response time. It also provides a 35% increase in F_1 score from the co-occurrence based baseline method, which also uses DNorm and tmChem.

Run	Response Time	Precision	Recall	F_1
Run 1	4,538.4	42.47	31.99	36.49
Run 2	4,471.4	39.30	31.71	35.10
Baseline		16.43	76.45	27.05

Table 3. Official evaluation results

To assess the effect of different components of the system on performance, we also evaluated our system on the development dataset (Table 4). The runs that used gold standard mention/concept annotations are indicated with (A) and those using DNorm and tmChem mention/concept annotations are indicated with (B). Using DNorm and tmChem yielded a precision of 72.59, a recall of 69.74, and an F_1 score of 71.14 for concept recognition on the development set.

Run	Precision	Recall	F_1
Run 1 (A)	54.64	59.76	57.09
Run 1 (B)	39.04	32.15	35.26
Rule-based (A)	60.94	37.08	46.11
Sentence-based (A)	54.13	54.34	54.23
No number restriction (A)	43.43	63.60	51.62

Table 4. Results on the development set

These results show that our relation extraction performance was influenced significantly by chemical/disease concept recognition.The system performance in F_1 score is reduced from 57.09 to 35.26 (approx. 38%), when it takes as concepts those extracted by DNorm and tmChem. The results also confirm the difficulty of extracting causal relationships, even when the entities involved are known. Trigger-centric rule-based approach achieves reasonable precision (60.94); however, it yields low recall (37.08), suggesting that most causal relations are expressed with more complex means than typical verb- or nominalization-anchored constructions. Augmenting the rule-based extraction with entence-based classification (ML_SENTENCE) is helpful in addressing the low recall to some extent (47% increase) at the expense of some precision loss (11%). One of the interesting aspects of the CID task is the prevalence of discourse-level relationships; we found that about 15% of all relationships were expressed only at this level. The discourse-level classifier improved F_1 score from 54.23 to 57.09 (approx. 5% increase), without hurting precision or recall. Overall, SVM models improved the F_1 score by about 24% (46.11 to 57.09). Restricting when the models are used had a significant effect on the results, since they tended to overgenerate relationships. Without any restriction on the number of relationships generated at the previous steps, the overall performance drops by about 10%, to 51.62, even though the recall is highest at 63.6.

4 Conclusion

We presented a hybrid approach for chemical-disease relationship extraction task. The rule-based extraction component yields reasonable precision with low recall. Augmenting this component with sentence-based and discourse-based relationship classifiers improves overall results significantly. We used relatively simple features for classifiers; more sophisticated syntactic and semantic features did not seem to provide much benefit. We also experimented with coreference resolution and relationship pruning based on negation/speculation detection; however, we did not observe a positive effect due to these components.We believe that focusing on better named entity recognition/normalization and discourse-level understanding is likely to be fruitful for improving the state-of-the-art in chemical-disease relationship extraction task.

Acknowledgments. This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

References

1. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874 (2008)
2. Kilicoglu, H.: Embedding Predications. Ph.D. thesis, Concordia University (2012)
3. Kilicoglu, H., Bergler, S.: Biological Event Composition. BMC Bioinformatics 13 (Suppl 11), S7 (2012)
4. Kim, J.D., Nguyen, N., Wang, Y., Tsujii, J., Takagi, T., Yonezawa, A.: The Genia Event and Protein Coreference tasks of the BioNLP Shared Task 2011. BMC Bioinformatics 13 Suppl 11, S1 (2012)
5. Leaman, R., Dogan, R.I., Lu, Z.: DNorm: disease name normalization with pairwise learning to rank. Bioinformatics 29(22), 2909–2917 (2013)
6. Leaman, R., Wei, C.H., Lu, Z.: tmChem: a high performance approach for chemical named entity recognition and normalization. Journal of Cheminformatics 7(S-1), S3 (2015)
7. Li, J., Sun, Y., Johnson, R., et al.: Annotating chemicals, diseases, and their interactions in biomedical literature. In: Proceedings of the fifth BioCreative challenge evaluation workshop (2015)
8. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 55–60 (2014)
9. Mihaila, C., Ohta, T., Pyysalo, S., Ananiadou, S.: BioCause: Annotating and analysing causality in the biomedical domain. BMC Bioinformatics 14(1), 2+ (2013)
10. Prasad, R., Dinesh, N., Lee, A., Miltsataki, E., Robaldo, L., Joshi, A., Webber, B.: The Penn Discourse TreeBank 2.0. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008) (2008)
11. Ripple, A.M., Mork, J.G., Knecht, L.S., Humphreys, B.L.: A retrospective cohort study of structured abstracts in MEDLINE, 1992-2006. Journal of the Medical Library Association : JMLA 99(2), 160–163 (2011)
12. Wei, C.H., Peng, Y., Leaman, R., et al.: Overview of the BioCreative V Chemical Disease Relation (CDR) Task. In: Proceedings of the fifth BioCreative challenge evaluation workshop (2015)