The biomedical relationship corpus of the BioRED track at the BioCreative VIII challenge and workshop

Rezarta Islamaj[1], Chih-Hsuan Wei[1], Po-Ting Lai[1], Ling Luo[2], Cathleen Coss[1], Preeti Gokal Kochar[1], Nicholas Miliaras[1], Olga Printseva[1], Oleg Rodionov[1], Keiko Sekiya[1], Dorothy Trinh[1], Deborah Whitman[1], and Zhiyong Lu[1,*]

[1]National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, Maryland, United States of America
[2]School of Computer Science and Technology, Dalian University of Technology, 116024, Dalian, China

*Corresponding author: E-mail: Zhiyong.Lu@nih.gov

Abstract
The automatic recognition of biomedical relationships is an important step in the semantic understanding of the information contained in the unstructured text of the published literature. The BioRED track at BioCreative VIII aimed to foster the development of such methods by providing to the participants the training BioRED corpus, a collection of 600 PubMed documents manually curated for diseases, gene/proteins, chemicals, cell lines, gene variants, and species, as well as pair-wise relationships between them being: disease-gene, chemical-gene, disease-variant, gene-gene, chemical-disease, chemical-chemical, chemical-variant, and variant-variant. Furthermore, relationships are categories into these semantic categories: positive correlation, negative correlation, binding, conversion, drug interaction, comparison, co-treatment, and association. Unlike the previous publicly available corpora, all relationships are expressed at the document level as opposed to the sentence level, and as such they are marked by their corresponding database concept identifiers. As such, diseases and chemicals are normalized to MeSH, genes (and proteins) to NCBI Gene, species to NCBI Taxonomy, cell lines to Cellosaurus, and gene/protein variants to dbSNP. Finally, each annotated relationship is categorized as novel depending on whether it is a novel finding, or experimental verification in the publication it is expressed in, so that it is distinguished from other relationships in the same text that provide known facts and/or background knowledge. The BioRED track at BioCreative VIII further provided 400 newly published articles annotated as above, to serve at the testing data for the challenge. All articles were manually annotated by expert biocurators at the National Library of Medicine (NLM), where each article is doubly annotated in a three round annotation process until full agreement is reached between all curators. This document details the characteristics of this novel resource for biomedical entity and relationship recognition. Using this new resource, we have demonstrated improvements in the biomedical entity and relationship recognition algorithms.

Keywords: corpus annotation; inter-annotator agreement; biomedical entity recognition and normalization, biomedical relation extraction, novel findings, text mining

Introduction
Biomedical entities appear throughout the biomedical research literature, in studies from chemistry, biology, genetics to various other disciplines such as medicine and pharmacology. As such, biomedical entity names are one of the most searched entity types in PubMed (1). Therefore, correctly identifying biomedical entities has a significant impact on information retrieval, helping scientists retrieve the relevant literature, and helping biocurators identify and catalogue into the knowledge sources the structure, uses, and other characteristics and features of each of the entities that could be important to medicine, biology, and human

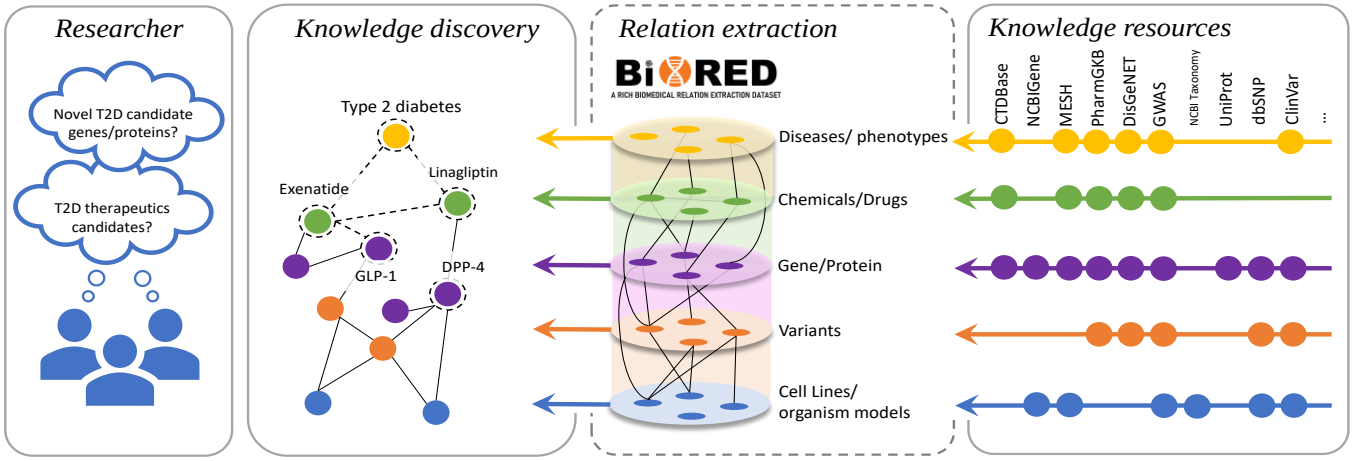


Figure 1. Overview of the relationship extraction in the biomedical domain. This illustration depicts the importance of factual and correct knowledge discovery based on verifiable facts.

health. Furthermore, most research questions directly rely not only on a correct understanding of the structure of these entities but equally important on the understanding of their interactions, their usage, and their implications with respect to disease and human health, as depicted in Figure 1. This is particularly important in the area of large learning models, where developing automated algorithms that correctly capture the semantic relationships contained in the text is crucially important to eliminate hallucinations and other spurious correlations which may have significant problems in the medical and biomedical domains.

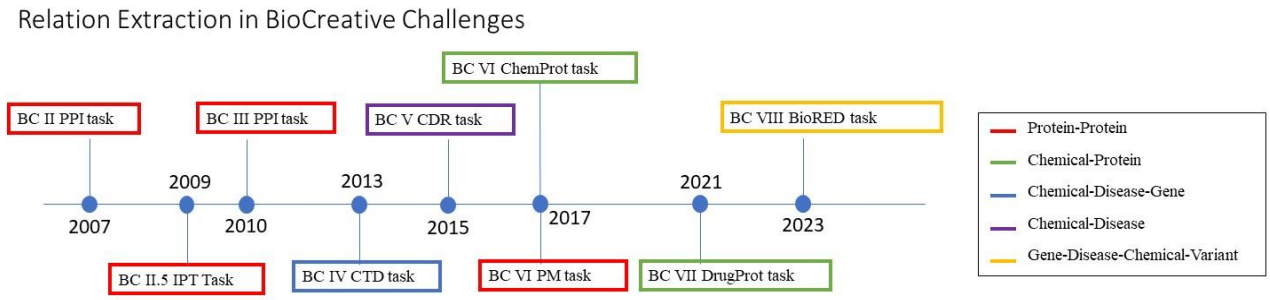


Figure 2 A timeline of BioCreative challenges and the relation extraction community tasks

Relation extraction is a known task in biomedical text mining, and the BioCreative challenges have embraced these tasks since BioCreative II was held in 2007. Figure 2 shows a schematic overview of the historical BioCreative relation extraction challenges for which BioCreative has provided the community with manually annotated resourced to foster development of automated methods targeting these challenges.

The BioCreative II PPI task (2) focused on the automated extraction of protein-protein interaction information from biomedical literature. The BioCreative II.5 IPT task (3) explored the integration of user interaction with automated systems for improved information extraction. Moving forward, the BioCreative IV CTD and BioCreative V CDR tasks (4) delved into identifying chemical-induce-disease relationships. The BioCreative VI PM task (5) centered on precision medicine, aiming to extract relevant information from biomedical texts. In 2021, the BioCreative VII DrugProt task (6) addressed drug-protein interactions in the context of drug discovery.

However, publicly available corpora in the biomedical literature often do not appear to annotate semantic relationships between entities of multiple types, or at the document level. Moreover, often, the annotated entity is not linked to a database concept identifier, and the previously annotated corpora do not distinguish relationships between novel findings and background or previously known associations. Developing relation extraction systems that accurately address these challenges requires a manually annotated corpus of multiple biomedical entity types, and relationships between them, with sufficient examples in the document text and enough articles for system training and an accurate evaluation of their performance. The corpus that supported the BioRED track at BioCreative VIII is a rich corpus, manually annotated to address these challenges.

The BioRED track at BioCreative VIII consisted of two tasks (7):
Sub-task 1: Given the abstract and human-annotated entities, the goal is to identify all the relationships between them in specific types.
Sub-task 2: Given the abstract, the goal is to develop an end-to-end system to identify all the asserted relationships and classify them into specific types. Participants are required to develop their own methods for named entity recognition.

The BioRED initiative started in 2021, with the development of the BioRED corpus (8), a collection of 600 PubMed abstracts pooled from previously annotated corpora (9-13) rich in gene, disease, variant and chemical entity mentions, and manually annotated for all chemical, gene, disease, species, gene variants and species entities mentioned in the abstracts. The manual annotation of entities involved their normalization to the corresponding vocabularies/ontologies, these being: for diseases and chemicals, MeSH(14), for genes and proteins, NCBI Gene (15), for species, NCBI Taxonomy (16), for cell lines, Cellosaurus (17), and for gene/protein variants, dbSNP (18). Furthermore, the manual annotations included the pairwise relationships between those entities, the semantic type of the relationship (i.e., positive correlation, negative correlation, binding, conversion, drug interaction, comparison, cotreatment, and association), and the novelty factor, namely identifying whether the relationship constitutes a novel finding in the article, or whether it is part of the background knowledge. This collection was used in the LitCoin challenge in 2022 (https://ncats.nih.gov/funding/challenges/winners/litcoin-nlp/details) and later on for the development of the BioREx (19).

To support the BioCreative VIII BioRED track, we worked on an even more ambitious endeavor, in that we aimed to equip the BioRED track participants with an adequate and current testing dataset. The testing dataset was selected to be rich in all the above-mentioned biomedical entities and pair-wise relationships, with a similar distribution and composition to the training dataset. Additionally, it was designed to simulate a real-life application, to reflect the current research trends in biomedical literature. Each article was doubly annotated in a three-round annotation process, where annotator discrepancies were discussed after each round until the full consensus was reached. Figure 3 shows an example article in the BioRED-BCVIII corpus, and a snapshot of the entity and relationship annotations for that article.

This resource consists of 1,000 PubMed articles, with no restricting conditions on their sharing and distributions, fully available for the development of algorithms targeting relationship extraction in the biomedical domain, and we expect it to be a significant contribution to this research field.

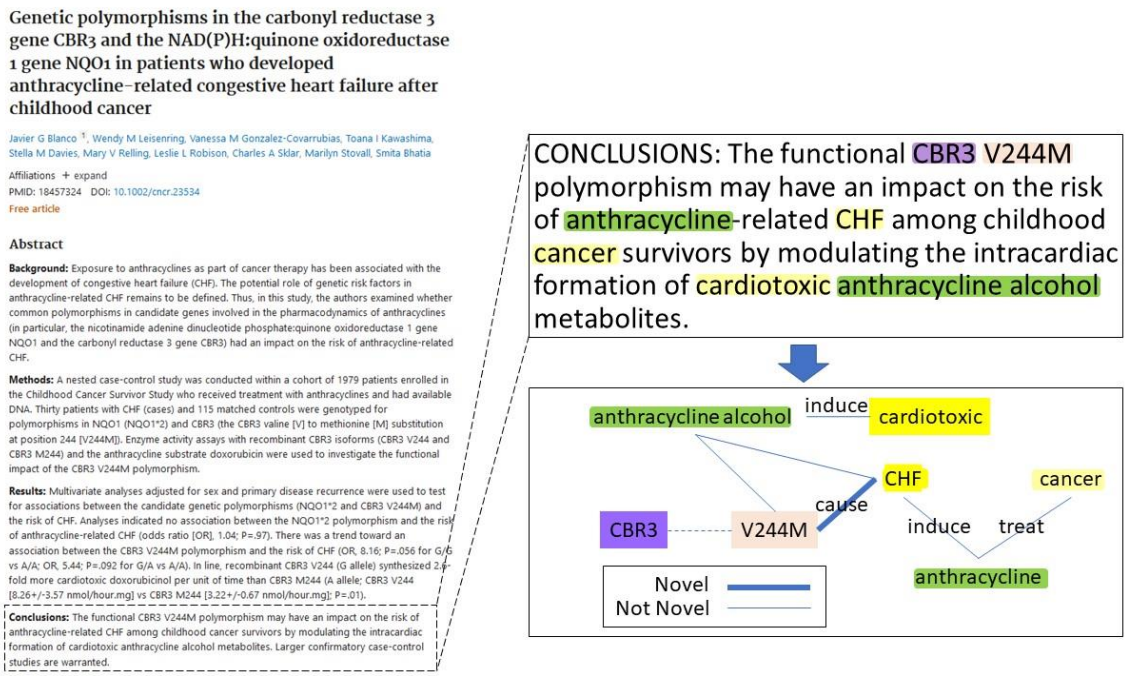


Figure 3 An example of article annotation data in the BioRED-BCVIII corpus

Methods

Document Selection Procedure

The biomedical relationships corpus of the BioRED BioCreative VIII track had these targets:

- be representative of biomedical literature publications that contain biomedical mentions and biomedical relationships.
- target articles that have no restrictions in sharing and distribution
- be instrumental in training biomedical relation extraction algorithms to produce high-quality results even in new, never-before-seen articles.

To select the articles most suitable for algorithm testing, we focused on recently published articles. The new algorithms, in a real-life application setting, would be most valuable for the incoming flux of published literature. As we experienced with the COVID-19 pandemic, correctly identifying what is discussed in the articles, and grouping those articles by the relevant topics, is most crucial, especially in the race to find an effective cure and a timely vaccine.

The 400 PubMed articles that constituted the BioRED track testing set were selected to be as similar as possible to the BioRED original set of 600 articles, to be complementary, balancing, and a suitable test set, that can also serve as a stand-alone corpus (BioRED-BCVIII). The selection criteria included: maximization of journal coverage to assure variety, a similar distribution of all entity mentions and identifiers per article, a similar distribution of all biomedical relationships per article, and similar language models.

Annotation Guidelines

The complete BioRED corpus annotation guidelines are publicly available (8). Here we give a quick summary.

Our guidelines specify which biomedical entities are considered, which entity pairs are considered, which semantic type categories are used for annotation, how to assign the entity pairs to the corresponding semantic category, and how to decide on a relationship novelty factor based on the article context. It also gives specific examples for each case.

Creating high-quality guidelines that fit the annotation task required a multi-step iterative process, starting from an initial draft that was revised until clear and refined guidelines were obtained. The guidelines were prepared by professional expert annotators with degrees in Chemistry, Biochemistry, Biological Sciences, and Molecular Biology, and with ample experience in annotating PubMed literature.

Annotation procedure for the BioRED-BCVIII resource

The BioRED-BCVIII articles are doubly annotated by eight NLM experts in two annotation phases: Phase 1, focusing on the annotation of entities, and Phase 2, focusing on the annotation of relationships. Each phase consisted of three annotation rounds. All manual annotations were performed using the TeamTat annotation tool (20).

For Phase 1, all articles were pre-annotated with entity mentions and their corresponding identifiers using the PubTator Central tools (21), and for Phase 2, all articles we pre-annotated with the BioREx relation extraction method (19). For each phase, in the first round of annotations, each annotator worked independently, reviewed, edited the annotations provided by the applied annotation tool, added new annotations that were missed, or deleted erroneous annotations, as needed. In the second round of annotations, each annotator reviewed all assigned articles that included their own annotations and their partner's annotations of round 1. In the third annotation round, the annotator identities were revealed to their partners, and they discussed each article until they reached 100% consensus.

In addition, to further facilitate the work of the annotators, during the relationship annotation phase, a specific visualization tool was built to help visualize the differences in the annotation of the relationships as suggested by each annotator. This visualization showed the pair of entities in a relationship, the relationship type and the novelty factor for each annotated relationship and included visual cues to identify the agreement between annotators and remaining differences.

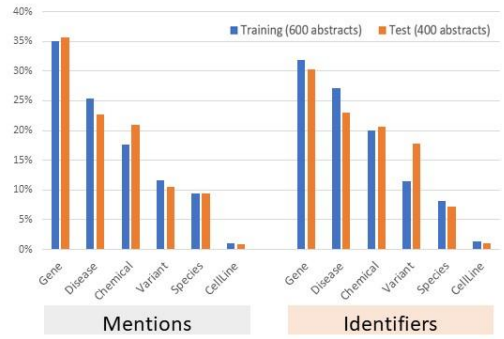
Corpus Document Format

While annotations can be represented in various formats, the BioRED-BCVIII dataset is available in: the BioC (22) (XML and JSON) format, and the PubTator format. These formats were chosen because they provided the following advantages: they support annotations representing both mention span (location) and entity identifier, articles in the PMC text mining subset (23) are available in the BioC format, the TeamTat tool uses the BioC format, the text mining tools we used for pre-annotation support these formats, PubTator Central API allows retrieval of any PubMed/PMC Open access article with precomputed annotations in any of these formats, and finally, these formats are simple and easy to modify, allowing additional analysis tools to be applied rapidly as needed.

Results and Discussion

Corpus characteristics

The BioRED-BCVIII track relationship extraction resources are rich in manual bio-entity mention and relationship annotations and currently the largest corpus, compatible with previously annotated corpora, targeted for developing similar text mining tools. The BioRED training dataset of 600 PubMed articles contains 20,419 manual biomedical mention annotations; corresponding to 6,502 manually annotated pair-wise relationships. The BioRED testing dataset of 400 PubMed articles contains 15,400 manual biomedical mention annotations; corresponding to 6,034 manually annotated pair-wise relationships.



	Mentions		Identifiers	
	Training	Test	Training	Test
Gene	6,697	5,715	2,346	1,713
Disease	5,545	3,641	1,949	1,332
Chemical	4,429	2,592	1,287	938
Variant	1,381	1,770	702	1,069
Species	2,192	1,507	659	427
CellLine	175	140	90	61

Type	Training	Training ∩ Test	Test	∩/Test
Disease	531	281	352	44.39%
Species	34	13	20	39.39%
Chemical	487	171	438	28.08%
CellLine	62	10	40	20.00%
Variant	658	20	955	2.05%
Gene	1390	308	907	25.35%

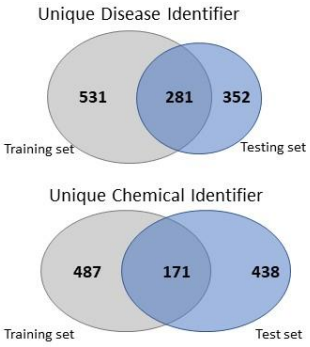


Figure 4 Entity Composition in the BioRED training and testing datasets.

Figures 4 and 5 illustrate the entity and the relationship composition of the BioRED training and testing datasets and demonstrate that the resources are: 1) compatible – to foster re-use, acknowledge and build on previous efforts of experts, and 2) complementary – to expand on previous knowledge and cover new areas of training data.

Entity pair	Training	Test
Gene, Gene	1,227	936
Variant, Chemical	76	199
Variant, Variant	25	2
Gene, Chemical	923	1,121
Disease, Chemical	1,237	779
Gene, Disease	1,633	1,610
Chemical, Chemical	488	412
Variant, Disease	893	975

Relation type	Train	Test
Association	3,387	2,759
Positive Correlation	1,766	1,751
Negative Correlation	1,150	1,192
Cotreatment	55	172
Drug Interaction	13	-
Conversion	4	13
Comparison	39	13
Bind	89	136

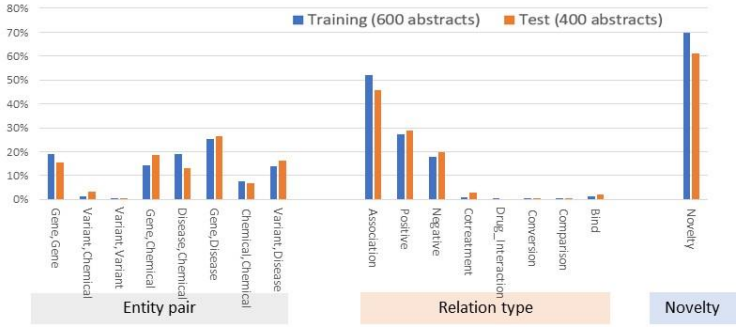


Figure 5 Relationship composition in the BioRED training and testing datasets.

Corpus technical validation

Table 1 shows the results of our benchmark method on Sub-task 1. Our benchmark is based on our previously published method with the original BioRED corpus. The first implementation was trained on 500 articles as the training data and tested on 100 BioRED articles designated as the testing data. This was the original LitCoin challenge setting. The second implementation used the complete 600 articles as the training

dataset and the BioCreative-BC8 BioRED testing data as the testing dataset (400 articles). The new testing data consists of recently published articles, and contains previously unseen entities, so differences in performances are expected. As expected, the performance is robust, while leaving room for further improvement.

Table 2 shows similar results on Sub-task 2. For this task, we needed to first identify the entities mentioned in the title and abstracts in the articles in the testing dataset, and to perform this step we used two tools: the current PubTator Central, and the newest available tool PubTator 3. As displayed, better entity recognition predictions result in better relation extraction performance. As expected, the BioRED-BCVIII testing dataset is an adequate testing dataset that complements the BioRED training dataset.

Table 1: Sub-task 1 results when applying our PubMedBERT-based benchmark method on the BioRED (2021 version) test set, and the BioRED at BioCreative VIII testing dataset.

	Entity pair			**Entity pair +Relation type**			**Entity pair +Novelty**			**ALL**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
BioRED test	76.13	75.41	75.77	58.51	57.95	58.23	64.58	63.97	64.28	48.52	48.07	48.29
BioRED-BCVIII test	72.03	76.71	74.29	50.34	53.61	51.93	54.54	58.08	56.25	37.71	40.16	38.90

Table 2: Sub-task 2 results when applying our PubMedBERT based benchmark method on the BioRED (2021 version) test set, and the BioRED at BioCreative VIII testing dataset. Because in Sub-task 2 we are not given the entity annotations, we use the PubTator Central and PubTator 3 tool outputs as our entity predictions.

	Entity pair			**Entity pair +Relation type**			**Entity pair +Novelty**			**ALL**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
PubTator Central – BioRED test	55.06	33.71	41.81	43.26	26.48	32.85	43.26	26.48	32.85	32.58	19.95	24.75
PubTator Central– BioRED-BCVIII	45.28	37.12	40.80	32.10	26.32	28.92	33.96	27.84	30.60	24.12	19.77	21.73
PubTator3 – BioRED test	57.40	41.36	48.08	44.99	32.42	37.68	46.78	33.71	39.18	35.92	25.88	30.08
PubTator3 – BioRED-BCVIII	55.49	47.98	51.46	39.67	34.30	36.79	41.69	36.05	38.66	29.76	25.73	27.60

Conclusions

The relationship corpus for the BioRED track at BioCreative VIII is a high-quality corpus and consists of: 1) The training dataset, previously published as BioRED, that contains 600 PubMed abstracts, and the testing dataset, that contains 400 recently published PubMed abstracts and specifically annotated for the BioCreative VIII challenge. The BioRED-BCVIII corpus thus contains 1000 PubMed articles, fully annotated for diseases, genes, chemicals, species, gene variants, and cell lines. All entities are also annotated with their corresponding database identifiers: NCBI gene for genes and proteins, NCBI taxonomy for species, dbSNP for gene/protein variants, MeSH for diseases and chemicals, and Cellosaurus for cell lines. Furthermore, all articles are fully annotated for pair-wise relationships between disease-gene, chemical-gene, disease-variant, gene-gene, chemical-disease, chemical-chemical, chemical-variant, and variant-variant, and their semantic categories: positive correlation, negative correlation, binding, conversion, drug

interaction, comparison, cotreatment, and association. Finally, each relationship is categorized for novelty, distinguishing whether that relationship is a significant finding for that article, or whether it describes previously known facts or background knowledge. The BioRED-BCVIII corpus is a gold-standard corpus for biomedical relationship extraction in biomedical articles.

Gold-standard data is crucial for the development of robust models, and we demonstrated that in the BioRED-BCVIII corpus both training and testing datasets have matching entity composition and distributions, and matching relationship composition and distributions. The training and testing datasets introduce new, previously unseen elements, and they complement each other to create a larger, richer gold-standard dataset for semantic relation extraction in biomedical journals. We tested the new corpus with our best performing relation extraction tool, and for Sub-task 2 we paired it with both PubTator Central and PubTator 3 predicted entities.

The BioRED-BCVIII relationship resource provides these contributions: 1) High-quality manual annotation of six types of biomedical entities, 2) All entity mentions are normalized to standardized vocabularies and ontologies, which via UMLS, can be easily mapped to other terminologies, if needed, 3) document-level semantic annotation of pair-wise relationships, and 4) novel relations based on article context. Annotations were performed by the expert literature biocurators at the National Library of Medicine. The annotation guidelines are compatible with previously annotated corpora; therefore, these corpora can be used as additional data. This resource will be invaluable for advancing text-mining techniques for biomedical relation extraction tasks in biomedical text.

Acknowledgment

This research was supported by the NIH Intramural Research Program, National Library of Medicine. It was also supported by the Fundamental Research Funds for the Central Universities [DUT23RC(3)014 to L.L.].

Reference

1. Islamaj Dogan, R., Murray, G.C., Névéol, A. and Lu, Z. (2009) Understanding PubMed® user search behavior through log analysis. *Database*, **2009**, bap018.
2. Krallinger, M., Leitner, F., Rodriguez-Penagos, C. and Valencia, A. (2008) Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome biology*, **9**, 1-19.
3. Leitner, F., Mardis, S.A., Krallinger, M., Cesareni, G., Hirschman, L.A. and Valencia, A. (2010) An overview of BioCreative II. 5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **7**, 385-399.
4. Wei, C.-H., Peng, Y., Leaman, R., Davis, A.P., Mattingly, C.J., Li, J., Wiegers, T.C. and Lu, Z. (2016) Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. *Database*, **2016**, baw032.
5. Islamaj Doğan, R., Kim, S., Chatr-Aryamontri, A., Wei, C.-H., Comeau, D.C., Antunes, R., Matos, S., Chen, Q., Elangovan, A. and Panyam, N.C. (2019) Overview of the BioCreative VI Precision Medicine Track: mining protein interactions and mutations for precision medicine. *Database*, **2019**, bay147.
6. Miranda-Escalada, A., Mehryary, F., Luoma, J., Estrada-Zavala, D., Gasco, L., Pyysalo, S., Valencia, A. and Krallinger, M. (2023) Overview of DrugProt task at BioCreative VII: data and methods for large-scale text mining and knowledge graph generation of heterogenous chemical–protein relations. *Database*, **2023**.
7. Islamaj, R., Lai, P.-T., Wei, C.-H., Luo, L. and Lu, Z. (2023), *Proceedings of the eighth BioCreative challenge evaluation workshop 2023*, New Orleans, LA.
8. Luo, L., Lai, P.-T., Wei, C.-H., Arighi, C.N. and Lu, Z. (2022) BioRED: A Rich Biomedical Relation Extraction Dataset. *Briefings in Bioinformatics*, **23**, bbac282.
9. Doğan, R.I., Leaman, R. and Lu, Z. (2014) NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, **47**, 1-10.
10. Islamaj, R., Wei, C.-H., Cissel, D., Miliaras, N., Printseva, O., Rodionov, O., Sekiya, K., Ward, J. and Lu, Z.J.J.o.b.i. (2021) NLM-Gene, a richly annotated gold standard dataset for gene entities that addresses ambiguity and multi-species gene recognition. **118**, 103779.

11. Wei, C.-H., Kao, H.-Y. and Lu, Z. (2015) GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed research international*, **2015**.
12. Li, J., Sun, Y., Johnson, R.J., Sciaky, D., Wei, C.-H., Leaman, R., Davis, A.P., Mattingly, C.J., Wiegers, T.C. and Lu, Z. (2016) BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, **2016**.
13. Wei, C.-H., Allot, A., Riehle, K., Milosavljevic, A. and Lu, Z. (2022) tmVar 3.0: an improved variant concept recognition and normalization tool. *Bioinformatics*, **38**, 4449-4451.
14. Rogers, F.B. (1963) Medical subject headings. *Bull Med Libr Assoc*, **51**, 114-116.
15. Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., Funk, K., Kelly, C. and Kim, S. (2022) Database resources of the national center for biotechnology information. *Nucleic acids research*, **50**, D20.
16. Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., McVeigh, R., O'Neill, K., Robbertse, B. *et al.* (2020) NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)*, **2020**.
17. Bairoch, A. (2018) The Cellosaurus, a Cell-Line Knowledge Resource. *J Biomol Tech*, **29**, 25-38.
18. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, **29**, 308-311.
19. Lai, P.-T., Wei, C.-H., Luo, L., Chen, Q. and Lu, Z. (2023) BioREx: Improving Biomedical Relation Extraction by Leveraging Heterogeneous Datasets. *Journal of biomedical informatics*, **146**, 104487.
20. Islamaj, R., Kwon, D., Kim, S. and Lu, Z. (2020) TeamTat: a collaborative text annotation tool. *Nucleic acids research*, **48**, W5-W11.
21. Wei, C.-H., Allot, A., Leaman, R. and Lu, Z. (2019) PubTator central: automated concept annotation for biomedical full text articles. *Nucleic acids research*, **47**, W587-W593.
22. Comeau, D.C., Islamaj Dogan, R., Ciccarese, P., Cohen, K.B., Krallinger, M., Leitner, F., Lu, Z., Peng, Y., Rinaldi, F., Torii, M. *et al.* (2013) BioC: a minimalist approach to interoperability for biomedical text processing. *Database (Oxford)*, **2013**, bat064.
23. Comeau, D.C., Wei, C.-H., Islamaj Doğan, R. and Lu, Z. (2019) PMC text mining subset in BioC: about three million full-text articles and growing. *Bioinformatics*, **35**, 3533-3535.