DATABASE
The Journal of Biological Databases and Curation

Database, 2016, 1–10
doi: 10.1093/database/baw068
Original article

Original article

BioCreative V CDR task corpus: a resource for chemical disease relation extraction

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] and Zhiyong Lu[3,*]

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

*Corresponding author: E-mail: zhiyong.lu@nih.gov; Tel: (+1) 301-594-7089; Fax: (+1) 301-480-2288

Citation details: Li,J., Sun,Y., Johnson,R.J. *et al.* BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database* (2016) Vol. 2016: article ID baw068; doi:10.1093/database/baw068

Received 25 December 2015; Revised 8 March 2016; Accepted 11 April 2016

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 was found to be sufficient for the task. Thus, we developed our own corpus called BC5CDR during the challenge by inviting a team of Medical Subject Headings (MeSH) indexers for disease/chemical entity annotation and Comparative Toxicogenomics Database (CTD) curators for CID relation annotation. To ensure high annotation quality and productivity, detailed annotation guidelines and automatic annotation tools were provided. The resulting BC5CDR corpus consists of 1500 PubMed articles with 4409 annotated chemicals, 5818 diseases and 3116 chemical-disease interactions. Each entity annotation includes both the mention text spans and normalized concept identifiers, using MeSH as the controlled vocabulary. To ensure accuracy, the entities were first captured independently by two annotators followed by a consensus annotation: The average inter-annotator agreement (IAA) scores were 87.49% and 96.05% 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.

Database URL: http://www.biocreative.org/tasks/biocreative-v/track-3-cdr/

Introduction

Relations between chemicals and diseases (Chemical-Disease Relations or CDRs) play critical roles in drug discovery, biocuration, drug safety, etc. (1). Because of their critical significance, CDRs are being manually curated by resources such as the Comparative Toxicogenomic Database (CTD; http://ctdbase.org) (2,3). 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 (4,5) including the automatic extraction of CDRs (6). These attempts have met with limited success, however, due in part to the lack of a large-scale training corpus. Through BioCreative V in 2015, one of the major formal evaluations for text-mining research (7), a new challenge was organized to advance the state-of-the-art in extraction of CDRs (8). The challenge included two subtasks: disease named entity recognition (DNER) task and 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 (9) for a brief review) including a few specifically targeting diseases (10–12) and chemicals (13), there were none that fulfilled the following content criteria: (i) inclusion of instances of chemical-disease relation annotations that are asserted from both within and across sentence boundaries; (ii) abstracts containing complete chemical, disease and relation annotations; (iii) chemical/disease annotations 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 (14) and drug-drug interaction (15). It is also significantly different from the previously constructed corpora for mining adverse drug reaction/effects in terms of the annotation scope (CID relations), requirements (see above) and size (1500 articles). As shown in Table 1, the EU-ADR corpus contains a total of 300 PubMed articles with 739 drugs, 812 diseases and 300 drug-disease associated relations at sentence level (16). The ADE corpus consists of 2972 PubMed articles with sentence-level statements of 5776 adverse effects related to 5063 drugs (17). The corpus developed by (18) served for disease and adverse effect named entity recognition tasks rather than relation extraction.

Methods and materials

Article selection

We selected a total of 1500 articles for the CDR task, split into three 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 between CTD and Pfizer, and comprises over 150 000 chemical-disease relations from 88 000 articles (19).

To ensure we have some unseen data for the task participants, the remaining 100 articles of the test set were annotated during the challenge (i.e. not selected from the previous CTD-Pfizer corpus) and their curation was not made public until the BioCreative V challenge was complete. We used the following method to select the 100 articles to ensure they would have a similar distribution of words as the training and development sets. For each of the 1000 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 must 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 (20,21) 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 selected an article from the training or development sets, obtained its similarity score, and then selected the new article with the closest similarity score. The

Table 1. Comparison with the previous chemical disease relation corpora

Corpus	Annotation scope	Size	Entity annotation—Mention	Entity annotation—Concept	Relation annotation
BC5CDR	Abstract	1500	Yes	Yes	Yes
EU-ADR (16)	Sentence	300	Yes	Yes	Yes
ADE (17)	Sentence	2972	Yes	No	Yes
Corpus (18)	Abstract	400	Yes	Yes	No

resulting articles are approximately as well related to the articles in the training and development sets, in terms of similarity scoring, as the articles in the training and development sets are related to one another.

Annotation tasks

We performed manual annotation of all chemicals and diseases mentioned in the 1500 articles. For each entity occurrence, we not only annotated its text span but also assigned a relevant concept identifier from MeSH (22). As shown in Figure 1, three diseases mentioned in the abstract were highlighted by our automated tool for potential consideration by the MeSH annotators, 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 1400 of the 1500 articles with few changes: (i) we removed relations that required entities not found in abstracts; (ii) we removed relations that were not disease specific (e.g. 'Drug-Related Side Effects and Adverse Reactions' (D064420)); and (iii) 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

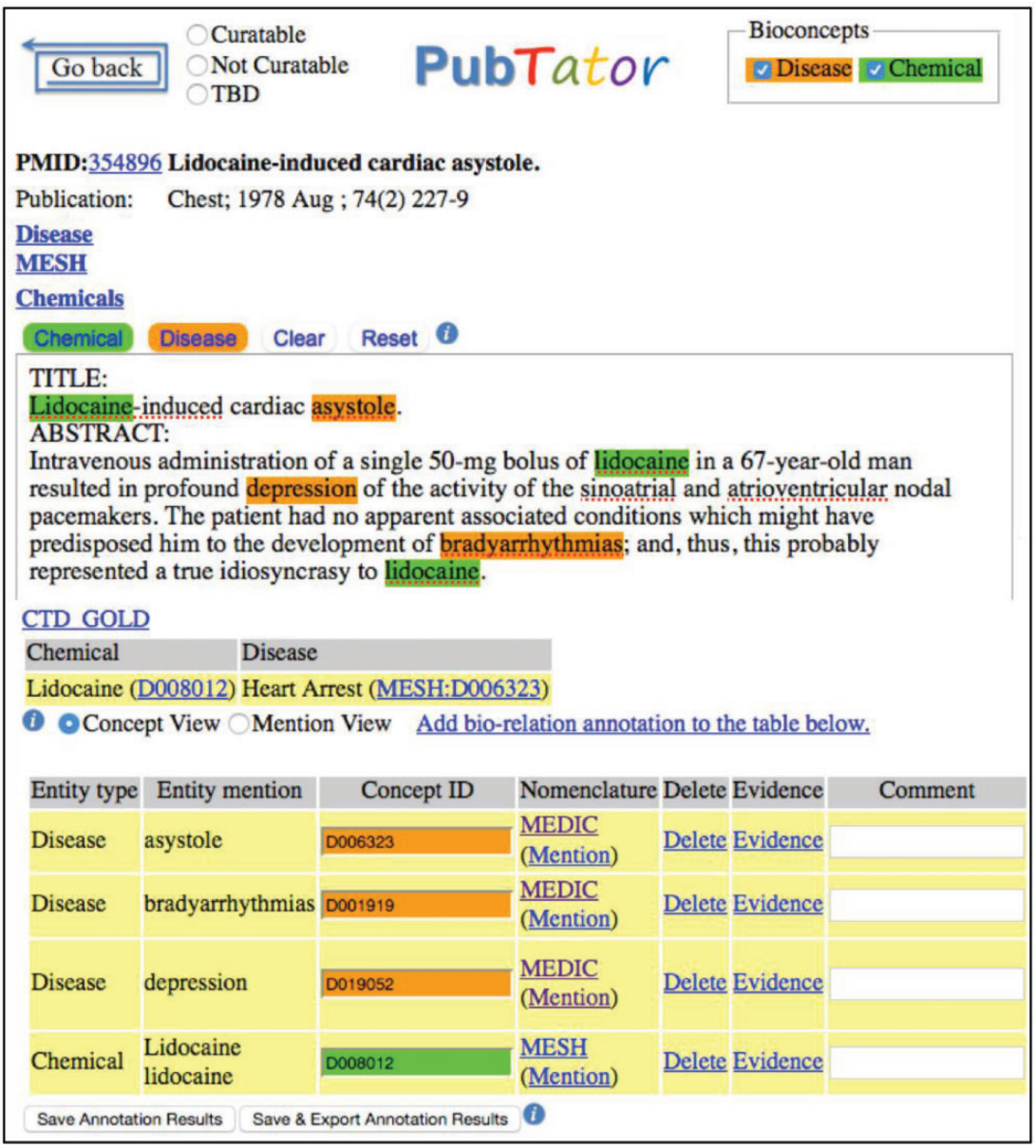


Figure 1. Annotation example shown in our annotation tool, PubTator.

set. For the BioCreative V challenge task, the CID relations refer to two types of relationships between a chemical and a disease in CTD:

- *Putative mechanistic relation* between a chemical and a 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 CTD curated relationship between Lidocaine and Heart Arrest (disease term for the synonym 'asystole' used by the authors in the abstract).
- *Biomarker relation* between a chemical and a 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 (23). Curation was limited to the title and abstract except in cases where reference to the full text was required for clarification; abstracts that required full-text curation were removed from the corpus. In addition to CDR curation, all observed interactions and relationships applicable to CTD were 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.

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 curated the relationships between chemicals and diseases.

Annotation guidelines

The task organizers followed the usual practice of biomedical corpus annotation for entity annotation: the MeSH 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 CID relation annotation, the standard CTD curation protocol was followed (23).

Annotation tools

Manual annotation of disease and chemical entities was performed using PubTator (4,5) (see Figure 1). To accelerate manual annotation (24), text-mined disease and chemical results were pre-computed using DNorm (25) and tmChem (26) 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 (23) was used for all relation curation.

Annotation data formats

All annotated data were made available to participants in both PubTator and BioC formats. The PubTator format consists of a straightforward tab-delimited text file. Figure 2 shows the tab-delimited file for the article (PMID: 354896) in training set. The BioC (27) format is an XML standard recently proposed for biomedical text-mining and data output. For the same article, its BioC format is shown in Figure 3.

```
354896|t|Lidocaine-induced cardiac asystole.
354896|a|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.
354896  0   9   Lidocaine   Chemical    D008012
354896  18  34  cardiac asystole    Disease D006323
354896  90  99  lidocaine   Chemical    D008012
354896  142 152 depression  Disease D003866
354896  331 347 bradyarrhythmias    Disease D001919
354896  409 418 lidocaine   Chemical    D008012
354896  CID D008012 D006323
```

Figure 2. PubTator format annotation (PMID: 354896).

```
<document>
<id>354896</id>
<passage>
<infon key="type">title</infon>
<offset>0</offset>
<text>Lidocaine-induced cardiac asystole.</text>
<annotation id='0'>
<infon key="type">Chemical</infon>
<infon key="MESH">D008012</infon>
<location offset='0' length='9' />
<text>Lidocaine</text>
</annotation>
<annotation id='1'>
<infon key="type">Disease</infon>
<infon key="MESH">D006323</infon>
<location offset='18' length='16' />
<text>cardiac asystole</text>
</annotation>
</passage>
<passage>
<infon key="type">abstract</infon>
<offset>36</offset>
<text>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.</text>
<annotation id='2'>
<infon key="type">Chemical</infon>
<infon key="MESH">D008012</infon>
<location offset='90' length='9' />
<text>lidocaine</text>
</annotation>
<annotation id='3'>
<infon key="type">Disease</infon>
<infon key="MESH">D003866</infon>
<location offset='142' length='10' />
<text>depression</text>
</annotation>
<annotation id='4'>
<infon key="type">Disease</infon>
<infon key="MESH">D001919</infon>
<location offset='331' length='16' />
<text>bradyarrhythmias</text>
</annotation>
<annotation id='5'>
<infon key="type">Chemical</infon>
<infon key="MESH">D008012</infon>
<location offset='409' length='9' />
<text>lidocaine</text>
</annotation>
</passage>
<relation id='R0'>
<infon key="relation">CID</infon>
<infon key="Chemical">D008012</infon>
<infon key="Disease">D006323</infon>
</relation>
</document>
```

Figure 3. BioC format annotation (PMID: 354896).

Table 2. The overall corpus statistics

Task dataset	Articles	Disease		Chemical		CID relation
		Mention	ID	Mention	ID	
Training	500	4182	1965	5203	1467	1038
Development	500	4244	1865	5347	1507	1012
Test	500	4424	1988	5385	1435	1066

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 (28). As shown below, if we defined A as the set of mentions of team A, B as the set of mentions of team B, then the Jaccard agreement score, $S_{A,B}$, could be calculated by counting the number of agreements and disagreements. Mentions with the same PMID, start and end point and concept identifier were counted as a case of agreement. For example, if one annotator annotated 'tardive dystonia' with concept ID of D004421, another annotated '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 relation curation was previously described in (29); no further experiments were conducted in this study.

Results and discussion

Corpus overview

The corpus consists of three separate sets of articles with diseases, chemicals and their relations annotated. The training (500 articles) and development (500 articles) sets were released to task participants in advance to support text-mining method development. The test set (500 articles) was used for final system performance evaluation. As shown in Table 2, 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 table also shows that while there are more chemical mentions than disease mentions in the corpus, there are more disease entities (IDs) than chemical entities (IDs).

Inter-annotator agreement for mention annotation

Table 3 shows the inter-annotator agreement (IAA) scores of three separate subsets for both disease and chemical annotations. The IAA scores over the entire corpus are 87.49% (diseases) and 96.05% (chemicals), which suggests higher agreement on chemical than disease mentions. Additionally, the IAA scores are slightly higher on the test set than the training and test sets.

Table 3. Inter-annotator agreement (IAA) scores of the three sets

Task dataset	Disease	Chemical
Training	0.8600	0.9523
Development	0.8742	0.9577
Test	0.8875	0.9630
All	0.8749	0.9605

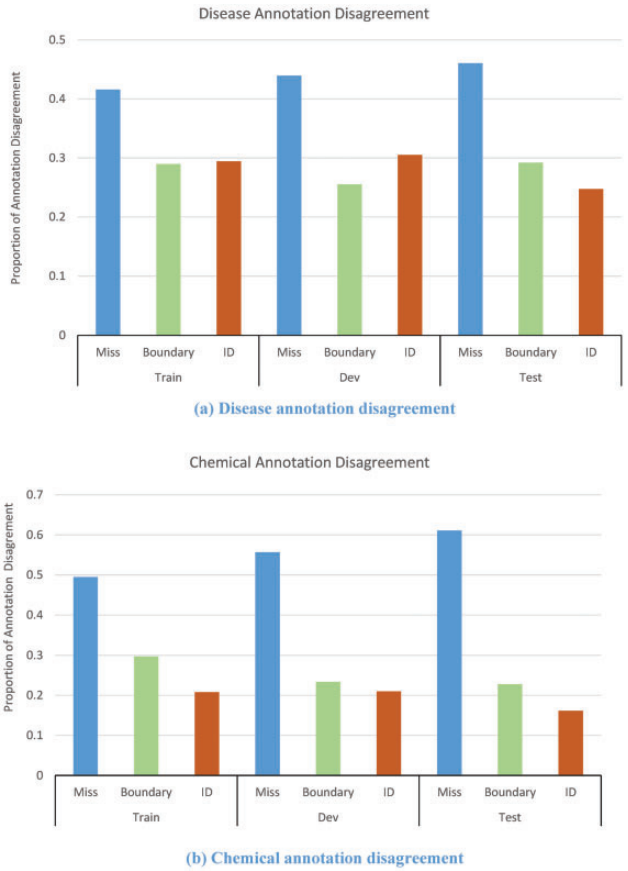
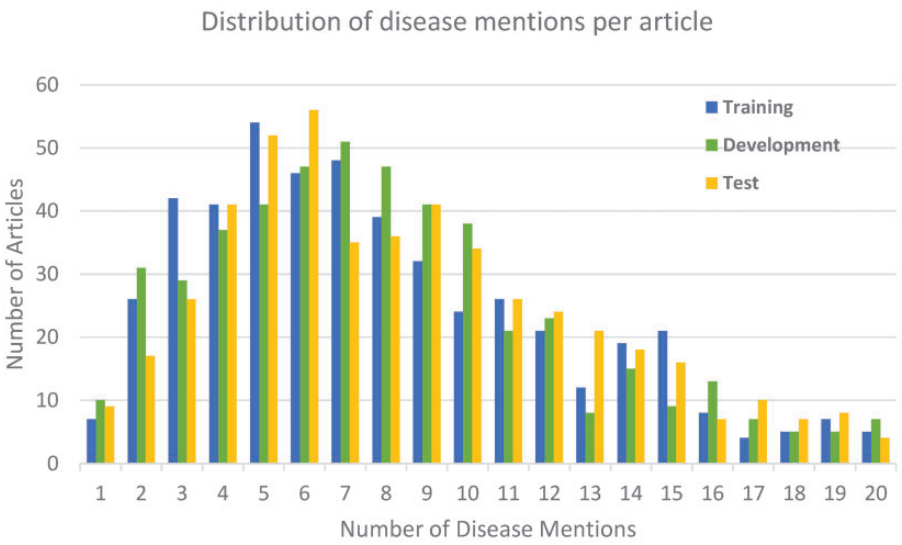
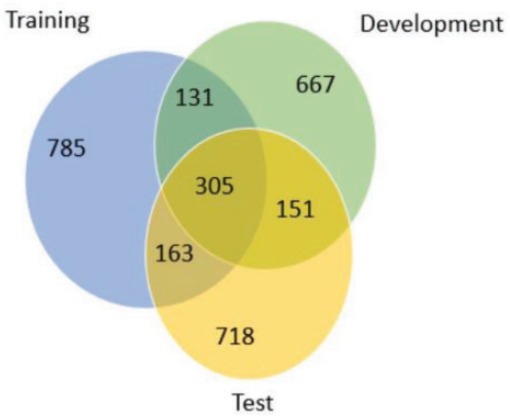


Figure 4.. Disagreements of disease and chemical annotations.

Analyzing the disagreements, we found that many were caused by missed annotation, boundary disagreement and inconsistent identifier assignment. Figure 4 shows the proportion of disease and chemical annotation disagreements in the training, development and test sets, respectively. The most of disagreements (~50%) were due to missed annotation, where one annotator failed to identify the disease/chemical mentions recognized by the other. 28% of disease annotation disagreements were related to the boundary issue. For example, in the article (PMID 20466178) entitled 'Rosaceiform dermatitis associated with topical tacrolimus



(a) Distribution of disease mentions per article



(b) Overlap of disease mentions

Figure 5. Distribution of disease mentions in the corpus.

treatment', it was difficult to judge whether annotate 'rosaceiform dermatitis' as 'rosacea' (MeSH ID: D012393) or simply annotate 'dermatitis' (MeSH ID: D003872).

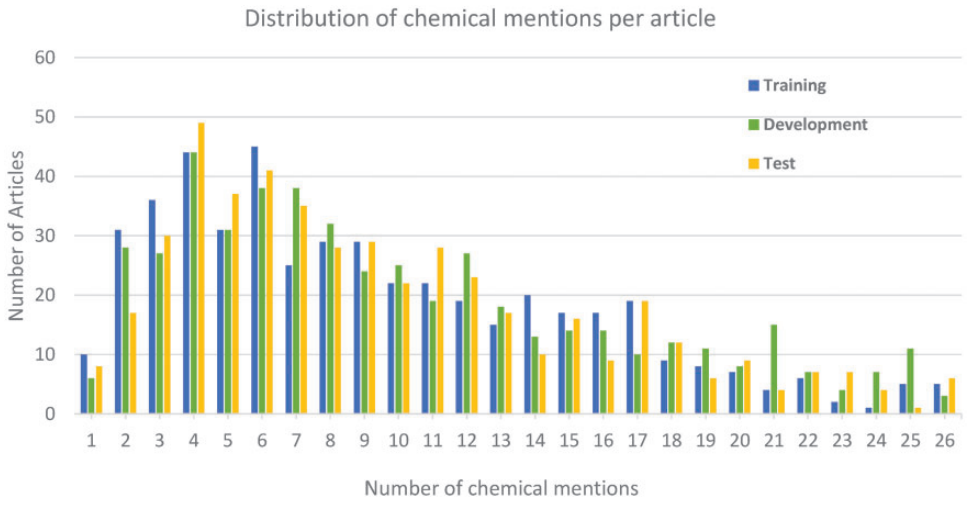
There were also many cases of disagreement over the concept identifier of diseases, especially for the mentions where the text did not exactly match any MeSH term. In some cases, it was hard to judge whether to assign an unknown concept identifier of '-1' or an ancestor concept identifier. For example,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

Disease mention distribution

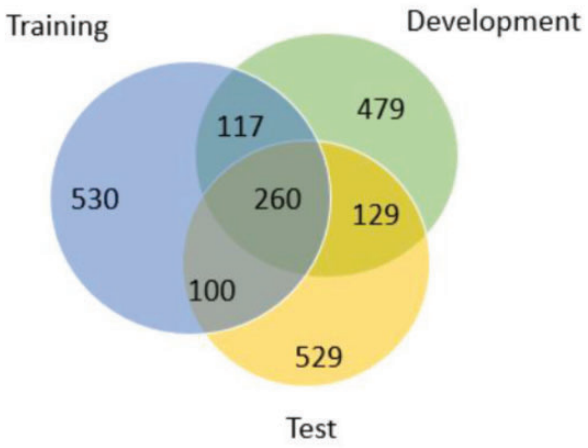
On average, the corpus contains 8.57 non-distinct disease mentions per PubMed abstract. Figure 5(a) shows the breakdown of the number of disease mentions per article in the training, development and test sets, respectively. The three data sets have similar disease mention distribution. In addition, we compared the overlap of unique disease mentions in the three data sets as shown in Figure 5(b). It can be seen that 718 disease mentions out of 1337 in the test set never appear in the training set or the development set. The similar distribution and differential disease mentions in the three sets make the corpus more useful for training models.

Chemical mention distribution

Compared with diseases, the corpus contains more non-distinct chemical mentions (10.62) per PubMed abstract on average. The chemical mention distribution in the three sets (Figure 6(a)) and the overlap of unique chemical



(a) Distribution of chemical mentions per article



(b) Overlap of chemical mentions

Figure 6. Distribution of chemical mentions in the corpus.

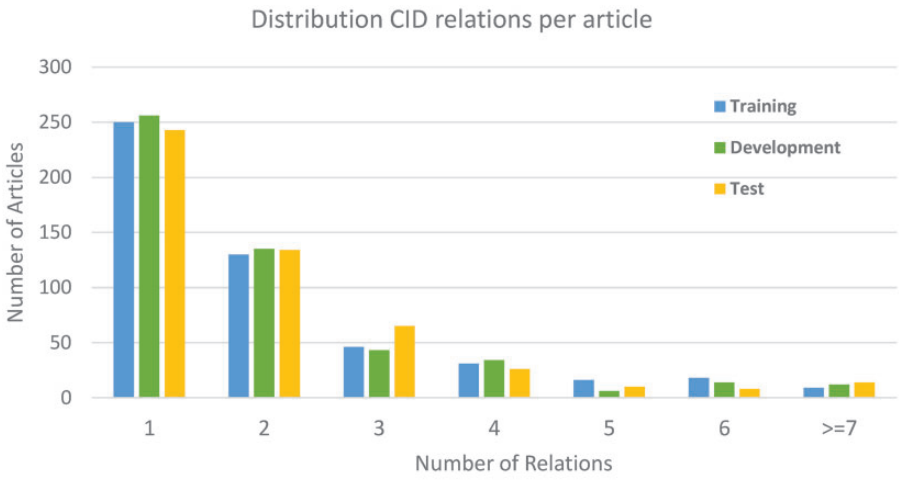
mentions in the three sets (Figure 6(b)) demonstrate that, like the disease mentions, the corpus has excellent characteristics to support model training for chemical entity recognition. Here, we used MeSH identifiers to normalize the chemical mentions because CID relations were previously annotated already in MeSH which is designed for literature indexing and has been used in similar annotation projects (26,30). The CTD chemical vocabulary (http://ctdbase.org/downloads/#allchems) can facilitate mapping the MeSH identifiers to other chemical resource accessions for further chemical-related studies.

CID relation distribution

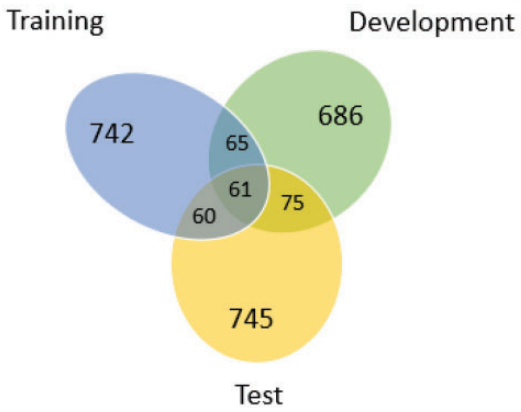
For CID relationships, the average number of relations per PubMed abstract is 2.08. About 50% of articles in the corpus have only one CID relation per article, and 86.8% of articles have no more than three relations (Figure 7(a)). Unlike the disease and chemical mentions, the overlap of unique CID relations in the three sets is as low as 61 relations; 79.17% (745 out of 941) of the relations in the test set have never appeared in the training or development set (Figure 7(b)). This makes relation extraction task more challenging.

Corpus usage in BioCreative V

The corpus successfully supported the BioCreative V Chemical Disease Relation (CDR) task (8,31). A total of 34 teams worldwide participated in the task: 16 teams participated in the in the DNER task, and 18 teams participated in the CID task. As reported in the BioCreative V, the best system performance F-score was 86.46% for the DNER task (32) and 57.03% for the CID task (33). This corpus provides a benchmark set to facilitate further improvement for biomedical text-mining method



(a) Distribution of CID relations per article



(b) Overlap of CID relations

Figure 7. Distribution of chemical-induced disease relations in the corpus.

development, especially as it relates to semantic relationship extraction.

Conclusions

We developed a corpus for both named entity recognition and chemical-disease relations in the literature. A total of 1500 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.

Acknowledgment

We would like to thank Yifan Peng for his help on the manuscript.

Funding

The research was supported by the National Population and Health Scientific Data Sharing Program of China, the China Knowledge Centre for Engineering Sciences and Technology (Medical Centre), the Fundamental Research Funds for the Central Universities (Grant No. 13R0101), the National Institute of Environmental Health Sciences (ES014065 and ES019604). Funding for open access charge: The National Institutes of Health Intramural Research Program.

Conflict of interest. None declared.

References

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

3. Davis,A.P., Grondin,C.J., Lennon-Hopkins,K. *et al.* (2015) The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res.*, 43, D914–D920.
4. Wei,C.H., Kao,H.Y. and Lu,Z. (2013) PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res.*, 41, W518–W522.
5. Wei,C.H., Harris,B.R., 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.
6. Wiegers,T.C., Davis,A.P. and Mattingly,C.J. (2014) Web services-based text-mining demonstrates broad impacts for interoperability and process simplification. *Database (Oxford)*, 2014, bau050.
7. Huang,C.C. and Lu,Z. (2015) Community challenges in biomedical text mining over 10 years: success, failure and the future. *Brief. Bioinf.*, 17, 132–144.
8. Wei,C.H., Pan,Y., Leaman,R. *et al.* (2015) Overview of the BioCreative V Chemical Disease Relation (CDR) Task. In: *Proceedings of the fifth BioCreative challenge evaluation workshop*. pp. 154–166.
9. Neves,M. (2014) An analysis on the entity annotations in biological corpora. *F1000Res*, 3, 96.
10. Leaman,R., Miller,C. and Gonzalez,G. (2009) Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark. In: *The 2009 Symposium on Languages in Biology and Medicine*, Jeju Island, South Korea, pp. 82–89.
11. Doğan,R.I., Leaman,R. and Lu,Z. (2014) NCBI disease corpus: a resource for disease name recognition and concept normalization. *J. Biomed. Inf.*, 47, 1–10.
12. Dogan,R.I. and Lu,Z. (2012) An improved corpus of disease mentions in PubMed citations. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), Montreal, Canada, pp. 91–99.
13. Krallinger,M., Rabal,O., Leitner,F. *et al.* (2015) The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J. Cheminf.*, 7, S2.
14. 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.
15. Herrero-Zazo,M., Isabel,S.-B., Martínez,P. *et al.* (2013) The DDI corpus: an annotated corpus with pharmacological substances and drug–drug interactions. *J. Biomed. Inf.*, 46, 914–920.
16. Mulligen,E.M.V., Fourrier-Reglat,A., Gurwitz,D. *et al.* (2012) The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. *J. Biomed. Inf.*, 45, 879–884.
17. Gurulingappa,H., Rajput,A.M., Roberts,A. *et al.* (2012) Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J. Biomed. Inf.*, 45, 885–892.
18. Gurulingappa,H., Klinger,R., Hofmann-Apitius,M. *et al.* (2010) An Empirical Evaluation of Resources for the Identification of Diseases and Adverse Effects in Biomedical Literature. In: *The 2nd Workshop on Building and evaluating resources for biomedical text mining*. Valetta, Malta.
19. Davis,A.P., Wiegers,T.C., Roberts,P.M. *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.
20. PubMed Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005–. PubMed Help. [Updated 2015 Aug 7]. http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.Computation_of_Similar_Articl
21. Lin,J. and Wilbur,W.J. (2007) PubMed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics*, 8, 423. (2007)
22. Lipscomb,C.E. (2000) Medical subject headings (MeSH). *Bull. Med. Library Assoc.*, 88, 265.
23. Davis,A.P., Wiegers,T.C., Rosenstein,M.C. *et al.* (2011) The curation paradigm and application tool used for manual curation of the scientific literature at the Comparative Toxicogenomics Database. *Database (Oxford)*, 2011, bar034.
24. Névéol,A., Doğan,R.I. and Lu,Z. (2011) Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. *J. Biomed. Inf.*, 44, 310–318.
25. Leaman,R., Doğan,R.I. and Lu,Z. (2013) DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29, 2909–2917.
26. Leaman,R., Wei,C.H. and Lu,Z. (2015) tmChem: a high performance approach for chemical named entity recognition and normalization. *J. Cheminf.*, 7, S3.
27. Comeau,D.C., Doğan,R.I., Ciccarese,P. *et al.* (2013) BioC: a minimalist approach to interoperability for biomedical text processing. *Database (Oxford)*, 2013, bat064.
28. Levandowsky,M. and Winter,D. (1971) Distance between sets. *Nature*, 234, 34–35.
29. Wiegers,T.C., Davis,A.P, Cohen,K.B. *et al.* (2009) Text mining and manual curation of chemical-gene-disease networks for the Comparative Toxicogenomics Database (CTD). *BMC Bioinformatics*, 10, 326.
30. Huang,M., Neveol,A. and Lu,Z. (2011) Recommending MeSH terms for annotating biomedical articles. *J. Am. Med. Inf. Assoc.*, 18, 660–667.
31. Li,J., Sun,Y., Johnson,R.J. *et al.* (2015) Annotating chemicals, diseases and their interactions in biomedical literature. In: *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*. pp. 173-182
32. Lee,H.C., Hsu,Y.Y. and Kao,H.Y. (2015) An enhanced CRF-based system for disease name entity recognition and normalization on BioCreative V DNER Task. In: *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*, pp. 226-233.
33. Xu,J., Wu,Y., Zhang,Y. *et al.* (2015) UTH-CCB@BioCreative V CDR Task: Identifying Chemical-induced Disease Relations in Biomedical Text. In: *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*, pp. 254-259.