RELigator: Chemical-disease relation extraction using prior knowledge and textual information

E. Pons*, B.F.H. Becker*, S.A. Akhondi, Z. Afzal, E.M. van Mulligen, J.A. Kors

Dept. of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

{e.pons, b.becker, s.ahmadakhondi, m.afzal, e.vanmulligen, j.kors}@erasmusmc.nl

Abstract. The Erasmus MC team participated in the chemical-disease relation (CDR) task in the BioCreative V challenge. The CDR task consists of two subtask: automatic disease named entity recognition and normalization (DNER) and extraction of chemical-induced diseases (CID) from Medline abstracts. For the DNER subtask, we used our concept recognition tool Peregrine, in combination with several optimization steps. For the CID subtask, our system – RELigator – was trained on a rich feature set, including features derived from a graph database containing prior knowledge about chemicals and diseases, and linguistic and statistical features derived from the training corpus abstracts. We describe the systems that we developed and used, provide evaluation results for both CDR subtasks on the reference set, and compare the performance of our systems with baseline systems provided by the challenge organizers.

Keywords. Chemical-induced disease; Named entity recognition; Normalization; Relation mining; Knowledge graph; Natural Language Processing; Machine learning

1 Introduction

The extraction of chemicals, diseases, and their relationships from unstructured scientific publications is relevant for many areas of biomedical research, e.g., pharmacovigilance and drug discovery. The manual extraction of these entities and relations, and their storage in structured databases is cumbersome and expensive, and it is impossible for researchers or curators to keep pace with the growing number of papers that are being published. Automatic extraction of chemical-disease relations (CDRs) would solve these problems, but previous attempts have met limited success. Among the difficulties that have to be addressed are the identification of relevant concepts, i.e., chemicals and diseases. Concept identification goes beyond concept recognition in that not only the mention of a chemical or disease has to be recognized, but also a unique identifier has to be assigned, which links it to a source that contains further information about the concept. Also the detection of relationships between the identified chemicals and diseases remains challenging.

In BioCreative V, one of the challenge tasks is the automatic extraction of CDRs from biomedical literature [1]. The CDR task consists of two subtasks. The first subtask involves automatic disease named entity recognition and normalization (DNER)

* These authors contributed equally.

from a set of Medline documents, and can be considered as a first step in CDR extraction. The second subtask consists of extracting chemical-induced diseases (CID) and providing the chemical-disease pairs per document.

The Erasmus MC team participated in both CDR subtasks. For the DNER subtask, we used our concept recognition tool Peregrine [2], in combination with several optimization steps. For the CID subtask, we applied the optimized Peregrine system for disease concept recognition; for chemical concept recognition, we used tmChem [3], a chemical concept recognizer that was provided by the challenge organizers. A relation extraction module was trained on a rich feature set, including features derived from a graph database containing prior knowledge about chemicals and diseases, and linguistic and statistical features derived from the training corpus documents.

In the following, we describe the systems that we developed and used, provide evaluation results for both CDR subtasks on the training and development corpus, and compare the performance of our systems with baseline systems provided by the challenge organizers.

2 Methods

Training data

The CDR task data consisted of a training set and a development set, each containing 500 Medline documents, each consisting of a title and abstract text. Chemicals and diseases in the documents were annotated in the form of text offset, text span, and MeSH identifier. Chemical-disease interactions were annotated at the abstract level, but only if the abstract provided evidence for a mechanistic relationship between a chemical and disease. Therapeutic relationships between chemicals and diseases were not annotated.

Recognition and normalization of chemicals

The chemical concept recognition was performed using the tmChem chemical recognizer system [3]. The tmChem system was the best performing system in the previous BioCreative chemical named entity recognition (CHEMDNER) challenge [4], and also includes a dictionary look-up to map recognized chemicals to MeSH identifiers. tmChem is an ensemble system that combines two CRF-based systems, of which we only used the one that performed best in the CHEMDNER challenge. We trained this system on the 1000 documents in the CDR training data.

Recognition and normalization of diseases

For the recognition of diseases, we constructed a thesaurus with concepts and corresponding terms taken from the vocabularies MeSH, MedDRA, Snomed-CT, and ICD10-CM, as contained in the Unified Medical Language System (UMLS) 2015AA edition. We restricted the terms to those that are flagged as non-suppressible and belong to one the semantic group “Disorders” [5].

All documents were processed by our concept recognizer Peregrine [2]. Moreover, we extracted all abbreviations and their corresponding long forms [6], and made sure

that any combination of abbreviation and long form was tagged with the same concept. If two recognized terms were adjacent and identified as the same concept, they were merged. If a recognized term was followed by the word 'syndrome', the term was expanded to include 'syndrome'. Finally, a list of suppress concepts was used to filter out concepts that should not have been tagged. This list was composed of concepts found by Peregrine but never annotated in the training set.

The resulting list of concepts with UMLS identifiers was mapped to MeSH identifiers with the IntraMap tool [7] developed by the National Library of Medicine. IntraMap consists of a precompiled mapping table containing the semantically closest MeSH header for each UMLS concept.

Relation extraction: problem definition

We formulated the relation extraction task as a binary decision problem on all possible pairs of chemicals and diseases found in each document. For the development of the relation extraction algorithm we distinguished between perfect entity annotations from the reference standard and imperfect entity annotations produced by Peregrine and tmChem.

Given perfect entity annotations, 10639 chemical-disease pairs were constructed as training instances. Co-occurrence pairs were allowed to cross the title-abstract border. Each co-occurrence was considered an instance and was labeled positive (n=2049) or negative (n=8644) depending on whether the chemical-disease relation had been annotated. For each instance, three types of features were generated, based on prior knowledge, and on statistical and linguistic information from the document.

Knowledge-based features

To generate features based on existing, prior knowledge, we used a graph database, BRAIN [8], developed by Euretos [9]. This graph database contains entities and relations from (curated) structured databases and texts (Medline) for almost every concept in the UMLS. Each connection between entities can have a set of named relations or predicates. Attached to each connection is the amount of provenance and the source of the provenance. Different sources have been assigned different weights. BRAIN provides an application programming interface that can be used to query for paths between two given concepts. A path can be direct (i.e., the concepts have a direct relationship) or indirect (the concepts are connected through an intermediate concept). For each path, a confidence score is computed that indicates how strongly the concepts are connected (based on the variety of sources and the number of references to texts and database records). We fed each chemical-disease pair to BRAIN and determined whether there was a direct or indirect path, the confidence score, the list of different predicates with their amount of provenance, and the total number of unique predicates. Provenance counts for absent predicates were treated as missing values.

Statistical features

The statistical feature set contained the number of occurrences of the chemical, the disease and the pair in the document, as well as their ratios to the numbers of occurrences of all chemicals, diseases and chemical-disease pairs in the document. Two features captured the minimal sentence distance between the chemical and the disease, and the minimal word distance. Binary features indicated whether the chemical, the disease, or both were mentioned in the document title.

NLP features

We used the Stanford CoreNLP parser to generate dependency trees of the sentences of each document. The semantic role of the concepts was assumed to be reflected by the "governing verb" in the parse tree. We defined the governing verb of a word as the first verb that is encountered when ascending the parse tree from the word towards the root.

Two sets of NLP features were derived. For the first set, the closest pair of occurrences of the chemical and the disease in the document was considered. The features consisted of the governing verbs of the two words, the word that relates the chemical and disease if they co-occur in a sentence, and the governing verb of the relating word. An additional feature described whether the chemical was mentioned before the disease, and if another chemical-disease pair was found lower in the parse tree. The second set of NLP features aggregated information about the governing verbs of all possible chemical-disease co-occurrence pairs in a document. This set contained one numeric feature for each governing verb encountered in the reference set. Each feature indicated how many times that word was found as governing word of the chemical and disease.

Machine learning

A total of 1454 features were generated for all instances. Various machine learning algorithms were explored, utilizing Weka machine learning libraries [10]. Performance was estimated by ten-fold cross-validation.

In a preliminary analysis in which we compared various classification algorithms, support vector machines (SVM) proved to have superior performance. Therefore we continued to optimize parameters for the SVM classification model. We used C-SVC classification with radial basis function kernel type, and initially with default settings for cost (1.0) and gamma (0.0).

All numeric features were normalized to scale between zero and one. Because of the class imbalance the cost matrix of the SVM was set to 5:1, giving extra weight to the minority class. Utilizing the best performing feature set, we tuned the cost and gamma parameters by performing a grid search, again using cross-validation. During the grid search, we used a fixed decision threshold of 0.5 for the SVM. We subsequently varied the decision threshold to optimize the F-score of the SVM.

3 Results

DNER task

The performance of different systems in the DNER task is shown in table 1. The challenge baseline system (i.e., dictionary look-up using names from CTD) resulted in an F-score of 0.523. Our Peregrine-based system obtained an F-score of 0.788 for the recognition of disease and an F-score of 0.794 for the normalization on the training data. On the final DNER test set, the system reached an F-score of 0.757.

Table 1. Performance of the challenge baseline system and Peregrine for disease entity recognition and normalization.

System	Data set	Precision	Recall	F-score
DNER baseline - Norm.	Test	0.427	0.674	0.523
Peregrine - Recognition	Training + development	0.833	0.748	0.788
Peregrine - Normalization	Training + development	0.829	0.762	0.794
Peregrine - Normalization	Test	0.737	0.772	0.757

CID task

Table 2 shows the performance results of different relation extraction systems on the CDR training and development data, using the gold-standard chemical and disease annotations to generate all possible chemical-disease pairs. A baseline system based on sentence co-occurrence of entities gave an F-score of 0.437 with a recall of 0.725, indicating that more than a quarter of the relations spanned more than one sentence. The use of prior knowledge, assuming that a relation was present if a chemical and disease was directly connected in BRAIN by a non-treatment predicate, resulted in an F-score of 0.503. Further improvements were gained by training an SVM with different feature sets. The best F-score of 0.692 was obtained when all features were used. This model was further improved by a grid search for optimal cost and gamma parameters. Best performance (F-score 0.760) was obtained for a cost of 2.2 and a gamma of 0.10, in combination with a decision threshold of 0.30. This system, named RELigator, was used for one of our submission runs; the other two runs used the same model but with thresholds of 0.20 and 0.40.

Table 2. Relation extraction performance for different systems on the CDR training data, given perfect entity annotations.

System	Threshold	Recall	Precision	F-score
Co-occurrence (sentence)	n/a	0.725	0.313	0.437
Prior knowledge (direct path, non-treat predicate)	n/a	0.664	0.405	0.503
SVM, prior knowledge features	0.36	0.903	0.481	0.628
SVM, statistical + NLP features	0.37	0.677	0.648	0.663
SVM, all features	0.28	0.833	0.592	0.692
SVM, all features, optimized	0.30	0.840	0.693	**0.760**

Table 3 shows the performance results of RELigator, using Peregrine and tmChem for entity normalization, on the final CDR test set. Our best submission had an F-score of 0.539. This compares favorably with the F-score of 0.271 of the challenge baseline system, which used DNorm [11], tmChem, and entity co-occurrence.

Table 3. Relation extraction performance of the challenge baseline system and the RELigator system for three decision thresholds, on the CDR test set.

System	Threshold	Recall	Precision	F-score
Challenge baseline	n/a	0.765	0.164	0.271
RELigator - run 1	0.2	0.513	0.539	**0.526**
RELigator - run 2	0.3	0.559	0.478	0.515
RELigator - run 3	0.4	0.589	0.409	0.483

4 Discussion

We described our Peregrine-based system for disease normalization, and the RELigator system for chemical-disease relation extraction. On the final CDR test sets, both systems were shown to clearly outperform the baseline systems made available by the challenge organizers.

The recall of our disease normalization system may be further improved by applying rewrite rules to UMLS terms [12]. Precision could be increased by the use of part-of-speech and chunking information to remove erroneously recognized terms [13]. Finally, we noticed that the mapping of UMLS identifiers to MeSH was not always correct; a detailed error analysis might suggest some improvements in the mapping.

Regarding the chemical-disease relation extraction, our results indicate that knowledge-based features and text-based features both contributed to the final system performance, and thus contain at least partly complementary information. Further improvement of the knowledge-based features may be possible by considering paths between a chemical and disease that span more than one intermediate concept. Also different confidence scoring schemes or grouping of predicates may yield more powerful features.

Finally, the chemical-disease relation annotations that we used to train our models were provided at the document level. We did not attempt to annotate the relation mentions in the document texts, which might have yielded stronger features. Such an exercise is left for future research.

Acknowledgment

We gratefully acknowledge Euretos for providing the BRAIN system.

References

[1] Wei CH, 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; Sevilla, Spain, 2015.

[2] Schuemie MJ, Jelier R, Kors JA. Peregrine: lightweight gene name normalization by dictionary lookup. In: Proceedings of the BioCreAtIvE II Workshop; Madrid, Spain, 2007. p. 131-3.

[3] Leaman R, Wei CH, Lu Z. tmChem: a high performance approach for chemical named entity recognition and normalization. J Cheminform 2015, 7:S3.

[4] Krallinger M, Leitner F, Rabal O, et al. CHEMDNER: The drugs and chemical names extraction challenge. J Cheminform 2015, 7:S1.

[5] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Medinfo 2001, 10:216-20.

[6] Schwartz AS, Hearst MA. A simple algorithm for identifying abbreviation definitions in biomedical text. Pac Symp Biocomp 2003, 4:451-62.

[7] Fung KW, Bodenreider O, Aronson A, et al. Combining lexical and semantic methods of inter-terminology mapping using the UMLS. Stud Health Technol Inform 2007, 129:605.

[8] Bio-IT World. Big BRAIN: Finding Connections in the Literature Flood with Euretos BRAIN[Internet]. Available from: http://www.bio-itworld.com/2014/7/1/big-brain-finding-gems-literature-flood-euretos-brain.html

[9] Euretos[Internet]. Available from: http://www.euretos.com.

[10] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[11] Li J, Sun Y, Johnson RJ, Sciaky D, et al., (2015) Annotating chemicals, diseases, and their interactions in biomedical literature, in Proceedings of the fifth BioCreative challenge evaluation workshop, Sevilla, Spain

[12] Hettne KM, van Mulligen E, Schuemie MJ, et al. Rewriting and suppressing UMLS terms for improved biomedical term identification. J Biomed Semantics 2010, 1:5.

[13] Kang N, Singh B, Afzal Z, et al. Using rule-based natural language processing to improve disease normalization in biomedical text. J Am Med Inform Assoc 2013, 20:876-81.