Extracting structured chemical-induced disease relations from free text via crowdsourcing

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

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

Abstract. Relationships between chemicals and diseases are important for biomedical research. Assembling databases of these relations is costly and often relies on expert curation. Here, we describe a crowdsourcing workflow for extracting structured chemical-induced disease (CID) relations from free text. Using the CrowdFlower platform, we recruited workers with no specialized training or education. For each candidate relation, five workers were shown the concepts in context and asked to judge whether the text supported a CID relation. Maximal agreement with the gold standard was achieved at a threshold of four or more positive votes. On a test set of 100 abstracts, the crowd attained an F-score of 0.587 (0.528 precision, 0.661 recall). For the 500 abstract evaluation dataset, the F-score was 0.505 (0.475 precision, 0.540 recall).

Key words: Crowdsourcing, relation extraction, natural language processing

1 Introduction

Relations between chemicals and diseases are of major importance in biomedical research, and represent one of the top ten queried associations in PubMed[1]. In the drug development field, chemical-disease interactions guide target discovery, drug repurposing, and toxicity prediction[2].

While the biomedical literature is rich in information regarding chemical-disease interactions, this information is typically represented as free text, a format that is difficult to access computationally. As the number of scientific publications grows exponentially[3], keeping abreast with the literature becomes an increasingly difficult task.

Structured representations of the literature, which can be computationally accessed, queried, and analyzed, standardize and organize existing information, resulting in improved information access[4]. However, existing efforts at creating these structured representations are expensive endeavours requiring large amounts of expert manual labor[5].

Automated relation extraction methods employing pattern, knowledge, or machine learning based approaches have been developed as alternatives to manual expert curation[6–8]. While performance can be quite good, one major limitation of many automated methods is that extracted relations are sentence-constrained.

In recent years, crowdsourcing has emerged as an alternative to both expert curation and automated methods for performing biomedical NLP tasks[9–11]. Here we describe a crowdsourcing workflow for extracting structured chemical-induced disease (CID) relations from free text.

2 Method

The objective for the CID track of the BioCreative V challenge was to extract CID relations as pairs of Medical Subject Heading (MeSH) identifiers from the free text of a scientific abstract[12]. This task had two separate subtasks: performing named entity recognition (NER) on the raw text to produce MeSH annotations, and linking MeSH identifiers together as CID relations. The BioCreative organizers provided 1000 PubMed abstracts manually annotated with chemical and disease mentions and CID relations, evenly divided into training and development sets[13]. We randomly selected a subset of 100 abstracts to develop our crowd-based approach.

Our crowdsourcing method focused exclusively on the relation extraction subtask (fig. 1). First, tmChem[14] and DNorm[15] were used to process the raw text and generate a set of MeSH annotations. We resolved acronyms without MeSH IDs by a simple rule-based pattern match against other annotations identified in the text. Then, all possible unique chemical-disease identifier pairs were calculated and divided into three mutually exclusive categories. Pairs which followed a simple CID pattern (chemical annotation occurs no more than 15 characters before disease annotation, and the text between them contains "induce") were identified and automatically judged to always be true, and were never shown to a crowd.

The remaining pairs were divided into those which never co-occurred within any sentence, and those which co-occurred at least once within any sentence. Lingpipe was used to split abstracts into individual sentences[16]. A separate crowdsourcing task was used to process the two sets of relation pairs. In both tasks, five workers were shown one relation identifier pair in the original context (either a single sentence or the full abstract) and asked to make a judgment about whether the provided text supported a CID relation between the chemical and disease. All annotations of the chemical and disease were highlighted in the text. For the abstract-level task, the judgment contained two choices, "true" or "false". For the sentence-level task, previous testing showed that workers were falsely annotating chemical-disease relationships following a "[chemical]-induced [intermediate disease] causes [disease]" pattern as CID relations. A third choice was included to capture these low frequency relations (which were treated as "false" during evaluation). An example of a sentence task is given in fig. 2.

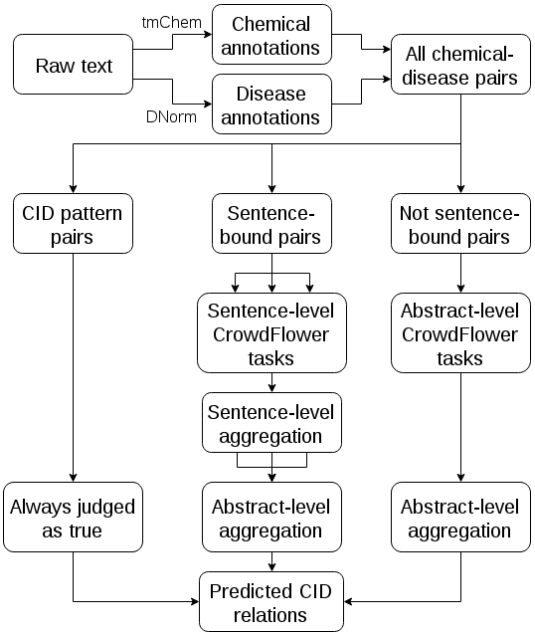

Fig. 1. Workflow diagram for extracting CID relations from free text.

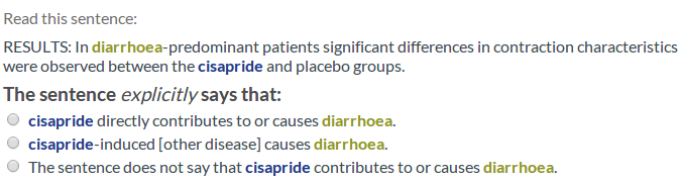

Fig. 2. An example of a sentence-level task with chemical and disease highlighted.

We used the CrowdFlower platform for all of our crowdsourcing jobs. Workers for both tasks had to pass an initial six question quiz and then maintain an overall minimum accuracy of 70% on a separate set of test questions hidden in the task stream. Those who fell below this limit were automatically removed and had their judgments discarded. Workers were paid 2 and 4 cents, and required to spend a minimum of 3 and 10 seconds per sentence- and abstract-level task respectively.

To produce a list of CID relations for each abstract, worker judgments were aggregated as follows: for pairs which did not co-occur within a sentence, positive votes for each relation were counted, and relations which received four or more positive votes were judged to be CID relations. Result aggregation for the sentence co-occurring pairs took into account that multiple sentences from the abstract could contain the same relation. Vote aggregation first occurred at the sentence level to produce a signal of whether a particular sentence supported the CID relation. Next, the sentence task with the maximum number of positive votes was used to represent whether the relation was true at the abstract level. This aggregation scheme assumed that the relation was true for the entire abstract if at least one sentence supported the relation.

CID relation extraction performance against the gold standard was calculated according to the scoring metric used by the official task creators[12]. Specifically, the set of gold standard data points for all abstracts was compared to the set of data points generated by the crowd. A data point was defined as a 3-tuple consisting of the document ID, chemical MeSH ID, and disease MeSH ID. True positives were defined as the intersection between the gold standard triples and the predicted CID triples. False positives were defined as the predicted triples minus the gold triples, and false negatives as the gold triples minus the predicted triples.

3 Results

The sentence- and abstract-level tasks for the 100 abstract test set completed within three hours, and cost $98.35 and $162.72 respectively ($2.61 per abstract). The crowd's ability to exactly match the gold standard CID relations as a function of the number of positive votes is given in figure 3A. Peak performance occurred at a threshold of four or more positive votes, resulting in an overall F-score of 0.587 (0.568 precision, 0.661 recall). CID, sentence-bound, and non-sentence bound relation pairs represented 4.1%, 42.3%, and 53.6% of the set of all possible NER generated relation pairs respectively. Performance on each category of relation pairs at the threshold of four or more votes with respect to the full gold standard was 0.232 F_1 for CID relations (0.690 precision, 0.140 recall), 0.442 F_1 for sentence-bound relations (0.526 precision, 0.381 recall), and 0.211 F_1 for non-sentence bound pairs (0.432 precision, 0.140 recall). Some concept identifiers in the gold standard were not identified in the NER step, and therefore the maximum possible recall for CID relations was 0.864.

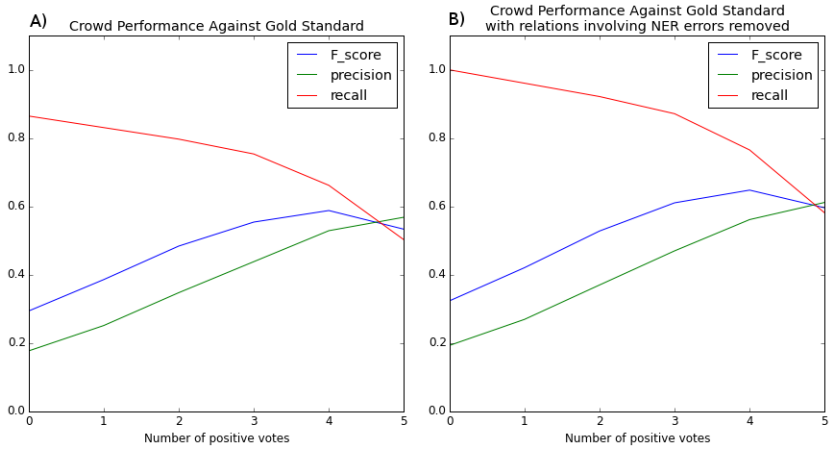

Fig. 3. A) Performance against the gold standard as a function of the number of votes crowd relations received for 100 abstracts. B) Performance against the gold standard when relations involving errors in NER were removed.

Since the crowd was never asked to perform relation extraction *de novo*, 13.6% of relations were impossible to verify due to the correct identifiers never appearing in the annotations produced by the NER tools. In our relation verification task, these relations were always counted as false negatives. Similarly, relations which used identifiers nonexistent in the gold standard were always judged to be false positives. To examine the crowd's performance on the relation extraction subtask only, we also performed an analysis in which both the false positives and false negatives from NER were filtered out. On this subset of relations, the crowd performed with 0.647 F-score (0.561 precision, 0.765 recall) (fig. 3B).

The same crowdsourcing interfaces and workflow were used for the official 500 abstract evaluation set. The sentence- and abstract-level tasks completed within 6 and 7 hours, and cost $439.63 and $851.04 respectively ($2.58 per abstract). Our performance on the evaluation dataset was 0.505 F-score (0.475 precision, 0.540 recall). A pure tmChem/DNorm-annotated co-occurrence model achieved a base performance of 0.267 F-score (0.162 precision, 0.751 recall). Based on the maximum recall of the co-occurrence model, it seems that tmChem and DNorm had lower NER performance on the official dataset, which may explain the corresponding decrease in F-score of the crowd's performance.

4 Discussion

This crowdsourcing approach is heavily influenced by the NER step. Concepts missed during NER create an upper bound on recall. False positives also occur when different identifiers are used to annotate the same concepts.

Our crowd also identified several instances where the gold standard appeared to be incorrect. For example, in PMID 18997632, the gold standard included a CID relation between "caffeine" and "seizure". However, the abstract stated "intravenous caffeine is commonly used to improve seizure duration". This seemed to be an error in the gold standard, since the text stated that caffeine reduced (and not induced) seizure duration. In our data, 5/5 workers rated this relation to be negative. Another example is PMID 15867025, from which the gold standard asserted a CID relation between "hepatitis B surface antigen" and "hepatitis B". However, the paper assessed vaccine adoption rates in hospitals, and never claimed hepatitis B surface antigen caused hepatitis B. Again, our crowd unanimously rated this as a false relation. Examples like these demonstrate that the crowd may also provide a useful check on expert-generated gold standards.

Many techniques can be made to boost crowd performance. Firstly, crowd involvement in the NER step would allow for *de novo* relation extraction. Secondly, a hierarchical post-crowd normalization could be used to remove false positives. For example, in PMID 982002, the crowd annotated both "acute renal failure" and "renal failure" in CID relations with "rifampcin". The MeSH ontology could be used to eliminate the false positive "rifampcin induces renal failure" because "acute renal failure" is more specific. Finally, using more advanced methods to aggregate worker judgments would likely lead to better performance[17].

In conclusion, we have demonstrated crowdsourcing as an inexpensive and fast method for CID relation extraction from free text which performs relatively well compared to a co-occurrence model.

Acknowledgments. We would like to thank Dr. Zhiyong Lu for sending us the evaluation dataset. This work was supported by grants from the National Institute of Health (GM114833, GM089820, TR001114).

References

1. Dogan, R.I., Murray, G.C., Névéol, A., Lu, Z.: Understanding PubMed user search behavior through log analysis. Database (Oxford). 2009:bap018. DOI:10.1093/database/bap018 (2009)
2. Hopkins, A.L.: Network pharmacology: the next paradigm in drug discovery. Nat. Chem. Biol. 4, 682–690 (2008)
3. Lu, Z.: PubMed and beyond: a survey of web tools for searching biomedical literature. Database (Oxford). 2011:baq036. DOI:10.1093/database/baq036 (2011)
4. Ananiadou, S., Kell, D.B., Tsujii, J.: Text mining and its potential applications in systems biology. Trends Biotechnol. 24, 571–579 (2006)
5. Davis, A.P., Wiegers, T.C., Roberts, P.M., King, B.L, Lay, J.M., Lennon-Hopkins, K., Sciaky, D., Johnson, R., Keating, H., Greene, N., Hernandez, R., McConnell, K.J., Enayetallah, A.E., Mattingly, C.J.: A CTD-Pfizer collaboration: manual curation of 88000 scientific articles text mined for drug-disease and drug-phenotype interactions. Database (Oxford). 2013:bat080. DOI:10.1093/database/bat080 (2013)
6. Xu, R., Wang, Q.: Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature. J. Biomed. Inform. 51, 191–199 (2014)

7. Kang, N., Singh, B., Bui, C., Afzal, Z., van Mulligen, E.M., Kors, J.A.: Knowledge-based extraction of adverse drug events from biomedical text. BMC Bioinformatics. 15, 64 (2014)
8. Gurulingappa, H., Mateen-Rajput, A., Toldo, L.: Extraction of potential adverse drug events from medical case reports. J. Biomed. Semantics. 3, 15 (2012)
9. Good, B.M., Su, A.I.: Crowdsourcing for bioinformatics. Bioinformatics. 29, 1925–1933 (2013)
10. Khare, R., Burger, J.D., Aberdeen, J.S., Tresner-Kirsch, D.W., Corrales, T.J., Hirchman, L., Lu, Z.:Scaling drug indication curation through crowdsourcing. Database (Oxford). 2015:bav016 (2015)
11. Mortensen, J.M., Minty, E.P., Januszyk, M., Sweeney, T.E., Rector, A.L., Noy, N.F., Musen, M.A.: Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of SNOMED CT. J. Am. Med. Inform. Assoc. 22, 640–648 (2015)
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, Sevilla, Spain (2015)
13. 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, Sevilla, Spain (2015)
14. Leaman, R., Wei, C.H., Lu, Z.: tmChem: a high performance approach for chemical named entity recognition and normalization. J. Cheminform. 7(Suppl 1): S3 (2015)
15. Leaman, R., Doğan, R.I., Lu, Z.: DNorm: disease name normalization with pairwise learning to rank. Bioinformatics. 29, 2909–2917 (2013)
16. Alias-i. LingPipe 4.1.0. `http://alias-i.com/lingpipe` (2008)
17. Simpson, E., Roberts, S., Psorakis, I.: Bayesian combination of multiple, imperfect classifiers. In: NIPS 2011, Spain. (2011)