Combining machine learning, crowdsourcing and expert knowledge to detect chemical-induced diseases in text

Àlex Bravo[1], Tong Shu Li[2], Andrew I. Su[2], Benjamin M. Good[2], Laura I. Furlong[1]

[1]Research Programme on Biomedical Informatics (GRIB), IMIM, UPF, Barcelona, Spain
[2]Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, California
abravo@imim.es
tongli@scripps.edu
asu@scripps.edu
bgood@scripps.edu
lfurlong@imim.es

Abstract. We have developed a system to extract Chemical-induced Disease (CID) relations corresponding to the Task 3.B of Biocreative V (BC5) by combining three strategies: machine learning, rule- and knowledge-based approaches. The first two approaches focus on identifying relations at the sentence-level, while the knowledge-based approach is applied both at sentence and abstract levels. The machine learning method is based on the BeFree system using two corpora as training data: the annotated data provided by the CID task organizers and a new CID corpus developed by crowdsourcing. Different combinations of results from the three strategies were selected for each run. In the Development set, the combined system approaches the highest Recall of the task with a Precision of 45%.

Keywords: Relation Extraction, Text Mining, Drug Side Effects, Crowdsourcing, Citizen Science.

1 Introduction

Text mining systems help us to extract, structure, integrate and automatically analyze information contained in millions of documents from the biomedical literature. The BC5 challenge designed a specific task for Chemical-induced Disease (CID) relations to promote the development of text mining solutions for the identification of drug side effects from scientific publications.

Previous approaches aimed at identifying drug side effects applied different strategies: co-occurrence based statistics [1], [2]; pattern-based approaches [3]; machine learning approaches [4]; and knowledge-based approaches [5]. In addition to scientific publications, clinical records and public forums were used as a source of CID relations. Most of them focused on CID relations stated at the sentence level. In this regard, the BC5 CID task represents a new challenge since the relations are both defined at the abstract and the sentence level.

In this paper, we present a system to extract CID relations at the document level for the BC5 task [6]. In addition, we present a new corpus focused on CID relations,

adfa, p. 1, 2011.
© Springer-Verlag Berlin Heidelberg 2011

which has been developed by crowdsourcing. We present the results obtained on the BC5 development and evaluation sets.

2 Methods

The task 3.B of Biocreative V (BC5) focuses on Chemical-induced Disease (CID) relations found in Medline abstracts. These relations can be expressed in a single sentence or span multiple sentences. We developed a system to identify CID relations both at the sentence and the abstract level, combining three strategies: machine learning, pattern-matching and a knowledge-based approach. The first two approaches aim at identifying relations at the sentence-level, while the knowledge-based approach is applied at both sentence and abstract levels. Fig. 1 shows a scheme of the system and the three configurations used for the challenge runs.

The NER systems DNorm [7], [8] and tmChem [9] were used to identify disease and chemical mentions, respectively, in the documents. Then, we obtained two sets of co-occurring entities, disease-chemical pairs co-occurring at the sentence level and disease-chemical pairs co-occurring at the abstract level.

2.1 Machine Learning approach

The machine learning method is based on our BeFree system [10], a tool to identify relations between biomedical entities in free text. BeFree is composed of a module for Biomedical Named Entity Recognition (BioNER) and a module for Relation Extraction (RE) based on Support Vector Machines (SVM). In this work, only the RE module was used. The RE module requires different morpho-syntactic features which are extracted for each sentence. The methods to obtain such features from a text can be classified in two groups: shallow and deep linguistic parsing. Due to the processing time constraints of the BC5 challenge, only shallow linguistic features were used (such as part-of-speech, lemma, and stem).

A predictive model for CID relations at the sentence level was developed using two different corpora: the annotation data provided by the CID task organizers (training and development sets) and a newly developed CID corpus (crowd-CID relation corpus).

Annotated data provided by BC5. All the abstracts from the training and development ($BC5_T$ and $BC5_D$) sets [11] were processed by the DNorm and tmChem NERs to identify disease and chemical entities. Then, the sentences including at least one co-occurrence between a disease and a chemical were selected as possible CID relations for training the BeFree RE module. If the identifiers of a specific co-occurrence were reported as true by the gold standard, the sentence and the CID pair was saved as a true CID relation, otherwise it was considered a false CID relation. According to this procedure, 3,632 true and 6,122 false CID relations were found in 3,728 sentences. The BeFree model trained on this set achieved a performance of 59.86%, 68.82% and 63.96% of precision, recall and F-score, respectively, by 10-fold cross-validation.

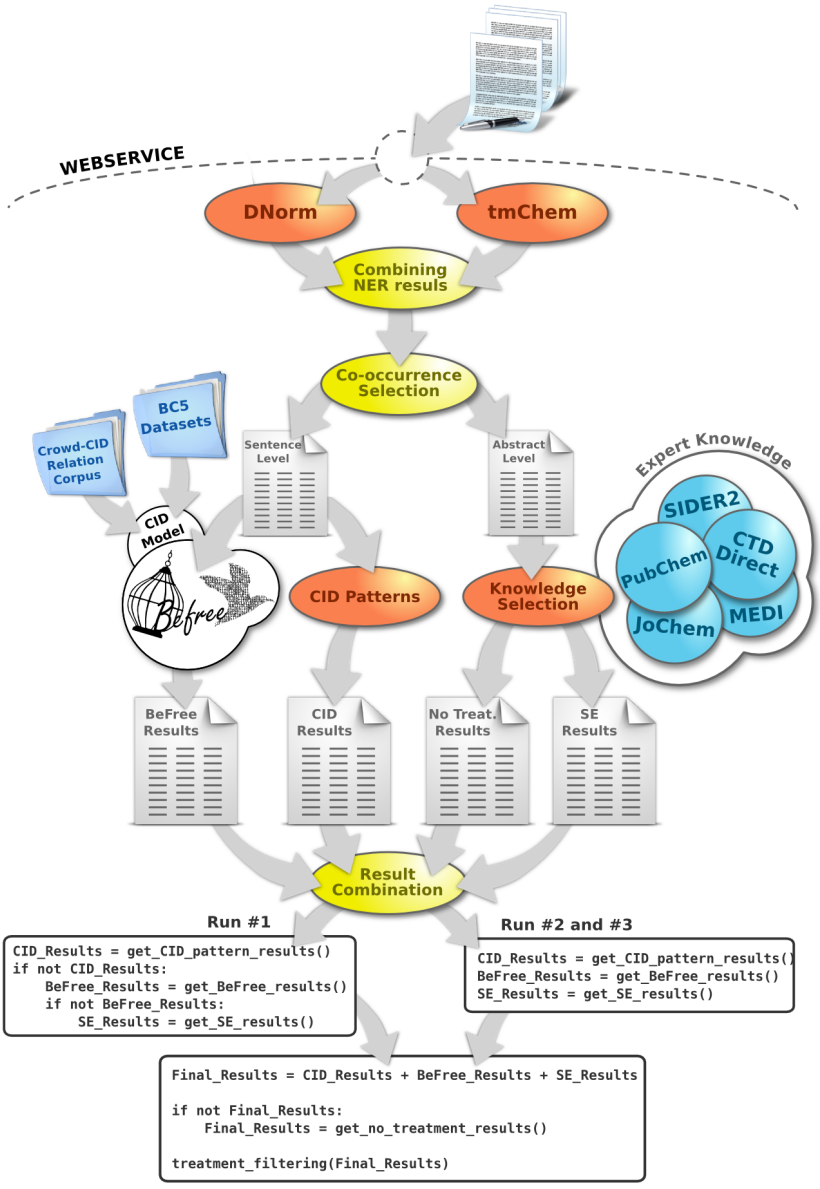

Fig. 1. The workflow diagram of the developed system for CID relation extraction

Crowd-CID relation corpus. A set of 1,990,686 abstracts focused on diseases, treatments and side effects (SEs) were retrieved from PubMed. Then, 3,000 abstracts were randomly selected for the annotation process. The DNorm and the tmChem NERs were used to identify disease and chemical mentions and sentences containing at least one co-occurrence between a disease and a chemical name were selected for annotation by crowd workers. Overall, a set of 2,756 PubMed abstracts contained a total of 17,198 unique sentences which at least one chemical or disease annotation. Of these 17,198 sentences, only 2,953 sentences contained both a chemical and a disease annotation. There were a total of 3,068 unique chemical-disease identifier pairs that resulted in 5,160 unique sentence-concept-pair triples representing crowd verification tasks.

Workers on the Crowdflower microtask platform were presented each sentence containing a putative CID relation and asked to judge whether the relation held. Five workers processed each sentence and were paid $0.02 per sentence completed. Workers were provided detailed instructions with multiple examples and were required to pass a quiz with 70% minimum accuracy before gaining access to the task. In addition, the system removed results from workers who failed hidden test questions embedded in the task stream when they dropped below 70% accuracy. Worker judgments were aggregated based on answer choice, resulting in the number of workers who said a relation was true, using majority rules. If more workers judged a CID relation to be correct than not, the relation was tagged as true, otherwise it was considered as false. Both true and false relations were included for training purposes.

A total of 6,276 CID relations from 2,899 sentences were included in the corpus where 4,336 annotations had complete agreement between the crowd workers (5/5).

The corpus was unbalanced between the number of true and false examples (only 869 examples were annotated as true CID relations). Consequently, the corpus was balanced by selecting at random a subset of false examples, obtaining a total of 1,838 examples from 1,251 sentences (869 true and 969 false examples).

The BeFree model trained with the crowd-CID relation corpus achieved a performance of 82.03%, 73.39% and 76.82% of precision, recall and F-score, respectively, by 10-fold cross-validation.

2.2 Rule-based approach

The rule-based approach is the most straightforward technique to identify relations between two entities at the sentence level. Xu & Wang 2014 reported the most frequent patterns used to express chemical-SE relations in FDA drug labels [12]. We used a subset of these patterns (such as “CHEMICAL-induced SE”, “CHEMICAL-associated SE”, “SE caused by CHEMICAL” and “SE during CHEMICAL”) to identify CID relations at the single sentence level.

2.3 Expert knowledge-based approach (EK)

We checked if a CID relation identified in the text (at the sentence or abstract level) was already known as a drug side effect, or as the therapeutic indication of the drug. For this purpose we developed a database of drug therapeutic indications and side effects by integrating information from different resources (CTD [13], SIDER2 [14] and MEDI [15] databases). JoChem [16] and PubChem [17] were used to map the chemical entities to MeSH identifiers. The UMLS Metathesurus MRCONSO table was used to map between MeSH and UMLS identifiers. The database contains 28,455 chemical-disase associations mapped to MeSH labelled as therapeutic, and 55,960 labelled as side efects. Associations labelled both as therapeutic and side effects were not included in the database.

2.4 Combined system

Fig. 1 shows a schema of our combined system to detect CID relations for the BC5 task. The machine learning and rule-based approaches focus on associations described at the single sentence level, while the expert knowledge-based approach was applied to the set of co-occurrences found at the abstract-level. Note that the set at the abstract-level did not include co-occurrences at sentence-level. The BeFree model was trained with the crowd-CID relation corpus and the $BC5_T$ set when assessing the system performance on the development set. For the challenge evaluation set, the model was trained on the crowd-CID relation corpus, the $BC5_T$ and $BC5_D$ sets.

3 Results and discussion

We present the results obtained on the development ($BC5_D$) and on the evaluation ($BC5_E$) sets. Different configurations of the systems were evaluated on the first 50 abstracts from the $BC5_D$ set. Then, the ones achieving better F-score were selected for the 3 runs on the challenge $BC5_E$ set.

From the 1,012 CID relations annotated in the $BC5_D$, approximately 70% of CID relations were relations found at the single sentence level. On the other hand, 70% of CID relations were true associations. Applying a simple co-occurrence approach on all the chemical and disease mentions identified by the provided BC5 NER systems on this gold standard results in 16.43% Precision and 76.45% Recall (16.46% Precision and 81.71% Recall were obtained on the subset of 50 abstracts from $BC5_D$). Thus, 76.45% is our upper limit in Recall in the $BC5_D$ (81.71% in the subset of 50 abstracts from $BC5_D$).

Table 1 shows the results obtained with our system. By only applying the expert knowledge-based approach to identify CID relations at abstract-level, we achieved 60.98%, 42.37% and 50.00% Recall, Precision and an F-score, respectively. The low Precision obtained with a knowledge-based approach is surprising. Without performing an error analysis we can speculate that issues related with identifier mapping between the knowledge base and the evaluation data, or different annotation criteria between the databases and the evaluation set might result in a large number of False

Positives. The low Recall can be explained by i) not considering associations at sentence-level, ii) limitations of the knowledge sources considered, iii) issues related with identifier mapping. On the other hand, the approaches aimed at detecting CID relations at the sentence-level were also individually evaluated. The rule-based approach resulted in a high precision and low recall (71.43% and 12.20% respectively), while the BeFree system achieved 51.16%, 53.66% and 52.38% of precision, recall and F-score, respectively (46.34% of recall, 38.78% of precision and 42.22% of F-score were the results obtained only using the $BC5_T$ set).

The different approaches were combined in different ways for the challenge (Fig. 1). In Run #1, the CID pattern approach was applied first, followed by BeFree and finally the knowledge-based approach. If a CID was found by the pattern based approach, the other two methods were not applied and the results reported were the ones resulting from pattern matching. In contrast, the three approaches were applied simultaneously in Run #2 and #3, and the results were the union of the results of each of them. The difference between these two runs is that in Run #3 the CID pattern approach is only applied to the title of the abstract and not to the remaining text. In the 3 runs, the final set of results was filtered by removing those CID relations that were annotated as therapeutic in our knowledge base.

Finally, the best performance obtained on the $BC5_E$ set was with Run #1, achieving 48.57% Precision, 38.27% Recall and 42.81% F-score. Compared to Run#1 on 50 documents of the $BC5_D$, the performance on the evaluation sets is significantly decreased. However, the results are not that different to the ones obtained on the 500 docuemnts of the $BC5_D$. Based on these results and those from comparisons between the 50 and 500 document set for Run #2 and #3 (data not shown), we can suggest that the differences in performance arise from distinct sub-sets within the data that have different characteristics. Another factor explaining the differences would be that the data used to train the BeFree model, which differs in these two settings (Development and Evaluation sets).

4 Conclusion and Future work

We have presented a new system to detect CID relations leveraging machine learning trained on expert knowledge and knowledge of the crowd, and simple pattern matching. The union of the results obtained by these approaches achieves the highest Recall of our system (79.27%), approaching the Recall upper limit of the Challenge (81%), at least in the Development set (50 abstracts). On the other hand, the low Precision obtained could be explained by our approach to detect relations at the abstract level, which considers all possible co-occurrences and relies on available expert knowledge to remove False Positives. An exhaustive error analysis will be performed when the $BC5_E$ set is available and will allow us to work towards improvement of the system. Finally, we also present a new crowd-CID relation corpus developed by a crowdsourcing annotation process, which improves the predictions of a machine learning system (BeFree).

Table 1. Performance of the different methods evaluated.

Exp.	Test Set	Method	Train Data	Level	R	P	F
1	$BC5_D$[1]	Co-occurrence	NA	Both	81.71	16.46	27.40
2	$BC5_D$[1]	EK	NA	Abst.	60.98	42.37	50.00
3	$BC5_D$[1]	CID-patterns	NA	Sent.	12.20	71.43	20.83
4	$BC5_D$[1]	CID-pat. + EK	NA	Both	63.41	42.62	50.98
5	$BC5_D$[1]	BeFree system	$BC5_T$	Sent.	47.56	38.61	42.62
6	$BC5_D$[1]	BeFree system	crowdCID	Sent.	54.88	33.33	41.47
7	$BC5_D$[1]	BeFree system	$BC5_T$ + crowdCID	Sent.	53.66	39.20	45.62
8	$BC5_D$[1]	Run #1	$BC5_T$	Both	57.31	63.51	60.25
9	$BC5_D$[1]	Run #1	crowdCID	Both	58.53	66.66	62.33
10	$BC5_D$[1]	Run #1	$BC5_T$ + crowdCID	Both	57.32	66.20	61.44
11	$BC5_D$[2]	Run #1	$BC5_T$ + crowdCID	Both	46.74	56.04	50.97
12	$BC5_E$[2]	Run #1	$BC5_{T+D}$ + crowdCID	Both	38.27	48.57	42.81
13	$BC5_D$[1]	Run #2	$BC5_T$ + crowdCID	Both	79.27	43.91	56.52
14	$BC5_D$[1]	Run #3	$BC5_T$ + crowdCID	Both	79.27	44.83	57.27

[1]subset of the first 50 abstracts.

[2]the full set (500 abstracts).

5 Acknowledgment

This work was supported by ISCIII-FEDER (CP10/00524 and PI13/00082), the IMI (115002 (eTOX), 115191 [Open PHACTS]), resources of which are composed of financial contribution from the EU-FP7 [FP7/2007-2013] and EFPIA companies' in kind contribution, and the European Union's Horizon 2020 research and innovation programme 2014-2020 under Grant Agreement No 634143. The Research Programme on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB). In addition, this work was also supported by grants from the National Institute of General Medical Science (GM114833, GM089820, TR001114).

REFERENCES

1. X. Wang, G. Hripcsak, M. Markatou, and C. Friedman, "Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: A Feasibility Study," *J. Am. Med. Informatics Assoc.*, vol. 16, no. 3, pp. 328–337, 2009.
2. R. Leaman and L. Wojtulewicz, "Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks," *Proc. ...*, no. July, pp. 117–125, 2010.
3. J. Liu, A. Li, and S. Seneff, "Automatic Drug Side Effect Discovery from Online Patient-Submitted Reviews: Focus on Statin Drugs," *IMMM 2011, First Int. Conf.*, 2011.
4. Y. Miura, E. Aramaki, T. Ohkuma, and M. Tonoike, "Adverse – Effect Relations Extraction from Massive Clinical Records," no. August, pp. 75–83, 2010.

5. N. Kang, B. Singh, C. Bui, Z. Afzal, E. M. van Mulligen, and J. A. Kors, "Knowledge-based extraction of adverse drug events from biomedical text.," *BMC Bioinformatics*, vol. 15, no. 1, p. 64, 2014.
6. L. R. Wei CH, Peng Y, "Overview of the BioCreative V Chemical Disease Relation (CDR) Task," *Proc. fifth BioCreative Chall. Eval. Work.*, 2015.
7. R. Leaman, R. Khare, and Z. Lu, "Challenges in clinical natural language processing for automated disorder normalization," *J. Biomed. Inform.*, vol. 57, pp. 28–37, 2015.
8. R. Leaman, R. I. Doğan, and Z. Lu, "DNorm: Disease name normalization with pairwise learning to rank," *Bioinformatics*, vol. 29, no. 22, pp. 2909–2917, 2013.
9. R. Leaman, C. Wei, and Z. Lu, "tmChem : a high performance approach for chemical named entity recognition and normalization," *J. Cheminform.*, vol. 7, no. Suppl 1, p. S3, 2015.
10. À. Bravo, J. Pinero, N. Queralt, M. Rautschka, and L. I. Furlong, "Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research," 2014.
11. J. R. Li J, Sun Y, "Annotating chemicals, diseases, and their interactions in biomedical literature, in Proceedings of the fifth BioCreative challenge evaluation workshop, Sevilla, Spain," 2015.
12. R. Xu and Q. Wang, "Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature," *Journal of Biomedical Informatics*, 2014.
13. "Curated chemical–disease data were retrieved from the Comparative Toxicogenomics Database (CTD), MDI Biological Laboratory, Salisbury Cove, Maine, and NC State University, Raleigh, North Carolina. World Wide Web (URL: http://ctdbase.org/). July, 2015." [Online]. Available: http://ctdbase.org/.
14. M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, "A side effect resource to capture phenotypic effects of drugs.," *Mol. Syst. Biol.*, vol. 6, p. 343, 2010.
15. W.-Q. Wei, R. M. Cronin, H. Xu, T. a Lasko, L. Bastarache, and J. C. Denny, "Development and evaluation of an ensemble resource linking medications to their indications.," *J. Am. Med. Inform. Assoc.*, vol. 20, no. 5, pp. 954–61, 2013.
16. K. M. Hettne, R. H. Stierum, M. J. Schuemie, P. J. M. Hendriksen, B. J. A. Schijvenaars, E. M. Van Mulligen, J. Kleinjans, and J. A. Kors, "A dictionary to identify small molecules and drugs in free text," *Bioinformatics*, vol. 25, no. 22, pp. 2983–2991, 2009.
17. E. E. Bolton, Y. Wang, P. A. Thiessen, and S. H. Bryant, "Chapter 12 PubChem: Integrated Platform of Small Molecules and Biological Activities," *Annual Reports in Computational Chemistry*, vol. 4. pp. 217–241, 2008.