Ontogene Term and Relation Recognition for CDR

Tilia Renate Ellendorff, Simon Clematide, Adrian van der Lek, Lenz Furrer, and Fabio Rinaldi

Institut für Computerlinguistik, Universität Zürich,
Binzmühlestrasse 14, 8050 Zürich
{ellendorff,clematide,vanderlek,
lenz.furrer,fabio.rinaldi}@uzh.ch
http://www.ontogene.org/

Abstract. For our participation in the CDR task of BioCreative 5, we have adapted the Ontogene System and optimized it for disease recognition (DNER Task) and identification of chemical-disease relationships (CID Task). For the DNER Task we have experimented with different changes to the term matching system. We describe the effects of an abbreviation detection tool as well as a selection of rules for term normalization.

Key words: Disease Named Entity Recognition, Biomedical Textmining, Toxicogenomics

1 Introduction

The CDR task of BioCreative 2015 was set up to promote the extraction of disease terms and the relations between chemical and disease entities from the biomedical literature[1]. The automatic identification of diseases and their relations to chemicals in the text of research article has the aim of replacing the costly and time consuming process of manual curation. Furthermore, it becomes the only way of keeping up with the massive increase of publications in the domain. Being able to extract such information from biomedical text helps to keep biomedical databases, such as the Comparative Toxicogenomics Database [2] up to date more efficiently and therefore is also beneficial for the acceleration of research in the domain of the life sciences. The Ontogene group participated in both sub-task of the CDR challenge.

As the DNER sub-task (Disease Named Entity Recogntion and Normalization) did not only include named entity recognition but also the grounding to database identifiers, we applied a dictionary look-up. We experimented with several settings of the dictionary, application of transformation rules and an abbreviation detection tool.

For the CID sub-task (chemical-induced disease relation extraction) we applied and extended a machine learning approach which is an adaptation of a system which we originally developed for discovering interactions of biomedical

entities, in particular protein-protein interactions [4], and was later applied to the PharmGKB and CTD databases. It was also used for participation in the CTD challenge of BioCreative 2013 [3].

Our approaches to the two sub-tasks are explained more in detail below.

2 DNER Task: Disease Named Entity Recognition and Normalization

For the DNER task we applied a dictionary look up in order to be able to identify the terms in the text as disease mentions and to ground these mentions to their respective identifiers. The disease terms were extracted from the official Medical Subject Headings (MeSH) database and transferred into an internal dictionary format designed to facilitate term look-up. We experimented with several approaches for making the dictionary look-up more efficient, which are described in the following subsections. We used the training set for dictionary development and error analysis and the development set for evaluation of the termmatcher.

2.1 Generation of Disease Term Dictionary

MeSH (Medical Subject Headings) contains a controlled vocabulary of keywords used to annotate PubMed abstracts and NLM's book database. This has the aim of providing the topics of a text and therefore facilitating the search of specific articles. Subject headings are available in the format of hierachically sorted descriptors and qualifiers. Each subject heading is connected to a tree number which defines its position in the hierarchy. Tree numbers provide information concerning the entity type which a term refers to. Branches on the highest level are referenced by letters with numbers referencing sub-branches and leaves. For building the disease dictionary we considered the branches C (diseases) and F03 (mental disorders), which is a branch of F (psychiatry and psychology). We considered these branches recursively, taking all sub-branches and leaves into account. For each term we also extracted all synonyms and transferred them into the internal dictionary format. The resulting dictionary contained disease terms for 2813 different MeSH identifiers. As a comparison, the CDR training data contains disease annotations for 665 different MeSH disease identifiers. The MeSH identifiers contained in our dictionary cover all MeSH identifiers encountered in the training data, except for the identifier -1 which, according to the BioCreative V CDR Task Data Annotation Guidelines document[1], is used by the annotators when a disease cannot be normalized.

In order to enhance the dictionary further for evaluation on the test set, we also included terms from training and development set.

[1] http://www.biocreative.org/media/store/files/2015/bc5_CDR_data_guidelines.pdf

2.2 Normalization

We applied basic normalization to every run: all terms were lower-cased and whitespace was removed. Additionally, we experimented with a selection of simple term transformations which had the aim of decreasing false positives.

During error analysis, we noticed that entity mentions in the text have not been found in the dictionary because the last part of a multi-word is different from what is recorded in the dictionary, as can be seen in the examples below.

Examples: (dictionary entry on the left, text variant on the right)
`respiratory arrest` vs. `respiratory depression`
(MESH:D012131; PubMed ID:10457883)
`neuromuscular blockage` vs. `neuromuscular manifestation`
(MESH:D020879 ; PubMed ID:10457883)

We tried to decrease similar errors by experimentally applying a list of simple transformation rules to the most common last parts of multi-word terms: *symptom(s)*, *sign(s)*, *toxicity*, *injury*, *lesion*, *dysfunction* and *insufficiency*. These were all mapped to the generic term *disorder* Furthermore we normalized *increase in* to *high*', *decrease in* to *low*, *renal* to *kidney* and *myocardial* to *heart* in order to deal with other errors observed during error analysis.

2.3 Abbreviation Detection

We developed a simple tool for the detection of abbreviated term variants of the same type as in the example below. After the termmatcher has finished matching the terms from the dictionary in the text of an abstract, the tool checks for the occurrence of a term which is followed by an item in brackets. If this is the case, the tool records the item and subsequently annotates it as a term. The corresponding MeSH identifier and term information is "inherited" from the term in front of the brackets.

Example:
`243 cocaine-dependent outpatients with` **`cocaine-induced mood disorder (CIMD),`** `other mood disorders, or no mood disorder were compared on measures of psychiatric symptoms.`
(PubMed ID 10365197)

3 CID: Chemical-induced disease relation extraction

As an extension to the approach that we already applied in BioCreative 2013 [3], we experimented with stem features in order to introduce contextual linguistic features that should add additional evidence for estimating the probability of concepts entering a relevant relation.

In order to be included as a stem feature for a certain concept, a stem had to co-occur at least two times within a window of one sentence with the concept that has been identified by the term recognizer. Function words and stems with less than 2 characters were generally excluded.

Over the development set, we measured a small but consistent improvement. Unfortunately, this improvement could not be transferred to the test set. There, our old model (retrained with a more recent set of CTD data) performed slightly better in run 1 than the model with the new features for co-occurring stems, used in run 2. It is difficult to explain why the improvement could not be seen on the test data. Run 3 used stem features and additionally retrieved meta-information (MeSH terms and chemical substances) from PubMed.

Our results for relation identification in terms of F-measure suffer from the fact that recall and precision remained unbalanced. Our simple selection criterion of using the best 5 relation candidates according to our relation score should have been replaced by a more selective criterion, for instance, by considering the distance between the best candidate and the lower-ranked ones, or by looking at the decay of the relation score between two successive ranks.

4 Results

All results were obtained using the official CDR task evaluation kit provided by the organizers. The baseline methods are described in [5].

4.1 Results on Development Set

Table 1. Results on Development Set

Run Setup	ID Evaluation			Mention Evaluation		
	P	R	F-Score	P	R	F-Score
MeSH Terms only	0.832	0.630	0.717	0.857	0.589	0.698
Abbreviation Detection	Same as above			0.853	0.640	0.730
Normalization	0.681	0.659	0.670	0.807	0.622	0.703
MeSH Terms and Training Terms	0.804	0.702	0.749	0.836	0.652	0.733
Development Terms	0.943	0.863	0.901	0.960	0.847	0.900

Table 1 illustrates the effects of normalization and abbreviation detection. In all these runs we included a list of stopwords[2] which we extended with the terms listed as 'general term' which should not be annotated in the CDR guidelines for disease mention and concept id annotation[3].

Abbreviation detection only slightly harms precision values but show a positive impact on recall, which also improves overall F-score (Table 1: Abbreviation Detection). As only abbreviations are detected if a corresponding terms

[2] ftp://ftp.ncbi.nih.gov/gene/DATA/stopwords_gene

[3] http://www.biocreative.org/media/store/files/2015/bc5_CDR_data_guidelines.pdf

has already been discovered in the abstract, abbreviation detection only changes mention evaluation scores but has no effect on ID evaluation.

Normalization detection shows a negative effect during ID evaluation, where the gain in recall does not justify the loss in precision (Table 1: Normalization). On the other hand, a slight positive effect during mention evaluation can be observed. However, generally speaking, the introduced transformation rules have not contributed much, and most likely need some refinement. Errors introduced by normalization detection are typically due to mapping to wrong identifiers. For instance in the abstract with PubMed id 15517007, *cardiac toxicity* is wrongly mapped to MeSH identifier D006331, which belongs to the much more general concept of *heart disease*. These cases have a negative impact on id evaluation, even though disease mentions have been correctly identified in the text.

To reproduce a run similar to the one submitted in official evaluation on the test set, where terms from training and development set were included into the dictionary as well, we expanded the dictionary generated from MeSH terms by also including terms from the training set (Table 1: MeSH Terms and Training Terms). In both ID evaluation and mention evaluation, the expanded dictionary clearly improved recall but harmed precision. On the level of ID evaluation, overall F-score slightly decreases, whereas it increases mention evaluation. Decrease in precision is due to abbreviations and composites that are annotated in the training data but that do not qualify as real term variants.

As a reference we included a run using a dictionary only containing the terms present in the development set (Table 1: Development Terms). Evaluating this run on the development set gives us an upper bound concerning the performance of the termmatcher. The observation that the evaluation values do not reach 100% can be mainly explained by the fact that the termmatcher so far is not able to detect composites of term mentions.

4.2 Official Results on Test Set

Table 2. Official Results on Test Set

Task	Run	Time (ms)	TP	FP	FN	P	R	F
DNER	1	4324	1692	604	296	73.69	85.11	78.99
	2	3470	1703	606	285	73.75	85.66	79.26
	baseline					42.71	67.46	52.30
CID	1	4575	595	1835	471	24.49	55.82	34.04
	2	4685	590	1840	476	24.28	55.35	33.75
	3	4929	596	1875	470	24.12	55.91	33.70
	baseline					16.43	76.45	27.05

Table 2 shows our official results on the test set. We omit DNER run 3 since, due to a configuration error, it turned out be identical to run 2. The main difference between run 1 and run 2 is that the former used the full pipeline, including the ME-based filtering of relations, and attempted to use the ranking provided by the ME method to filter the best concepts. Instead run 2 omitted this step and generated results directly after term matching without using ME-based filtering. Correspondingly, run 2 (and run 3) are faster than run 1. To our surprise, results are slightly worse in run 1, pointing to a potential problem in the way the ME filtering approach was used.

As for the CID results, as described before, the 3 runs were distinguished by the different type of ME model used. Run 1 used our previous approach using mainly features based on tokens, while runs 2 and 3 used models including stem features. Disappointingly, the results in this case were slightly worse.

References

1. Li, J., Sun, Y., Johnson, R., Sciaky, D., Wei, C., Leaman, Davis, A., Mattingly, C., Wiegers, T., Lu, Z.: Annotating chemicals, diseases, and their interactions in biomedical literature. In: Proceedings of the fifth BioCreative challenge evaluation workshop. Sevilla, Spain (2015)
2. Mattingly, C., Rosenstein, M., Colby, G., Forrest Jr, J., Boyer, J.: The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. Journal of Experimental Zoology Part A: Comparative Experimental Biology 305A(9), 689–692 (2006), http://dx.doi.org/10.1002/jez.a.307
3. Rinaldi, F., Clematide, S., Ellendorff, T.R., Marques, H.: OntoGene: CTD entity and action term recognition. In: Proceedings of the Fourth BioCreative Challenge Evaluation Workshop. vol. 1, pp. 90–94 (2013)
4. Rinaldi, F., Schneider, G., Kaljurand, K., Clematide, S., Vachon, T., Romacker, M.: OntoGene in BioCreative II.5. IEEE/ACM Transactions on Computational Biology and Bioinformatics 7(3), 472–480 (2010)
5. Wei, C., Peng, Y., Leaman, R.: Overview of the biocreative v chemical disease relation (cdr) task. In: Proceedings of the fifth BioCreative challenge evaluation workshop, Sevilla, Spain (2015)