Assessing the state of the art in biomedical relation extraction: evaluating ChatGPT, PubMedBERT and BioREx for the BioRED track at BioCreative VIII

Po-Ting Lai[1], Rezarta Islamaj[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
Biomedical relation extraction aims to identify and categorize relationships between biomedical entities in unstructured text. This is crucial for various biomedical NLP applications, from drug discovery to custom medical solutions. BioRED track at the BioCreative VIII challenge and workshop aimed to foster the development of novel algorithms for biomedical relation extraction. This challenge differed from previous relation extraction challenges because it addresses relation extraction at the document level, addresses relation extraction between five entity types, in eight semantic categories and asks for the classification of the relation on whether it is a novel finding according to the document or background knowledge. In this paper, we use the BioRED track training dataset, and build three different benchmarking systems using: BERT, GPT and BioREx methods. The BioRED track consisted of two sub-tasks: the first subtask provided the participants with article titles, abstracts and human annotated genes, diseases, chemicals, gene variants, cell lines, and species mentions in the text. All annotated entities were linked to database identifiers, the second sub-task did not provide entity annotations, and asked for end-to-end relation extraction systems. Although we discuss three different systems, we followed a similar overall strategy for both sub-tasks, and we considered them as multi-label classification problems. For sub-task 1 we used the provided human annotations as entity inputs, while for sub-task 2 we retrieved the PubTator output for all articles. Here we discuss our three different approaches, and offer our perspective on the advantages and limitations of these approaches. Our best performing system was the BioREx model with an F1-score of 75.68%, and 56.89% on recognizing the entity pairs, and relation types respectively, surpassing the median (73.56%, and 53.17%) and average scores (67.03% and 47.74%) of all participating teams.

Introduction
Biomedical relation extraction aims to automatically discern and categorize relationships between biomedical concepts from natural text data. This task stands central to biomedical natural language processing (NLP), fostering advancements in areas like drug discovery and personalized medicine.

The BioRED track at BioCreative VIII (1,2) challenged participants to engage with 600 articles of the BioRED dataset (3) used in the LitCoin challenge[1], with the goal to train systems capable to discern and categorize the same defined biomedical relationships in new published articles. Participants could optionally integrate any other publicly available biomedical resources, datasets, tools and APIs to supplement their efforts. Additionally, the BioRED track also asked for the classification of the identified relations into novel findings, distinguishing from relations describing pre-existing knowledge or background information. There were two sub-tasks differing on the type of the input data provided. Sub-task 1 provided the article text (title and abstract) and the human expert annotated entities, while sub-task 2 only provided the article text and therefore required the development of an end-to-end system capable of both detecting the entities and identifying the asserted relationships, their semantic categories and their novelty factor.

Material and Methods

Data and Tools Utilized to Develop Our Systems

For the development of our benchmarking systems, we leveraged a variety of datasets and tools:

- **Dataset:** BioRED training set was used for training and assessing our models. However, for the BioREx model (4), we also used these additional datasets, because they are used in the BioREx training: BC5CDR (5), DDI (6), DrugProt (7), AIMed (8).
- **PubTator Central API (9):** A web-based service for automated annotation of biomedical concepts in the literature that helps researchers identify and classify biomedical entities. It focuses on automatic concept identification and normalization. It has annotations for genes/proteins, genetic variants, diseases, chemicals, species, and cell lines.
- **PubTator 3 API[2]:** An enhanced version of PubTator Central with advanced NER and normalization features. It's important to note that PubTator3 enhances the NER/Normalization with AIONER (10) and GNorm2 (11). It also integrates BioREx for relation extraction.
- **GPT Models (GPT 3.5 (12) and GPT 4 (13)):** These advanced language models are from OpenAI. We deployed Azure's OpenAI API (version “2023-03-15-preview”), and we used it for both relation pair extraction and the novelty tasks. Specifically, we employed the “gpt-3.5-turbo” and “gpt-4” engines, setting temperature values at 0.1 to ensure responses were fine tuned to our input prompts.
- **BioREx:** BioREx is a biomedical relationship extraction tool that, in addition to the BioRED corpus, uses other publicly available datasets similar to the BioRED as additional training data. Compared to PubMedBERT (14), it performed significantly better. We utilized its pre-trained BioLinkBERT (15) model for PubTator3. However, BioREx cannot predict the novelty of a relation, so when using this model, we were unable to give predictions for the novelty factors.
- **PubMedBERT:** PubMedBERT is utilized to identify and categorize relationships between biomedical entities in the text. We employed its trained models, specialized in biomedical literature, to classify relationships into specific types based on context and

[1] LitCoin NLP Challenges: https://bitgrit.net/competition/13; https://bitgrit.net/competition/14
[2] https://www.ncbi.nlm.nih.gov/research/pubtator3/

entity involvement. PubMedBERT was also used for determining the novelty of a relationship. PubMedBERT is able to analyze the context to assess whether a described relationship is new finding or previously known information.

Benchmarking systems

The benchmarking systems developed for evaluating biomedical relation extraction include:

- PubMedBERT-based approach: We utilized the open-source implementation of PubMedBERT and its pre-trained models (4) for BioRED relation type and novelty label classification tasks. Figure 1 illustrates how this model operates. PubMedBERT treats both sub-tasks as multi-class classification tasks, where, given a pair of entities represented by their normalized IDs and the text (title and abstract), it classifies them into the corresponding class using the [CLS] tag. BioRED model produces two outputs, which are independent models: the relation type model and the novelty prediction model. In sub-task 2, we use the same setup with one difference: the entity predictions are extracted from the PubTator Central and PubTator 3 APIs.

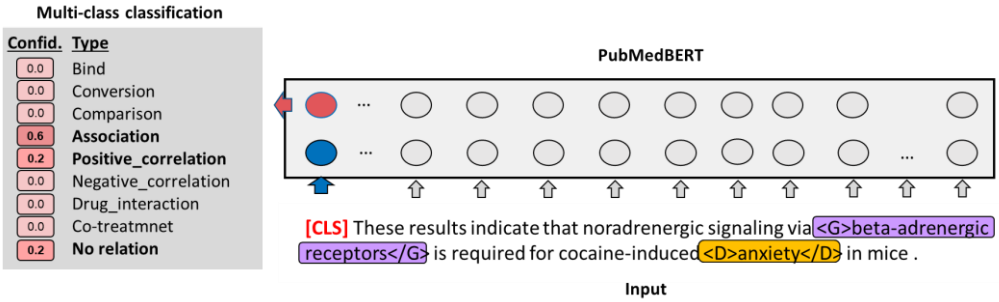

Figure 1. An illustration of relation task formulation of PubMedBERT.

- **BioREx-based approach**: We explain the BioREx based system in Figure 2. As shown in the Figure, BioREx is trained on nine combined datasets, including BioRED and eight relation extraction datasets. BioREx's rules do not deal with multi-label problem, where an entity pair can have a relation type and a novelty label at the same time. Therefore, it does not address the novelty label classification problem, and only focuses on relation type classification.
- **GPT-based approach:** As illustrated in Figure 3, The GPT prompts involved a series of explanations to setup the LLM for relation prediction. Initially we listed the entities, by first listing their type, then their corresponding normalized database identifier, and then following with the exact text mention as it appears in the given text (title and abstract). This information is labelled as "Passage 1 (Synonyms)". "Passage 2 (Main Article)" provides the given article. Following that, Questions 1 and 2 were designed to define the relation types and the novelty labels.

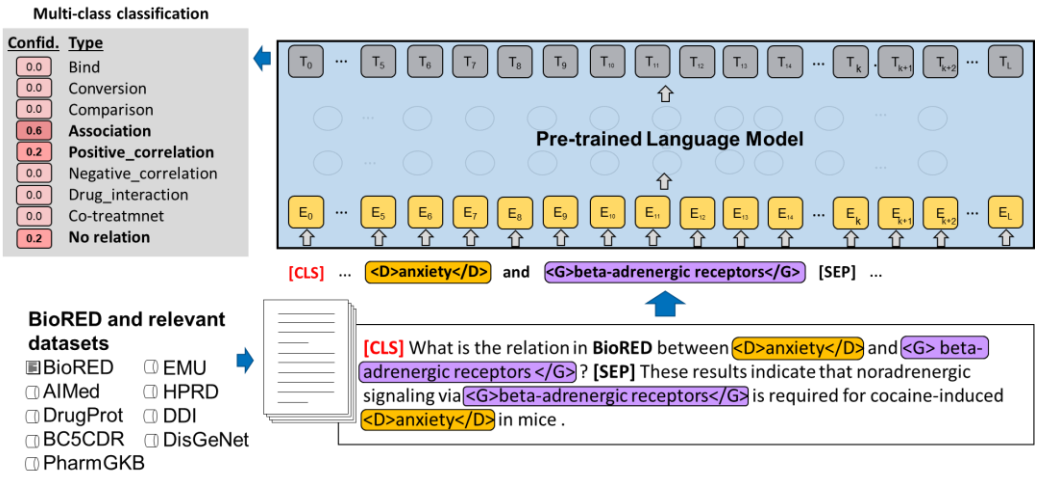

Figure 2. An illustration of relation task formulation of BioREx.

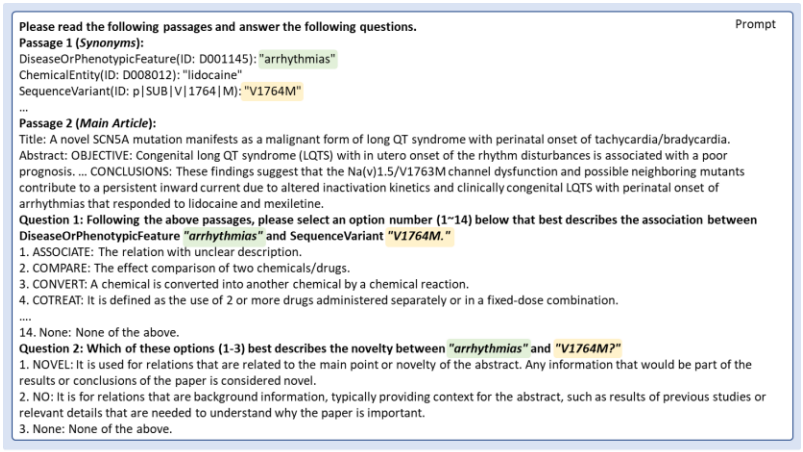

Figure 3. An illustration of prompt input for OpenAI GPT-3.5 and GPT-4.

Results

We list the F1-score for each of our evaluations. For both Sub-tasks, we list the evaluation by identifying the pair of entities (E) in a relationship, their relation type (R), and their novelty label (N). This evaluation used the official BioRED track evaluation script. In sub-task2, we also evaluate NER and ID performances, where we evaluate ID at the document level.

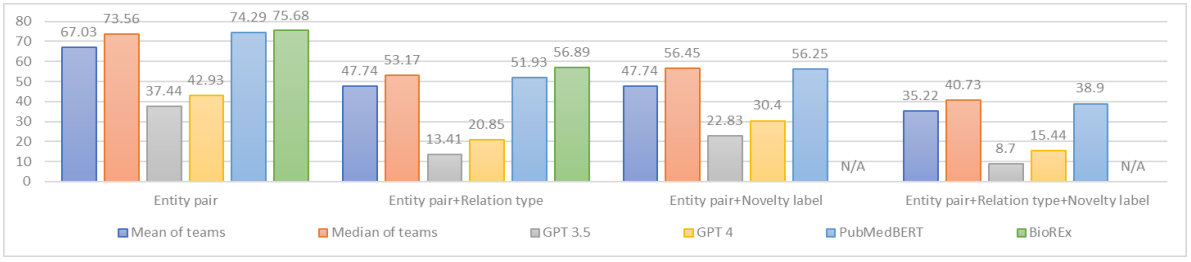

Figure 4. Sub-Task 1 performance evaluation (F-measure scores in %) for the GPT3.5, GPT4, PubMedBERT and BioREx models, as compared to the mean and median of all submissions to BioRED track at BioCreative VIII workshop.

Figure 4 shows the performances of our systems on Sub-task 1 along with the median and average scores from all teams' runs. We note that the GPT models achieved low F-scores. We plan to experiment with some additional prompt variations to see if it is possible to achieve higher results. However, we believe that a better approach would be to integrate the GPT output into the other models. For the identification of correct entity pairs in a relation, BioREx achieved the best F-score of 75.68, and both PubMedBERT and BioREx models performed higher than the mean and median of all participating teams. For the identification of the correct relation type, BioREx model again achieved the best F-score of 56.89, but the PubMedBERT model performance, while higher than the mean, scored lower than the median of all participating teams. For the novelty identification and relation extraction (all), again, our PubMedBERT model performed higher than the mean, but lower than the median of all participating teams.

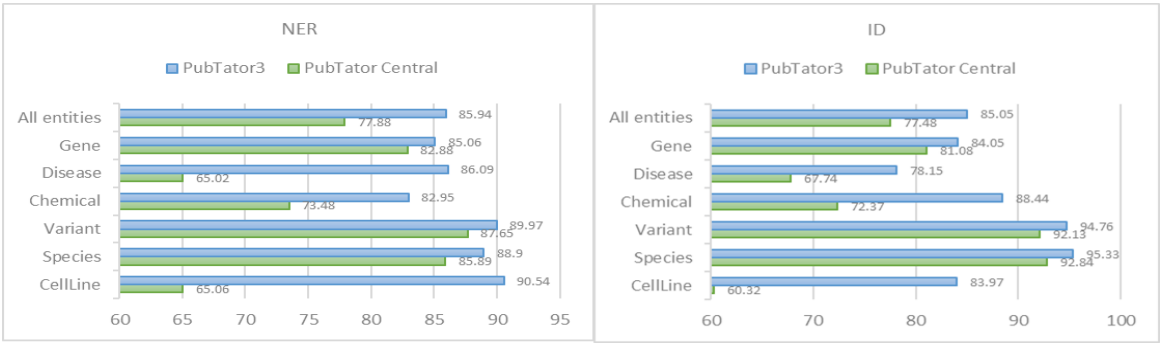

Figure 5. Performance comparison of the two API tools PubTator Central and PubTator3 (in F1%) which were used to extract the entity predictions for Sub-task 2.

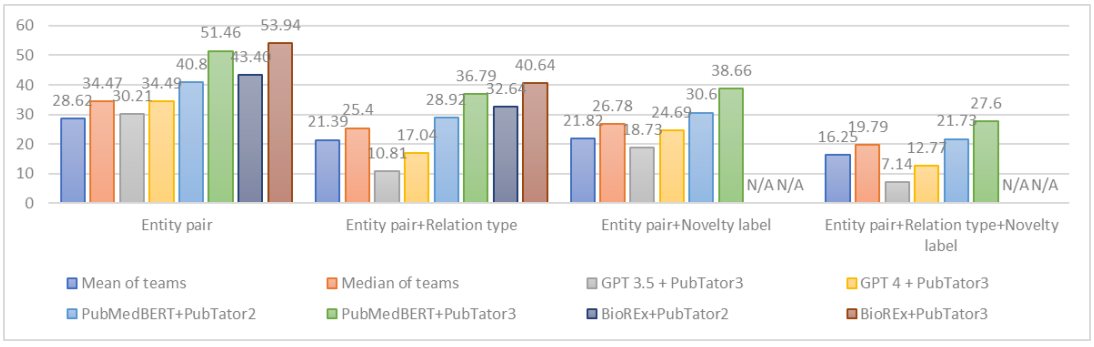

Figure 6. Sub-Task 2 performance evaluation (F1-scores in %) for the GPT3.5, GPT4, PubMedBERT and BioREx models when paired with PubTator Central, and PubTator 3 APIs in an end-to-end system, as compared to the mean and median of all submissions to BioRED track at BioCreative VIII workshop.

Figure 5 provides a detailed comparison of NER and ID performances for Sub-task 2 when we use the PubTator Central API versus when we use the PubTator 3 API. Consequently, Figure 6 illustrates the results for the end-to-end systems on Sub-Task 2 when we combine our models with either the PubTator Central API or the PubTator 3 API. Here, we note that the end-to-end

model depends on the high accuracy of the entity recognition system. Because the PubTator 3 system has a higher accuracy, that translates into a significant improvement of all our models when they use that output as their starting point. In fact, for Sub-task 2, we are able to produce better results than the mean and median values of the BioRED track participants for all evaluation metrics.

As expected, the GPT-4 model outperformed GPT-3.5. However, with respect to relation extraction GPT-4 performance with our current setting is not comparable to those of the PubMedBERT or BioREx models. This results aligns with similar findings from other biomedical RE tasks (16).

Limitations and Future work

Our work in organizing this track and also experimenting with large language models both GPT3.5 and GPT 4, as well as PubMedBERT and BioREx has given us a lot of insights on how we can improve in the future. First, the BioREx model concentrates only on the entity pair task, as this is typically where curators reach higher consensus, potentially leading to more accurate practical applications. It's worthwhile to note that the BioRED corpus is unique in that differently from the other annotated datasets, it includes annotations for *Novelty*. Given the BioREx's data-centric approach, one way to augment Novelty in BioREx could be to consider the novelty annotations as a new dataset. This would enable the application of data-centric methodology to predict novelty, thus providing valuable results and improvements. Additionally, on our application of the biomedical relation extraction in the GPT models, we did not leverage the GPT output to augment the PubMedBERT or BioREx outputs. We believe that combining these models to leverage each-other strengths could improve the performance. Additionally, we could further experiment with the prompt to the GPT models. We also observed that track participants teams reported multi-task models which showed better performance. We believe that this kind of approach could lead to a better solution for the relation extraction task.

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. 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.
2. Islamaj, R., Wei, C.-H., Lai, P.-T., Luo, L., Coss, C., Kochar, P.G., Miliaras, N., Printseva, O., Rodionov, O., Sekiya, K. *et al.* (2023), *Proceedings of the eighth BioCreative challenge evaluation workshop 2023*, New Orleans, LA.
3. 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*.

4. 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**.
5. 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**.
6. Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P. and Declerck, T. (2013) The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of Biomedical Informatics*, **46**, 914-920.
7. Miranda, A., Mehryary, F., Luoma, J., Pyysalo, S., Valencia, A. and Krallinger, M. (2021), *Proceedings of the seventh BioCreative challenge evaluation workshop*.
8. Bunescu, R., Ge, R., Kate, R.J., Marcotte, E.M., Mooney, R.J., Ramani, A.K. and Wong, Y.W. (2005) Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, **33**, 139-155.
9. 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.
10. Luo, L., Wei, C.-H., Lai, P.-T., Leaman, R., Chen, Q. and Lu, Z. (2023) AIONER: all-in-one scheme-based biomedical named entity recognition using deep learning. *Bioinformatics*, **39**.
11. Wei, C.-H., Luo, L., Islamaj, R., Lai, P.-T. and Lu, Z. (2023) GNorm2: an improved gene name recognition and normalization system. *Bioinformatics*.
12. OpenAI. (2023) Azure OpenAI Service API [gpt-35-turbo].
13. OpenAI. (2023) Azure OpenAI Service API [gpt-4].
14. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J. and Poon, H. (2021) Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, **3**, 1-23.
15. Yasunaga, M., Leskovec, J. and Liang, P. (2022). Association for Computational Linguistics, Dublin, Ireland, pp. 8003-8016.
16. Tian, S., Jin, Q., Yeganova, L., Lai, P.-T., Zhu, Q., Chen, X., Yang, Y., Chen, Q., Kim, W. and Comeau, D.C. (2023) Opportunities and Challenges for ChatGPT and Large Language Models in Biomedicine and Health. *arXiv preprint arXiv:2306.10070*.