The overview of the BioRED (Biomedical Relation Extraction Dataset) track at BioCreative VIII

Rezarta Islamaj[1], Po-Ting Lai[1], Chih-Hsuan Wei[1], Ling Luo[2] and Zhiyong Lu[1, *]

[1]National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), MD, 20894, Bethesda, USA
[2]School of Computer Science and Technology, Dalian University of Technology, 116024, Dalian, China

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

Abstract

The BioRED track at BioCreative VIII calls for a community effort to identify, semantically categorize, and highlight the novelty factor of the relationships between biomedical entities in unstructured text. Relation extraction is crucial for many Natural Language Processing (NLP) applications, from drug discovery to custom medical solutions. While previous community challenges focused on identifying relationships of a single type (i.e., protein-protein interactions), categorized relationships into different semantic categories but did not require entity normalization, or worked at the sentence level, real-world applications necessitate that entities are linked to specific knowledge base records, relationships encountered in any given document generally occur between different entity types, and perusal of the whole document provides valuable additional detail. In addition, journal publications often distinguish between novel findings and background information or prior knowledge.

The BioRED track consisted of two sub-tasks: 1) in sub-task 1, participants were given the article text and human expert annotated entities, and were asked to extract binary relation pairs, identify their semantic type and the novelty factor, and 2) in sub-task 2, participants were given only the article text, and were asked to build an end-to-end system that could identify and categorize the relationships and their novelty.

We received a total of 94 submissions from 14 teams worldwide. For each submission, we calculated four evaluations: relation identification computed whether the correct pair of entities was identified in an entity pair relationship, relation type computed whether the entity pair was categorized with the right relation type, novelty identification computed whether the entity pair was in a novel relationship based on the article context, and relation extraction (all) computed whether the correct entity pair was predicted with the correct relation type, and with the correct novelty factor.

The highest performance achieved for the sub-task 1 was 77.17% F-score when evaluating relation identification, 58.95% F-score when evaluating relation type, 59.22% F-score when evaluating novelty identification and 44.55% F-score when evaluating all of the above aspects of relation extraction. The highest performance achieved for the sub-task 2 was 55.84% F-score when evaluating relation identification, 43.03% F-score when evaluating relation type, 42.74% F-score when evaluating novelty identification and 32.75% F-score when evaluating all of the

above aspects of relation extraction. The BioRED track dataset and other challenge materials are available at (https://ftp.ncbi.nlm.nih.gov/pub/lu/BC8-BioRED-track/) and (https://codalab.lisn.upsaclay.fr/competitions/13377 and https://codalab.lisn.upsaclay.fr/competitions/13378)

This community challenge demonstrated 1) current substantial achievements in the large learning model technologies can be utilized to further improve automated prediction accuracy, but additional research is necessary to incorporate these models to their full potential, and 2) developing and end-to-end system is substantially more challenging. We look forward to further development of biomedical text mining methods for relation extraction, and we invite the community to utilize the BioRED-BCVIII corpus of 1,000 PubMed documents fully manually annotated for biomedical entities and the relationships between them.

Introduction

Biomedical relation extraction plays a pivotal role in discerning and categorizing relationships between biomedical concepts from textual data. This task stands central to biomedical natural language processing (NLP), fostering advancements in areas like drug discovery and personalized medicine. Previous work in relation extraction includes several community challenges at previous BioCreative workshops, such as the protein-protein interactions challenges in BioCreative II and III (1,2), the CTD challenges in BioCreative IV and V (3,4), precision medicine task in BioCreative VI (5), and drug-chemical challenges in BioCreative VI and VII (6,7). As noted, these challenges, as well as other publicly available benchmark datasets for relation extraction are limited to single-type relations often confined to a single sentence or short context.

The recently published BioRED dataset (8) encompassed a broader coverage than the previous corpora, in that it captured multiple entity types (like gene/protein, disease, and chemical) and document-level relation pairs amongst five types of entities (such as, gene–disease and chemical–chemical). Furthermore, beyond relation extraction, BioRED also included annotations for novel findings, distinguishing relations that constitute the main contributions of an article from pre-existing knowledge. However, the dataset consisted of historical articles focused on genes, diseases, chemicals, and others sampled from previously published relation extraction datasets.

During the COVID-19 pandemic the world witnessed the medical research in search for a cure, treatment and vaccine that can help ease the suffering worldwide. For the research community it presented the challenge of recognizing new, previously unseen biomedical entities in previously unseen context. Therefore, it is imperative to build accurate automatic methods that can robustly identify relations in previously unseen contexts.

The BioRED track at BioCreative VIII challenge constitutes of two distinct sub-tasks. Sub-task 1: Here, participants are encouraged to build automatic tools that can read a document (journal title and abstract) and the pre-annotated entities, and identify the pairs of entities in a relationship, the type of the relationship and its novelty factor. Sub-task 2: The objective broadens to an end-to-end system that is only provided with the text of the document, and requires the detection of the asserted relationships, their type and novelty factor.

This track expanded the BioRED dataset to 1,000 PubMed articles, by using the complete previous BioRED as training data, while providing an additional 400 articles published between September 2022 and March 2023 as a testing dataset.

Material and Methods

The BioRED-BCVIII Dataset

The BioRED-BCVIII dataset spans a variety of journals, are rich in the coverage of biomedical entities, cover a plethora of biomedical-related topics to be representative of biomedical literature publications that contain relationships between biomedical entities. We describe the BioRED-BCVIII dataset in detail in (9), but here we give a brief overview.

The original BioRED dataset is provided as the training dataset. To help the challenge participants, we setup two leaderboards (10,11) for system development at the CodaLab site, where the dataset is separated into training (500 articles) and validation (100 articles) datasets. Because the BioRED dataset is publicly available, we masked the article identifiers for all the articles, and the validation set articles were hidden amongst a set of 1,000 PubMed articles. The evaluation script was carefully tuned to only assess the validation set articles when computing the statistics to rank the teams. The challenge data was provided in three formats: PubTator, BioC-XML, and BioC-JSON (12), and participant teams could use any of those.

The test dataset articles were specifically selected so that the articles 1) have no restrictions on sharing and distribution, 2) are useful for biomedical entity and relation extraction, and other related text mining tasks, and 3) are suitable for testing real-world tasks, therefore focused on recently published articles.

For the BioRED track evaluation, we selected PubMed documents published in the recent six months before the challenge. We anticipated that these abstracts would reflect current trends in biomedical research. We collaborated with eight NLM biocurators (with 20 years biomedical indexing as their profession on average) to manually annotate 400 abstracts with biomedical entities linked to their corresponding database identifiers (gene/proteins normalized to NCBI GENE (13), diseases normalized to MeSH (14), chemicals to MeSH, gene variations to dbSNP (15), species to NCBI Taxonomy (16), and cell lines to Cellosaurus (17)) and all binary relationships between them, as specified in the BioRED annotation guidelines. Because all relationships are defined at the document level, each entity is identified with its corresponding database identifier, as opposed to its text mention. All articles were doubly annotated, and the corpus annotation was conducted in two phases: Phase 1) the annotation of all biomedical entities (text span and database identifiers) in a three-round annotation process, and then, after all articles were re-shuffled and re-distributed to pairs of biocurators, Phase 2) the annotation of biomedical relationships (semantic type and novelty factor) in a three-round annotations process. In Table 1 we show the data statistics of the BioRED track train and test dataset.

Table 1: Data statistics for the BioRED track train and testing dataset. For each element, we list total and unique numbers.

Annotations	BioRED train	BioRED test
Documents	600	400
Gene	6,697 (1,643)	5,728 (1,278)
Disease	5,545 (778)	3,641 (644)
Chemical	4,429 (651)	2,592 (618)
Variant	1,381 (678)	1,774 (974)
Species	2,192 (47)	1,525 (33)
Cell Line	175 (72)	140 (50)
All entities	20,419 (3,869)	15,400 (3,597)
Disease-Gene	1,633	1,610
Chemical-Gene	923	1,121
Disease-Variant	893	975
Gene-Gene	1,227	936
Chemical-Disease	1,237	779
Chemical-Chemical	488	412
Chemical-Variant	76	199
Variant-Variant	25	2
All pairs	6,502	6,034
Pairs describing novel relationships for the context as opposed to background knowledge	4,532	3,683

To ensure test data integrity, we compiled a large dataset of 10,000 documents, and the manually annotated 400 documents were concealed within this large set. Only the 400 articles were used to compute the evaluation metrics.

Benchmarking systems

In our previous work (18), we described an improved relationship extraction tool for biomedical entity recognition. This tool was based on a BERT model with data adjustment rules. BioREx proposes a data-centric approach that bridges annotation discrepancies between data sources and amalgamates them to construct a rich dataset of adequate size. Despite the diversity in annotation scope and guidelines, BioREx systematically consolidates disparate annotations into one large-scale dataset.

Intrigued with the prospect of testing these previous models on a real-world dataset, we evaluated both systems on the BioRED-BCVIII testing dataset. In addition, with the rise of large language model systems, we were also eager to test the GPT system for biomedical relation extraction. We describe in detail these three systems in (19) and here we give a summary in Table 2 and Table 3.

Evaluation Measures

The evaluation metrics used to assess team predictions were micro-averaged recall, precision and F-score (main evaluation metric). Three different result types were scored: False negative (FN) results corresponding to incorrect negative predictions; False positive (FP) predictions corresponding to incorrect positive predictions and True positive (TP) results corresponding to

Table 2: Results of benchmarking systems for Sub-Task 1 (in %).

Sub-Task 1	**Entity pair**			**Entity pair +Relation type**			**Entity pair +Novelty**			**ALL**		
Method	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
GPT 3.5	23.07	99.35	37.44	8.26	35.58	13.41	14.06	60.57	22.83	5.36	23.08	8.70
GPT 4	27.57	96.88	**42.93**	13.39	47.05	20.85	19.52	68.60	30.40	9.92	34.85	15.44
PubMedBERT	72.03	76.71	74.29	50.34	53.61	51.93	54.54	58.08	56.25	37.71	40.16	38.90
BioREx	76.83	74.56	**75.68**	57.76	56.05	56.89	-	-	-	-	-	-

Table 3: Results of benchmarking systems for Sub-Task 2. For Sub-Task 2, since we are not given manual annotation of entities, we used the PubTator Central or PubTator 3 to retrieve these predictions: (2) stands for PubTator Central, and (3) stands for PubTator 3 (in %).

Sub-Task 2	**Entity pair**			**Entity pair +Relation type**			**Entity pair +Novelty**			**ALL**		
Method	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
GPT 3.5 (3)	18.93	74.71	30.21	6.78	26.75	10.81	11.74	46.32	18.73	4.47	17.65	7.14
GPT 4 (3)	22.56	73.22	**34.49**	11.14	36.16	17.04	16.15	52.40	24.69	8.35	27.10	12.77
PubMedBERT(2)	45.28	37.12	40.80	32.10	26.32	28.92	33.96	27.84	30.60	24.12	19.77	21.73
PubMedBERT(3)	55.49	47.98	51.46	39.67	34.30	36.79	41.69	36.05	38.66	29.76	25.73	27.60
BioREx(2)	48.91	39.00	43.40	36.79	29.34	32.64	-	-	-	-	-	-
BioREx(3)	59.26	49.49	**53.94**	44.65	37.29	40.64	-	-	-	-	-	-

correct predictions. Recall R = TP/(TP + FN). Precision P = TP/(TP+FP). The F-measure F = 2· P · R / (P + R).

We measure the precision, recall, and F-scores for these settings: 1) Entity pair: to evaluate the correct pairs of entities that are in a relationship, 2) Entity pair + Relationship type: once the correct pairs have been identified, to evaluate the correct semantic relationship that the two entities engage in, and for novelty, 3) Entity pair + Novelty: once the correct pairs have been identified to evaluate the correct pairs in a novel relationship, and 4) All: once the correct pairs, and the correct semantic relationship type have been identified, to evaluate whether the novelty factor. For Sub-Task 2, all these evaluations require that the entities are correctly identified and normalized to their corresponding database identifiers.

The evaluation script was made available to all track participants, together with the data and other challenge materials via: FTP:http://ftp.ncbi.nlm.nih.gov/pub/lu/BC8-BioRED-track; CodaLab: https://codalab.lisn.upsaclay.fr/competitions/13377 and https://codalab.lisn.upsaclay.fr/competitions/13378.

Team Invitations and Challenge Participation

We announced the BioRED track at BioCreative VIII in Spring 2023. The BioRED corpus as the training dataset, and the codalab evaluation website containing challenge description, dataset, evaluation script and leaderboard were made available in May 2023. A webinar was held in June 2023 for interested teams to introduce them to the challenge motivation and data collection. The testing dataset, which complements the BioRED corpus, was manually annotated during April-August 2023.

Fourteen teams submitted a total of 56 runs for Sub-task 1, 19 of which were submitted after the deadline, and were considered unofficial. For Sub-task 2, nine teams submitted a total of 38 runs, 9 of which were considered unofficial because they were submitted after the deadline. Team participation is illustrated in Figure 1.

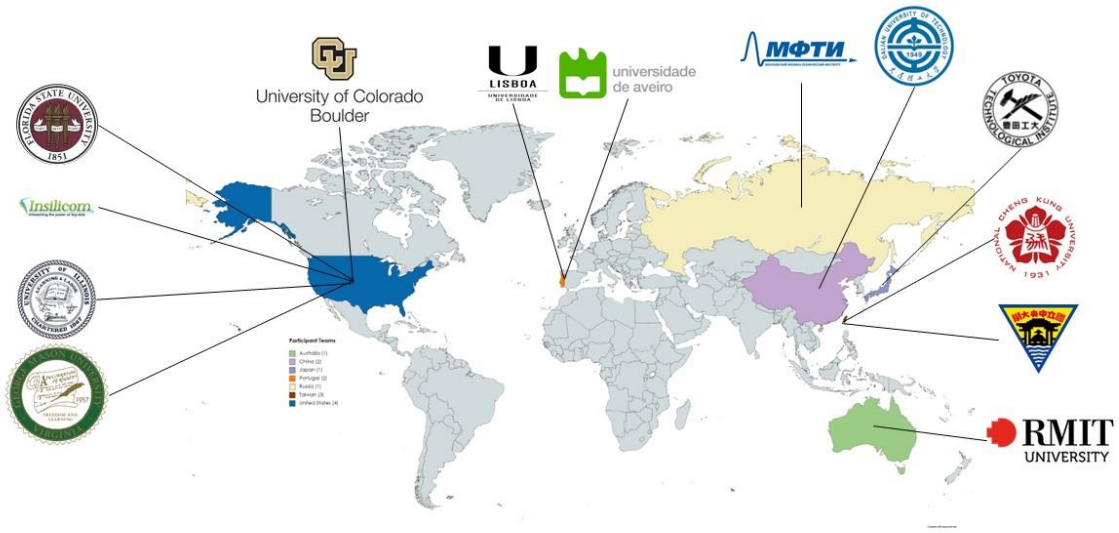

Figure 1. Team participation for the BioRED track at BioCreative VIII.

Results

We received 56 submissions from a total of 14 teams. The participating teams represent seven nations from Europe, Asia, Australia and North America. Two teams were from industry, with the remainder from universities. The teams reported sizes of 2 to 7 (average 4), typically with backgrounds in natural language processing, machine learning, information retrieval, and/or computer science.

Team Submissions

We report the performance for all valid submissions in Tables 4 through N. Tables 4-7 show the results for Sub-task 1, and Tables 8-11 show the results for Sub-task 2. Tables 4 and 8 respectively also list the median and mean results for all the submitted runs. In each table we select the best submitted run for each team (notice that the run numbers are different) for the corresponding evaluation measure, and we highlight all the results scoring higher than the mean. It is important to note that all teams reported increased F-scores in their unofficial submitted runs, which indicates that there is a lot of room for improvement for this task.

For Sub-task 2, the evaluation of the end-to-end systems depended on highly accurate predictions of the entities and their corresponding database identifiers. We noticed that teams

Table 4: Sub-task 1 evaluation results ranked by the F-score on recognizing the correct entity pairs in a relationship. For each team we selected their best run. The table lists the Median and Mean values of all runs (in %).

Team #	Best Run #	Entity Pair (F)	+Relation Type (F)	Entity+Novelty(F)	All (F)
156	5	**77.17**	58.86	58.24	44.39
129	4	**75.59**	56.67	59.20	44.41
127	1	**75.38**	55.93	57.69	43.04
142	1	**74.28**	51.90	56.26	38.91
114	2	**74.27**	54.76	58.54	43.16
138	5	**74.08**	52.82	56.85	40.55
157	1	**73.56**	52.76	56.05	39.71
118	4	**73.46**	55.31	56.45	42.53
116	2	56.41	26.54	40.94	18.26
148	1	52.67	37.26	39.46	28.29
155	1	48.96	28.78	30.37	17.61
154	1	32.48	7.27	12.90	2.96
111	1	24.20	8.66	15.48	5.33
Median		73.51	53.12	56.32	40.55
Mean		62.34	43.82	45.50	32.32

Table 5: Sub-task 1 evaluation results ranked by the F-score on recognizing the correct relationship type (in %).

Team #	Best Run #	Entity Pair (F)	+Relation Type (F)	Entity+Novelty(F)	All (F)
156	2	77.00	**58.95**	58.25	44.19
129	4	75.59	**56.67**	59.20	44.41
127	1	75.38	**55.93**	57.69	43.04
118	4	73.46	**55.31**	56.45	42.53
114	2	73.22	**55.05**	57.97	43.50
138	2	74.03	**53.31**	56.43	40.73
157	1	73.56	**52.76**	56.05	39.71
142	2	72.49	**52.58**	-	-
148	2	51.17	38.02	37.77	28.35
155	1	48.96	28.78	30.37	17.61
116	2	56.41	26.54	40.94	18.26
111	1	24.20	8.66	15.48	5.33
154	1	32.48	7.27	12.90	2.96

Table 6: Sub-task 1 evaluation results ranked by the F-score on recognizing the novelty factor for each pair in a relationship (in %).

Team #	Best Run #	Entity Pair (F)	+Relation Type (F)	Entity+Novelty(F)	All (F)
129	5	75.59	56.67	**59.22**	44.39
114	4	73.22	55.05	**58.54**	43.50
156	3	77.07	58.88	**58.27**	44.55
127	1	75.38	55.93	**57.69**	43.04
118	2	73.24	55.16	**57.32**	43.49
138	5	74.08	52.82	**56.85**	40.55
142	1	74.28	51.90	**56.26**	38.91
157	1	73.56	52.76	**56.05**	39.71
116	2	56.41	26.54	40.94	18.26
148	1	52.67	37.26	39.46	28.29
155	1	48.96	28.78	30.37	17.61
111	1	24.20	8.66	15.48	5.33

with higher resources and those that were able to achieve higher NER predictions, were able to report better relationship identification scores.

When we look at the methods and resources utilized by the participating teams on this task we notice that most teams relied on BERT-based models such as PubMed BERT, BioBERT, etc. For Sub-task 2, since most teams did not have enough time to develop their in-house NER systems, we noticed that the majority relied on PubTator API to retrieve the predicted entities.

Table 7: Sub-task 1 evaluation results ranked by the F-score on recognizing correctly the entity pairs in a relationship, their semantic type, and their novelty factor (in %).

Team #	Best Run #	Entity Pair (F)	+Relation Type (F)	Entity+Novelty(F)	All (F)
156	3	77.07	58.88	58.27	**44.55**
129	4	75.59	56.67	59.20	**44.41**
114	2	73.22	55.05	57.97	**43.50**
118	2	73.24	55.16	57.32	**43.49**
127	1	75.38	55.93	57.69	**43.04**
138	2	74.03	53.31	56.43	**40.73**
157	1	73.56	52.76	56.05	**39.71**
142	1	74.28	51.90	56.26	**38.91**
148	2	51.17	38.02	37.77	28.35
116	2	56.41	26.54	40.94	18.26
155	1	48.96	28.78	30.37	17.61
111	1	24.20	8.66	15.48	5.33
154	1	32.48	7.27	12.90	2.96

Table 8: Sub-task 2 evaluation results ranked by the F-score on recognizing the correct pair of entities in a relationship (in %).

Team #	Run #	NER (F)	ID (F)	Entity Pair (F)	+Relation Type (F)	Entity+Novelty(F)	All (F)
156	5	89.26	84.07	**55.84**	43.03	42.74	32.75
129	5	78.58	76.35	**41.27**	31.03	31.38	23.34
127	1	78.30	75.98	**39.45**	29.76	30.29	22.80
118	1	78.31	75.61	**39.03**	28.59	30.63	22.48
157	2	79.39	75.35	**38.08**	27.90	28.99	20.89
138	1	65.12	52.48	16.55	12.67	12.47	9.46
111	1	72.90	65.75	6.03	3.21	3.54	1.86
148	3	87.28	42.67	5.54	3.76	4.29	2.93
154	1*	69.98	29.58	5.10	1.31	2.47	0.69
Median		78.58	62.53	24.15	17.76	18.15	13.31
Mean		77.57	60.30	26.46	19.66	20.05	14.91

Table 9: Sub-task 2 evaluation results ranked by the F-score on recognizing the relationship type for each entity pair (in %).

Team #	Run #	NER (F)	ID (F)	Entity Pair (F)	+Relation Type (F)	Entity+Novelty(F)	All (F)
156	5	89.26	84.07	55.84	**43.03**	42.74	32.75
129	5	78.58	76.35	41.27	**31.03**	31.38	23.34
127	1	78.30	75.98	39.45	**29.76**	30.29	22.80
118	1	78.31	75.61	39.03	**28.59**	30.63	22.48
157	2	79.39	75.35	38.08	**27.90**	28.99	20.89
138	1	65.12	52.48	16.55	12.67	12.47	9.46
111	1	72.90	65.75	6.03	3.21	3.54	1.86
148	3	87.28	42.67	5.54	3.76	4.29	2.93

Table 10: Sub-task 2 evaluation results ranked by the F-score on recognizing correctly the novelty factor for each entity pairs in a relationship (in %).

Team #	Run #	NER (F)	ID (F)	Entity Pair (F)	+Relation Type (F)	Entity+Novelty(F)	All (F)
156	5	89.26	84.07	55.84	43.03	**42.74**	32.75
129	5	78.58	76.35	41.27	31.03	**31.38**	23.34
118	1	78.31	75.61	39.03	28.59	**30.63**	22.48
127	1	78.30	75.98	39.45	29.76	**30.29**	22.80
157	2	79.39	75.35	38.08	27.90	**28.99**	20.89
138	1	65.12	52.48	16.55	12.67	12.47	9.46
111	1	72.90	65.75	6.03	3.21	3.54	1.86
148	3	87.28	42.67	5.54	3.76	4.29	2.93

Table 11: Sub-task 2 evaluation results ranked by the F-score on recognizing the correct pair of entities in a relationship, their relationship type, and the novelty factor of that relationship (in %).

Team #	Run #	NER (F)	ID (F)	Entity Pair (F)	+Relation Type (F)	Entity+Novelty(F)	All (F)
156	5	89.26	84.07	55.84	43.03	**42.74**	**32.75**
129	5	78.58	76.35	41.27	31.03	**31.38**	**23.34**
127	1	78.30	75.98	39.45	29.76	**30.29**	**22.80**
118	1	78.31	75.61	39.03	28.59	**30.63**	**22.48**
157	2	79.39	75.35	38.08	27.90	**28.99**	**20.89**
138	1	65.12	52.48	16.55	12.67	12.47	9.46
111	1	72.90	65.75	6.03	3.21	3.54	1.86
148	3	87.28	42.67	5.54	3.76	4.29	2.93

NLTK and Spacy were popular tools for data pre-processing. We also noticed the use of additional resources such as CRAFT, AIONER, NCBI Entrez and OMIM, as well as the incorporation of GPT models in different capacities. We saw the use of large learning models for fine-tuning results and for data augmentation. Several teams used tools such as rhw, or nlpaug, to rewrite the input abstracts for data augmentation to their input models. Notably, all top performing teams relied on ensemble models.

Discussion and Conclusions

The BioRED track at BioCreative VIII had ambitious goals. We wanted to foster development of large-scale automated methods that could handle the task of relationship extraction 1) at the document level, 2) between multiple entity types, and 3) on a real-life setting, where we are faced with new, previously unseen research topics that might contain previously unseen entities.

To this end, we reviewed some of the submitted results to see if there are particular challenges that we could direct future systems to focus on. We selected all articles in the testing dataset on the topic of COVID-19, which was a topic unseen in the training data, and reviewed all submitted results for these articles. While we were very encouraged to see that for more than 80% of these articles there were systems reporting an F-score of 70% or higher, we noticed that systems still have difficulties when the entities are not expressed in the same sentence, and especially when they are positioned far apart in the document. The more co-references, and the more difficult the language in the article, the harder it is for automated system to find the right relationship. In conclusion, we notice the wide participation and even wider interest that this initiative received, and we believe that the BioRED dataset, and the BioRED evaluation

leaderboard at the codalab location will be widely used as a benchmark to evaluate better systems that are better able at handling these challenges.

Funding

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.].

References

1. Krallinger, M., Vazquez, M., Leitner, F., Salgado, D., Chatr-Aryamontri, A., Winter, A., Perfetto, L., Briganti, L., Licata, L. and Iannuccelli, M. (2011) The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC Bioinformatics*, **12**, 1-31.
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. 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: The Journal of Biological Databases and Curation*, **2016**.
4. 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*, **2014**.
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: The Journal of Biological Databases and Curation*, **2019**.
6. Krallinger, M., Rabal, O., Akhondi, S.A., Pérez, M.P., Santamaría, J., Rodríguez, G.P., Tsatsaronis, G., Intxaurrondo, A., López, J.A. and Nandal, U. (2017), *Proceedings of the sixth BioCreative challenge evaluation workshop*, Vol. 1, pp. 141-146.
7. Miranda, A., Mehryary, F., Luoma, J., Pyysalo, S., Valencia, A. and Krallinger, M. (2021), *Proceedings of the seventh BioCreative challenge evaluation workshop*.
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*.
9. Islamaj, R., Wei, C.-H., Lai, P.-T., Luo, L., Coss, C., Kochar, P.G., Miliaras, N., Printseva, O., Rodionov, O., Sekiya, K., Trinh, D., Whitman, D., and Lu, Z. (2023), *Proceedings of the eighth BioCreative challenge evaluation workshop 2023*, New Orleans, LA.
10. Islamaj, R., Lai, P.-T., Wei, C.-H., Luo, L. and Lu, Z. (2023) BioCreative VIII Track 1: BioRED (Biomedical Relation Extraction Dataset) Track Subtask 1. https://codalab.lisn.upsaclay.fr/competitions/13377.
11. Islamaj, R., Lai, P.-T., Wei, C.-H., Luo, L. and Lu, Z. (2023) BioCreative VIII Track 1: BioRED (Biomedical Relation Extraction Dataset) Track Subtask 2. https://codalab.lisn.upsaclay.fr/competitions/13378.
12. Comeau, D.C., Islamaj Doğan, R., Ciccarese, P., Cohen, K.B., Krallinger, M., Leitner, F., Lu, Z., Peng, Y., Rinaldi, F. and Torii, M.J.D. (2013) BioC: a minimalist approach to interoperability for biomedical text processing. **2013**.
13. Brown, G.R., Hem, V., Katz, K.S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K.D. and Maglott, D.R. (2015) Gene: a gene-centered information resource at NCBI. *Nucleic acids research*, **43**, D36-D42.

14. Lipscomb, C.E. (2000) Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, **88**, 265.
15. Smigielski, E.M., Sirotkin, K., Ward, M. and Sherry, S.T. (2000) dbSNP: a database of single nucleotide polymorphisms. *Nucleic acids research*, **28**, 352-355.
16. Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K. and Robbertse, B. (2020) NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*, **2020**, baaa062.
17. Bairoch, A. (2018) The cellosaurus, a cell-line knowledge resource. *Journal of biomolecular techniques: JBT*, **29**, 25.
18. 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**.
19. Lai, P.-T., Islamaj, R., Wei, C.-H., Luo, L., and Lu, Z. (2023), *Proceedings of the eighth BioCreative challenge evaluation workshop 2023*, New Orleans, LA.