Chemical compound and drug name recognition using CRFs and semantic similarity based on ChEBI

Andre Lamurias, Tiago Grego, and Francisco M. Couto

Dep. de Informática, Faculdade de Ciências, Universidade de Lisboa, Portugal
alamurias@lasige.di.fc.ul.pt, tgrego@fc.ul.pt, fcouto@di.fc.ul.pt
http://www.lasige.di.fc.ul.pt

Abstract. This document presents our approach to the BioCreative IV challenge of recognition and classification of drug names (CHEMDNER task). We developed a system based on Conditional Random Fields for recognizing chemical terms, and on ChEBI resolution and semantic similarity techniques for validating the recognition results. Our system created multiple classifiers according to different training datasets that we built. Each of the classifiers returned a confidence score that were combined to filter and rank the results. F-measure, precision and recall were estimated by using cross-validation on the CHEMDNER training and development data for each method used. The best f-measure, precision and recall estimated for the CEM subtask was 0.79, 0.95 and 0.76, respectively. Excluding results with low semantic similarity values enabled us to achieve higher levels of precision.

Key words: Ontologies; Semantic Similarity; Conditial Random Fields; Random Forests; ChEBI

1 Introduction

This paper presents our approach to the Task 2 of BioCreative IV (CHEMDNER task). Our approach was based in our system [2], used for the SemEval 2013 challenge [7], task 9.1, concerning chemical compound and drug name recognition in biomedical literature. For BioCreative IV, our system was improved by using the confidence score of each classifier for each prediction and by using the maximum semantic similarity value to other compounds in the same fragment of text. This value was obtained by performing ChEBI resolution on each chemical entity recognized and calculating the Gentleman's simUI semantic similarity measure [8] for each pair of successfully mapped entities. We used these values, along with the ChEBI mapping score, as features for a Random Forests model, to further improve our predictions.

Our team participated on both subtasks, deriving the predictions for the Chemical Document Indexing (CDI) subtask from our Chemical Entity Mention recognition (CEM) predictions. We submitted five runs for each subtask, using different methodologies and expecting different results for each one. With the CHEMDNER gold standard, we were also able to estimate the expected micro-averaged precision, recall and f-measure values for each methodology and subtask, by using cross-validation in the training and development sets.

2 Systems description and methods

2.1 Datasets used

In addition to the provided CHEMDNER dataset, for training our classifiers, we used the DDI corpus dataset provided for the Semeval 2013 challenge [1], and a patent document corpus released to the public by the ChEBI team [6]. The DDI dataset contains two sub-datasets. One that consists of MEDLINE abstracts, and other that contains DrugBank abstracts. All named chemical entities were labeled with their type which could be one of the following: Drug, Brand, Group and Drug_n. Based on the label, we created four datasets based on the DDI corpus dataset and seven datasets based on the CHEMDNER corpus, each containing only one specific type of annotated entities (Table 1).

Table 1. Number of documents and annotations available in each training corpus.

	Corpus	**Documents**	**Annotations**
Patents	Chemical	40	3717
DDI	Drug	593	9425
	Group	489	3399
	Brand	295	1437
	Drug_n	88	504
	All	714	14765
CHEMDNER	Abbreviation	1845	9059
	Family	2738	8313
	Formula	1739	8585
	Identifier	349	1311
	Multiple	258	390
	Systematic	3763	13472
	Trivial	3714	17802
	All	7000	59004

2.2 CRF entity recognition

For this competition, we adapted Mallet's implementation of Conditional Random Fields (CRFs) [10] to output the confidence score of each prediction. This information was useful to adjust the precision of our predictions, and to rank them according to how confident the system is about the extracted mention being correct.

We used the provided CHEMDNER corpus, the DDI corpus and the patents corpus for training multiple CRF classifiers, based on the different types of entities considered on each dataset (Table 1). Each title and abstract from the test set was classified with each one of these classifiers. In total, our system combined the results from fourteen classifiers.

2.3 ChEBI resolution

After having recognized the named chemical entities, our method performs their resolution to the ChEBI ontology. The resolution method takes as input the string identified as being a chemical compound name and returns the most relevant ChEBI identifier along with a mapping score.

To perform the search for the most likely ChEBI term for a given entity we employed an adaptation of FiGO, a lexical similarity method [5]. Our adaptation compares the constituent words in the input string with the constituent words of each ChEBI term, to which different weights have been assigned according to its frequency in the ontology vocabulary. A mapping score between 0 and 1 is provided with the mapping, which corresponds to a maximum value in the case of a ChEBI term that has the exact name as the input string.

Our resolution method was applied to the named chemical entities on the CHEMDNER training and development sets. We were able to find a ChEBI identifier for 69.2 % of these entities. The fraction of entities our method was unable to resolve for each type is shown in Table 2.

Table 2. Number of chemical entities from the CHEMDNER training and development set not mapped to ChEBI and total

Type	Systematic	Identifier	Formula	Trivial	Abbreviation	Family	Multiple
Unmapped	3382 (25.1%)	1156 (88.2%)	3972 (46.3%)	3622 (20.3%)	4181 (46.2%)	1690 (20.3%)	91 (23.3%)
Total	13472	1311	8585	17802	9059	8313	390

With the named chemical entities successfully mapped to a ChEBI identifier, we were able to calculate the Gentleman's simUI semantic similarity measure for each pair of entities on a text. This measure is a structural approach, which explores the directed acyclic graph (DAG) organization of ChEBI [9]. We then used the maximum semantic similarity value for each entity as a feature for filtering and ranking.

2.4 Filtering false positives with a Random Forests model

The output provided for each putative chemical named entity found is the classifier's confidence score, and the most similar putative chemical named entity mentioned on the same document through the maximum semantic similarity score. Using this information, along with the ChEBI mapping score, we were able to gather 29 features for each prediction. When a chemical entity mention is detected by at least one classifier, but not all, the confidence score for the classifiers that did not detect this mention was considered 0. These features were used to train a classifier able to filter false positives from our results, with minimal effect on the recall value. We used our predictions obtained by cross-validation on the training and development set to train different Weka [11] classifiers, using the different methods implemented by Weka. The method that returned better results was Random Forests, and so we used that classifier on our test set predictions.

2.5 Post-processing

A common English words list was used as an external resource in post-processing. If a recognized chemical entity was part of this list or one of the words on the list was part of the chemical entity, then we assumed that it was a recognition error and should be filtered out and not be considered a chemical entity. This list was tuned with the rules used on the annotations of the gold standard.

Some simple rules were also implemented in an effort to improve the quality of the annotations. For instance, if the recognized entity was found to be composed entirely by digits, then it should be filtered out because it is most certainly an annotation error.

With such naïve but efficient rules it was expected that the performance of entity recognition would improve.

2.6 Testing runs

Using different combinations of the developed methods, five runs were submitted for each subtask and are now described.

Run 1: With this run we used all available classifiers, whose results were used to build a Random Forests model to filter the predictions.

Run 2: With this run we used only the classifiers trained with the CHEMDNER corpus and filtered with a confidence score and ChEBI mapping score threshold of 0.8.

Run 3: With this run we used the results from all available classifiers, including those trained with the DDI and patents documents corpus.

Run 4: In this run we excluded the results obtained with the CHEMDNER corpus classifiers that had a semantic similarity measure lower than 0.6.

Run 5: This run is similar to run 4 but all classifiers were used.

Our predictions were ranked according to the confidence score obtained. On runs 2, 3, 4 and 5, this score was calculated for each entity by averaging the top 3 classifier scores, maximum semantic similarity value to other compounds in the same fragment of text and the ChEBI mapping score obtained for that entity mention. On run 1, we used the confidence given by Weka's Random Forests method for each prediction. Post-processing was applied to every run, except run 2.

With the results from each run, we were able to generate predictions for the CEM subtask, using every result, and for the CDI subtask, considering only unique entities for each document.

3 Discussion

The two largest corpus used on our system were originally released for two different competitions. The DDI corpus, released for SemEval 2013, task 9, is focused on drug and brand names. The objective of this task was to first recognize and classify drug names and then extract the interactions between these drugs. This corpus consisted of DrugBank and MEDLINE abstracts while the CHEMDNER corpus consists of only MEDLINE abstracts. The CHEMDNER task required only the recognition of all chemical entities on a text. Using only classifiers trained with the CHEMDNER corpus, we expect better precision, since the gold standard for the test set follows the same annotation criteria. We expect better recall when including classifiers trained with the DDI corpus since some of the drug names can be classified as Abbreviation, Trivial or Family types of CHEMDNER annotations.

The improvements made on our system since SemEval 2013 gave us more parameters to tune and filter our results. We were able to control which classifiers to use and define thresholds for the classifier confidence score, the ChEBI mapping score and the semantic similarity values. These parameters were combined according to f-measure, precision and recall estimates for the training data. The metrics for each set of predictions were calculated using the official evaluation script on the results of 3-fold cross-validation for the CHEMDNER training and development dataset (Table 3).

Table 3. Precision, Recall and F-measure estimates for each method used, obtained with cross-validation on the training and development dataset.

Run		P	R	F
1	CDI	0.84	0.72	0.79
	CEM	0.87	0.70	0.79
2	CDI	0.95	0.06	0.12
	CEM	0.95	0.07	0.11
3	CDI	0.52	0.80	0.63
	CEM	0.57	0.76	0.65
4	CDI	0.88	0.23	0.36
	CEM	0.90	0.21	0.34
5	CDI	0.88	0.23	0.36
	CEM	0.80	0.23	0.35

The method that returned the best f-measure value consisted in filtering the results of all classifiers with the Random Forests model described on section 2.4. With this filter, we were able to improve our f-measure estimate from 0.65 to 0.79, for the CEM subtask. Without the filter, we obtained our best recall (0.76), but with a drop in precision (0.57). Previously, we used a mapping score threshold of 0.8 to improve the precision of our results. This time, a confidence score threshold of 0.8 was also used, and the estimated precision for this method was 0.95.

To evaluate the importance of using semantic similarity values, we tested different thresholds for the results obtained with every classifier and just the CHEMDNER corpus classifiers. The threshold that returned the best f-measure was 0.6. This method could be further improved by improving the ChEBI resolution method and by using more efficient semantic similarity measures, for example by incorporating disjoint axioms information from ChEBI [12] or by calculating the shared information between all CHEBI terms recognized in a fragment of text [13].

Acknowledgments. The authors want to thank the Portuguese Fundação para a Ciência e Tecnologia through the financial support of the SPNet project (PTDC/EBB-EBI/113824/2009), the SOMER project (PTDC/EIA-EIA/119119/2010) and the PhD grant SFRH/BD/36015/2007 and through the LASIGE multiannual support.

References

1. Segura-Bedmar, I., Martínez, P. and Pablo-Sánchez, C.: Using a shallow linguistic kernel for drug-drug interaction extraction. Journal of Biomedical Informatics, 44(5): 789–804. (2006)
2. Grego, T., Pinto, F. and Couto, F. M.: Identifying Chemical Entities based on ChEBI. Software Demonstration at the International Conference on Biomedical Ontologies (ICBO) (2012).
3. Corbett, P. , Batchelor, C. and Teufel, S. Annotation of chemical named entities. Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, 57–64 (2007).
4. Grego, T., Pezik, P., Couto, F. M. and Rebholz-Schuhmann, D. Identification of Chemical Entities in Patent Documents. Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living, volume 5518 of Lecture Notes in Computer Science, 934–941 (2009).
5. Couto, F. M. , Coutinho, P. M. and Silva, M. J. Finding genomic ontology terms in text using evidence content. BMC Bioinformatics, 6 (Suppl 1), S21 (2005).
6. ChEBI Patents Annotations,`http://chebi.cvs.sourceforge.net/viewvc/chebi/chapati/patentsGoldStandard`
7. Grego T., Pinto, F., and Couto, F., LASIGE: using conditional random fields and chebi ontology, in 7th International Workshop on Semantic Evaluation (SemEval) (2013).
8. Gentleman, R. Visualizing and distances using GO. `http://www.bioconductor.org/docs/vignettes.html` (2005).

9. Couto, F., Sofia Pinto, H.: The Next Generation of Similarity Measures that fully explore the Semantics in Biomedical Ontologies. Journal of Bioinformatics and Computational Biology (2013).
10. McCallum., A. K. Mallet: A machine learning for language toolkit. `http://mallet.cs.umass.edu` (2002).
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H.: The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1 (2009).
12. Ferreira, J. D., Hastings, J., and Couto, F. M.: Exploiting disjointness axioms to improve semantic similarity measures, Bioinformatics, Oxford Univ Press, vol. in press (2013).
13. Couto, F. M. and Silva, M.: Disjunctive shared information between ontology concepts: application to Gene Ontology, Journal of Biomedical Semantics, vol. 2, no. 5, pp. 1-16, (2011)