Annotating chemicals, diseases and their interactions in biomedical literature

Jiao Li[1], Yueping Sun[1], Robin J. Johnson[2], Daniela Sciaky[2], Chih-Hsuan Wei[3], Robert Leaman[3], Allan Peter Davis[2], Carolyn J. Mattingly[2], Thomas C. Wiegers[2], Zhiyong Lu[3,*]

[1]Institute of Medical Information and Library, Beijing City, China 100020
[2]Department of Biological Sciences, North Carolina State University, Raleigh, NC, USA 27695
[3]National Center for Biotechnology Information, Bethesda, MD, USA 20894

li.jiao@imicams.ac.cn, sun.yueping@imicams.ac.cn, rjjohns3@ncsu.edu, dsciaky@ncsu.edu, chih-hsuan.wei@nih.gov, robert.leaman@nih.gov, apdavis3@ncsu.edu, cjmattin@ncsu.edu, tcwieger@ncsu.edu, *zhiyong.lu@nih.gov

Abstract. Community-run formal evaluations and manually annotated text corpora are critically important for advancing biomedical text mining research. Recently in BioCreative V, a new challenge was organized for the tasks of disease named entity recognition (DNER) and chemical-induced disease (CID) relation extraction. Given the nature of both tasks, a test collection is required to contain both disease/chemical annotations and relation annotations in the same set of articles. Despite previous efforts in biomedical corpus construction, none were found to be sufficient for the task. Thus, we developed our own corpus during the challenge by inviting a team of expert annotators from both MeSH and the Comparative Toxicogenomics Database (CTD), who performed manual annotation of entities (diseases/chemicals) and relations, respectively. To ensure high annotation quality and productivity, detailed annotation guidelines and automatic annotation tools were provided. The resulting corpus consists of 1,500 PubMed articles with 4,409 annotated chemicals, 5,818 diseases, and 3,116 chemical-disease interactions. Each annotation includes both the mention text spans and normalized concept identifiers (MeSH was used as the controlled vocabulary). To ensure accuracy, the entities were captured independently by two annotators; the average inter-annotator agreement (IAA) scores are 88.75% and 96.31% for the disease and chemicals, respectively, in the test set according to the Jaccard similarity coefficient. Our corpus was successfully used for the BioCreative V challenge tasks and should serve as a valuable resource for the text-mining research community.

Keywords. Corpus Annotation; BioNLP; Text Mining; CTD; PubTator

1 Introduction

Relations between chemicals and diseases (Chemical-Disease Relations or CDRs) play critical roles in drug discovery (toxicity), biocuration, pharmacovigilance, etc (1). Because of their critical significance, databases such as the Comparative Toxicogenomic Database (CTD) (2) are manually curating these relations. Due to the high cost of manual curation and rapid growth of the biomedical literature, several attempts have been made to assist curation using text-mining systems (3,4) including the automatic extraction of CDRs (5). These attempts have met with limited success, however, in part due to the lack of a large-scale training corpus. Through BioCreative V in 2015, one of the major formal evaluations for text mining research (6), a new challenge was recently organized to advance the state-of-the-art in extracting CDRs (7). The challenge included two subtasks: the disease named entity recognition (DNER) task and the chemical-induced disease (CID) relation extraction task.

To support both tasks, a text corpus of PubMed abstracts containing annotations of both chemical/diseases and their interactions is desirable. Despite the existence of many biomedical corpora (see (8) for a brief review) including a few specifically targeting diseases (9-11) and chemicals (12), there is none that fulfilled the following content criteria: a), our corpus needs to include instances of chemical-disease relation annotations that are asserted from both within and across sentence boundaries; b) complete chemical, disease and relation annotations need be present within each abstract; c) chemical/disease annotations need be grounded in concept identifiers via a controlled vocabulary. Thus, we proposed building a new corpus that satisfies these three requirements.

The proposed corpus is related to some previous efforts in corpus annotation for biomedical information extraction research, such as protein-protein interaction (13) and drug-drug interaction (14). It is also significantly different from the previously constructed corpora (15,16) for mining adverse drug reaction/effects in terms of the annotation scope, requirements (see above), and corpus size (a total of 1,500 articles in our corpus).

2 Methods and Materials

2.1 Article selection

We selected a total of 1,500 articles for the CDR task, split into 3 subsets: 500 each for the training, development, and test sets. The training, development and most (400) of the test set were randomly selected from the CTD-Pfizer corpus, which was generated via a previous collaboration curation between CTD and Pfizer, and consists of over 150,000 chemical-disease relations in 88,000 articles (17).

The remaining 100 articles for the test set had not been previously curated by CTD, but were selected using a process that ensured they would have a similar distribution of words as the training and development sets. For each of the 1,000 articles in the training and development sets, we retrieved the list of related articles using PubMed E-utilities. We removed from consideration any articles that did not meet our selection criteria. Specifically, the target article should be in English, contain an abstract, and be published in 2014 or later. For each new article, we computed an overall score by summing the similarity scores between the target article and each article in the training and development sets. We also determined an overall similarity score for each article in the training and development sets with a similarity score calculated using all other articles in the training and development sets. We then selected the final set by sampling with replacement from the similarity distribution of the training and development sets: we randomly select an article from the training or development sets, get its similarity score, and then select the new article with the closest similarity score. The resulting articles are about as well related to the articles in the training and development sets as the articles in the training and development sets are related to each other.

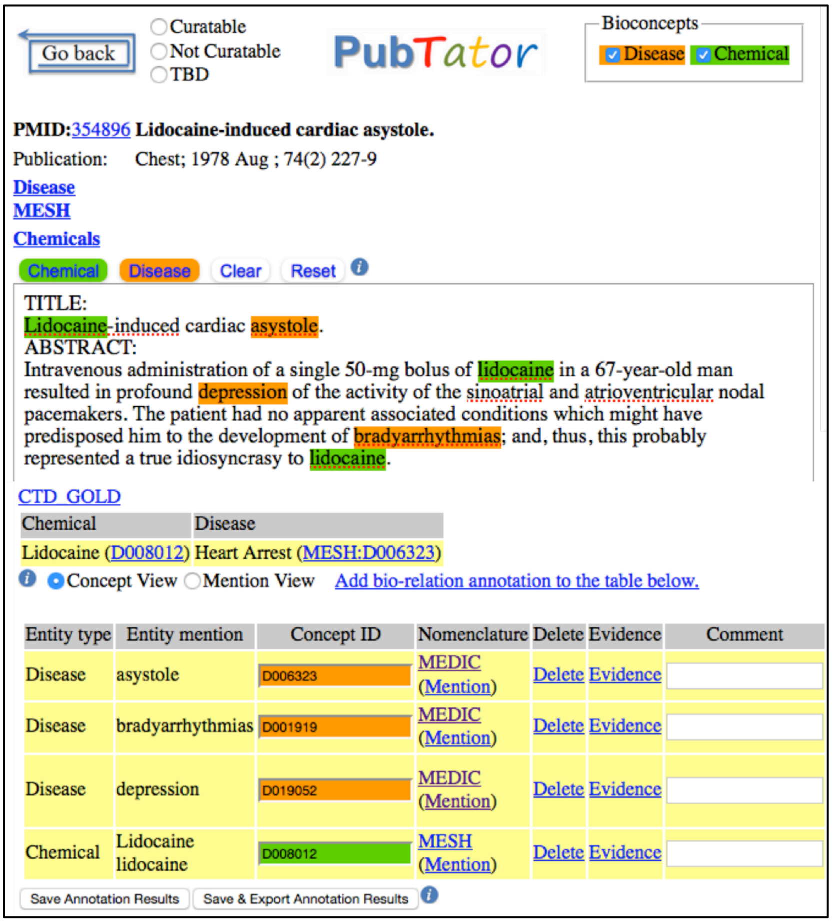
2.2. Annotation Tasks

We performed manual annotation of all chemical and diseases mentioned in the 1,500 articles. For each entity occurrence, we not only annotated its text span but also assigned a relevant concept identifier from MeSH (18), a controlled vocabulary of biomedical concepts. For instance, as shown in Figure 1, three diseases were annotated, along with three occurrences of the same chemical (Lidocaine).

As indicated above, we largely leveraged the previous annotation of chemical-disease relationships from the CTD-Pfizer dataset for 1,400 of the 1,500 articles with few changes: a) we removed relations that required entities not found in abstracts; b) we removed relations that were not disease specific (e.g., “Drug-Related Side Effects and Adverse Reactions” (D064420)); and c) we updated a few CTD relations due to the MeSH vocabulary changes (the CTD-Pfizer project was conducted in years 2011/12, and the MeSH vocabulary has changed since then).

We performed new manual annotation of chemical-disease relations for the remaining 100 articles in the test set. For the BioCreative V challenge task, the chemical-induced disease (CID) relations refer to two types of relationships between a chemical and a disease in CTD: *Putative mechanistic relationships* between a chemical and disease indicates that the chemical may play a role in the etiology of the disease (e.g., exposure to chemical X causes lung cancer). Figure 1 shows an example of such a relationship between Lidocaine and Heart Arrest. *Biomarker relations* between a chemical and disease indicates that the chemical correlates with the disease (e.g., increased abundance in the brain of chemical X correlates with Alzheimer disease).

CTD curators used their standard curation process for CDR curation (19). Curation was limited to the title and abstract except in cases where reference to the full text was required; abstracts that required full text curation were removed from the corpus. In addition to CDR curation, all observed interactions and relationships applicable to CTD where curated for each abstract. CTD triaged and/or curated 143 articles in conjunction with BioCreative V; the final 100 selected for inclusion in the Test Dataset represented abstract-only curation for CDRs.



PMID:354896 Lidocaine-induced cardiac asystole.

Publication: Chest; 1978 Aug ; 74(2) 227-9

Disease
MESH
Chemicals

TITLE:
Lidocaine-induced cardiac asystole.
ABSTRACT:
Intravenous administration of a single 50-mg bolus of lidocaine in a 67-year-old man resulted in profound depression of the activity of the sinoatrial and atrioventricular nodal pacemakers. The patient had no apparent associated conditions which might have predisposed him to the development of bradyarrhythmias; and, thus, this probably represented a true idiosyncrasy to lidocaine.

CTD_GOLD

Chemical	Disease
Lidocaine (D008012)	Heart Arrest (MESH:D006323)

● Concept View ○ Mention View Add bio-relation annotation to the table below.

Entity type	Entity mention	Concept ID	Nomenclature	Delete	Evidence	Comment
Disease	asystole	D006323	MEDIC (Mention)	Delete	Evidence	
Disease	bradyarrhythmias	D001919	MEDIC (Mention)	Delete	Evidence	
Disease	depression	D019052	MEDIC (Mention)	Delete	Evidence	
Chemical	Lidocaine lidocaine	D008012	MESH (Mention)	Delete	Evidence	

Figure 1 Annotation example shown in our annotation tool, PubTator.

2.3 Annotators

For entity annotation, we recruited four MeSH indexers, all of whom had a medical training background and curation experience. Each article was annotated independently by two annotators (i.e., double-annotation). Differences were resolved by a third and senior annotator (YS). Three CTD annotators conducted the relation annotation.

2.4 Annotation guidelines

The task organizers followed the usual practice of biomedical corpus annotation for entity annotation: the annotators were asked to follow an initial set of guidelines when annotating the first 100 sample articles. Annotation discrepancies and questions were discussed and settled by the senior annotator; the annotation guidelines were revised accordingly. Detailed guidelines are available on the task website. For relation annotation, the CTD standard curation protocol was followed.

2.5 Annotation tools

Manual annotation of disease and chemical entities was performed using PubTator (3,4) (See Figure 1). To accelerate manual annotation (20), text-mined disease and chemical results were pre-computed using DNorm (21) and tmChem (22) and displayed to the annotators. When necessary, the annotators added new annotations, and deleted or edited the automatic annotations, based on their judgment. The annotators were permitted to use public resources such as UMLS or Wikipedia to facilitate the annotation process. CTD's in-house Curation Tool (19) was used for all relation curation.

2.6 Annotation data formats

All annotation data was made available to participants in both PubTator and BioC formats. The PubTator format consists of a straightforward tab-delimited text file. The BioC (23) format is a recently proposed XML standard for biomedical text mining research data and tool output.

2.7 Inter-annotator Agreement (IAA) Analysis

To assess the consistency of the disease and chemical annotation, we measured pairwise agreement of duplicate annotations using the Jaccard score (24). As shown below, if we define A as the set of mentions of team A, B as the set of mentions of team B, then the Jaccard agreement score, $S_{A,B}$, can be calculated by counting the number of agreements and disagreements. Mentions with the same PMID, start and end point, and concept identifier are counted as a case of agreement. For example, if one annotator annotates "tardive dystonia" with concept ID of D004421, another annotates "dystonia" with concept ID of D004421, then that would count as two cases of disagreement and no case of agreement as different mentions were annotated.

$$S_{A,B} = \frac{|A \cap B|}{|A \cup B|}$$

IAA for CTD gene-chemical curation has been previously described (19); no further experiments for IAA as it relates to CDR curation were conducted.

3 Results and discussion

3.1 Corpus statistics

As shown in Table 1, the three data sets have similar distributions of chemical mentions, disease mentions, and CID relations, which makes the corpus more useful for training models. The corpus contained more chemical than disease mentions, but more disease than chemical entities (IDs).

Table 1 The overall corpus statistics

Task Dataset	Articles	Chemical		Disease		CID relation
		Mention	ID	Mention	ID	
Training	500	5,203	1,467	4,182	1,965	1,038
Development	500	5,347	1,507	4,244	1,865	1,012
Test	500	5,385	1,435	4,424	1,988	1,066

3.1 Inter-annotator Agreement for mention annotation

Table 2 shows the inter-annotator agreement (IAA) scores of three separate subsets for disease and chemical annotations. The IAA scores over the entire corpus are 87.49% (diseases) and 96.05% (chemicals), which suggest high levels of agreement.

Table 2 Inter-annotator agreement (IAA) scores of the three sets.

Data Set	Average IAAs - disease (%)	Average IAAs - chemical (%)
Training Set	0.8600	0.9523
Development Set	0.8742	0.9577
Test Set	0.8875	0.9630
All Sets	0.8749	0.9605

Annotator disagreements were manually adjudicated, and most were found to be disagreements over the annotation span. As an example, it is difficult to judge whether “rosaceiform dermatitis” should be annotated as “rosacea” or simply “dermatitis”.

Also, there were many cases of disagreements over the concept identifier of diseases, especially for the mentions whose text did not exactly match any MeSH term. In some cases, it is hard to judge whether to assign an unknown concept identifier of "-1" or an ancestor concept identifier. For example, in the article with PMID of 12093990, one annotator annotated "infection with hemorrhagic fever viruses" as "-1", while the other one selected "D006482" (Hemorrhagic Fevers, Viral"). In this case the adjudicating annotator chose the latter term.

4 Conclusions

We developed a corpus for both named entity recognition and chemical-disease relations in the literature. A total of 1,500 articles have been annotated with automated assistance from PubTator. Jaccard agreement results and corpus statistics verified the reliability of the corpus. Furthermore, our annotated data includes the CDR relations that are asserted across sentence boundaries (i.e. not in the same sentences). We believe this data set will be invaluable for advancing text-mining techniques for relation extraction tasks.

5 Acknowledgment

The research was supported by the National Key Technology Research and Development Program of China (Grant No. 2013BAI06B01), the National Population and Health Scientific Data Sharing Program of China, the Fundamental Research Funds for the Central Universities (Grant No. 13R0101), the National Institute of Environmental Health Sciences (ES014065 and ES019604), and the NIH Intramural Research Program, National library of Medicine (CW, RL, ZL).

REFERENCES

1. Dogan RI, Murray GC, Névéol A, et al. (2009) Understanding PubMed user search behavior through log analysis. *Database (Oxford)* **2009**: bap018.
2. Davis AP, Murphy CG, Saraceni-Richards CA, et al. (2009) Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical–gene–disease networks. *Nucleic Acids Res* **37**: D786-D792.

3. Wei C-H, Kao H-Y, Lu Z (2013) PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res* **41**: W518-W522.
4. Wei C-H, Harris BR, Li D, et al. (2012) Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database (Oxford)* **2012** bas041.
5. Wiegers TC, Davis AP, Mattingly CJ (2014) Web services-based text-mining demonstrates broad impacts for interoperability and process simplification. *Database (Oxford)* **2014**: bau050
6. Huang C-C, Lu Z (2015) Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in Bioinformatics*.
7. Wei C-H, Peng Y, Leaman R, et al. (2015) Overview of the BioCreative V Chemical Disease Relation (CDR) Task. The fifth BioCreative challenge evaluation workshop. Sevilla,Spain.
8. Nevesa M (2014) An analysis on the entity annotations in biological corpora. *F1000Research* **25;3:96**: 96.
9. Leaman R, Miller C, Gonzalez G (2009) Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark. The 2009 Symposium on Languages in Biology and Medicine. Jeju Island, South Korea. pp. 82-89.
10. Doğan RI, Leaman R, Lu Z (2014) NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics* **47**: 1-10.
11. Lu RIDZ (2012) An improved corpus of disease mentions in PubMed citations. Proceedings of the 2012 Workshop on Biomedical Natural Language. Stroudsburg, PA, USA. pp. 91-99
12. Krallinger M, Rabal O, Leitner F, et al. (2015) The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics* **7**: S2.
13. Krallinger M, Vazquez M, Leitner F, et al. (2011) The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC Bioinformatics* **12**: S3.
14. Herrero-Zazo M, IsabelSegura-Bedmar, Martínez P, et al. (2013) The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of Biomedical Informatics* **46**: 914-920.
15. Mulligen EMV, Fourrier-Reglat A, Gurwitz D, et al. (2012) The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. *Journal of Biomedical Informatics* **45**: 879-884.
16. Gurulingappa H, Rajput AM, AngusRoberts, et al. (2012) Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics* **45**: 885-892.
17. Davis AP, Wiegers TC, Roberts PM, et al. (2013) A CTD-Pfizer collaboration: manual curation of 88,000 scientific articles text mined for drug-disease and drug-phenotype interactions. *Database (Oxford)* **2013**: bat080.
18. Lipscomb CE (2000) Medical subject headings (MeSH). *Bulletin of the Medical Library Association* **88**: 265.
19. Davis AP, Wiegers TC, Murphy CG, et al. (2011) The curation paradigm and application tool used for manual curation of the scientific literature at the Comparative Toxicogenomics Database. *Database (Oxford)*: bar034.
20. Névéol A, Doğan RI, Lu Z (2011) Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. *Journal of Biomedical Informatics* **44**: 310-318.

21. Leaman R, Doğan RI, Lu Z (2013) DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics* **29**: 2909-2917.

22. Leaman R, Wei C-H, Lu Z (2015) tmChem: a high performance approach for chemical named entity recognition and normalization. *Journal of Cheminformatics* **7**: S3.

23. Comeau DC, Doğan RI, Ciccarese P, et al. (2013) BioC: a minimalist approach to interoperability for biomedical text processing. *Database (Oxford)* **2013**: bat064.

24. Levandowsky M, Winter D (1971) Distance between sets. *Nature* **234**: 34-35.