BioRED task NCU-IISR submission: Preprocessing-Robust Ensemble Learning Approach for Biomedical Relation Extraction

Wilailack Meesawad[1], Chun-Yu Hsueh[1], Yu Zhang[1], Jen-Chieh Han[1], and Richard Tzong-Han Tsai[1,2,*]

[1]Department of Computer Science and Information Engineering, National Central University, Taiwan
[2]Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan

[*]Corresponding author: E-mail: thtsai@csie.ncu.edu.tw

Abstract
We propose an end-to-end system for the BioCreative VIII Challenge Track 1: BioRED Track, focusing on biomedical relation extraction. In our study, we employ an ensemble learning approach, combining the PubTator API with multiple pre-trained Bidirectional Encoder Representations from Transformers (BERT) models. A variety of preprocessing inputs are utilized, including prompt questions, entity ID pairs, and co-occurrence contexts. Special tokens and boundary tags are added to enhance model understanding. In this study, PubMedBERT and the Max Rule ensemble learning mechanism are used to combine outputs from different classifiers. In subtask 1, the method achieves a F1 score of 43%, and in subtask 2, it achieves a score of 23%, demonstrating significant advancements in biomedical relation extraction.

Keywords: Relation extraction, Biomedical natural language processing, Fine-Tuning, Ensemble learning.

Introduction
Relation Extraction is a crucial task in biomedical natural language processing, which aims to identify and classify relationships between biomedical entities, such as genes, diseases, and proteins, within biomedical texts. During the past few years, BioCreative (1) (Critical Assessment of Information Extraction Systems in Biology) has been assessing the current state-of-the-art for specific purposes in the field of biomedical text mining and information extraction. It is a tradition for this group to hold a Challenge each year.

During BioCreative VIII Challenge Task 1, we implemented our approach in the BioRED (2) Track. Two subtasks are involved. The purpose of subtask 1 is to develop methods for relation extraction. Participants are also required to categorize relationships that represent novel findings (the key concepts of this track), rather than previous background knowledge or other available information. The second subtask requires the teams to develop an end-to-end system to identify relationships in free text. We investigate different preprocessing inputs for bidirectional encoder representations based on transformers (BERT). Through the use of ensemble learning, the performance of the system is further enhanced.

Material and Methods

Dataset

This year, the BioRED corpus was provided, a collection of 600 PubMed articles that contains manual annotations of biomedical concepts and binary relationships by domain experts. For training, 500 articles are used, while for validation, 1,000 articles are used. There are only 100 abstracts used for leaderboard evaluation.

In subtask 1, they provided both the abstract and human-annotated entities, whereas in subtask 2, they provided only the abstract. Approximately 10,000 documents are used as the test set for the final evaluation. A total of 400 test set documents are concealed in this collection. Performance will be evaluated only using these test set articles.

In BioRED datasets, the experts annotate relations at the document level. Entity spans have an entity ID. Additionally, some entity spans can have two or more entity IDs.

Problem formulation

The system we use in subtask 1 is based on the open-source RE system implementation of the BioRED paper (3). A relation candidate instance consists of two biomedical entities and the context of their co-occurrence. As some entity spans contain two or more entity IDs, the relation with those spans must be expanded to include two or more instances. The purpose of this work is to classify instances into a predefined relation extraction type or to classify them as being unrelated (i.e. "None" as a negative example). The aim of this task is also to identify whether these relations are novel findings or existing information, by categorizing the instances as "Novel", "No" (not novel findings), or "None" (negative examples). Figure 1 illustrates the model architecture.

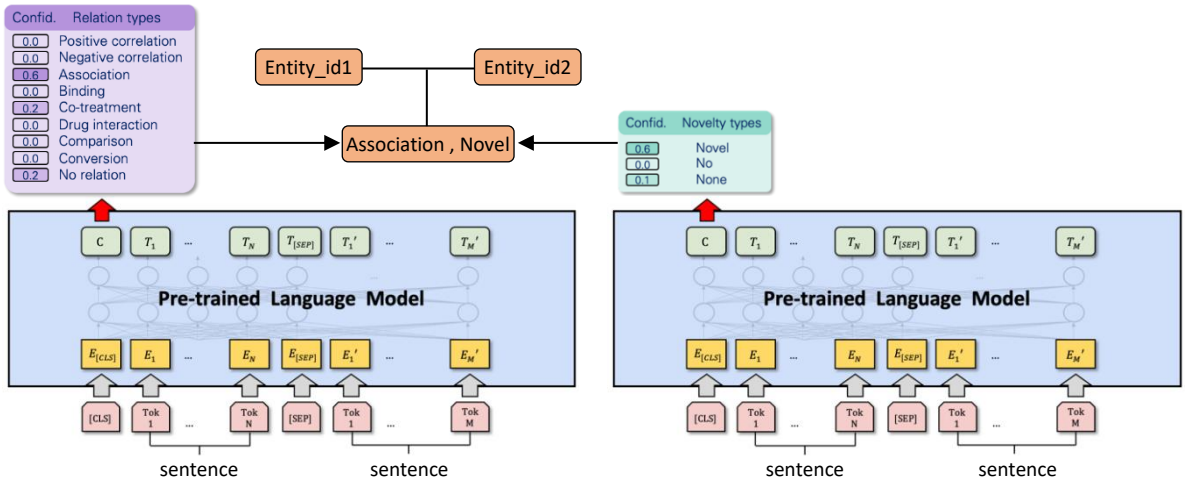

Figure 1: Model architecture.

The development of an end-to-end system (subtask 2) was initiated by leveraging PubTator API (4) to access essential biomedical concepts and entity IDs. In addition, these biomedical concepts and entity IDs were standardized to match the format of the datasets.

In order to normalize the dataset retrieved from PubTator API, specific terms were mapped to more generalized categories. We consolidated the expressions "ProteinMutation," "DNAMutation," and "SNP" into "SequenceVariant." Furthermore, "Chemical" was unified as "ChemicalEntity," "Disease" as "DiseaseOrPhenotypicFeature," "Gene" as "GeneOrGeneProduct," and "Species" as "OrganisationTaxon." With regard to entity IDs, we performed a series of transformations, including the removal of prefixes such as "MESH:," "tmVar," and hyphens ("-"). We also standardized notations, such as changing "CVCL:" to "CVCL_" and "RS#:" to "RS". Using the pre-trained models from subtask 1, we made predictions on the processed output. The performance of the models was further enhanced by applying an ensemble learning approach. Figure 2 illustrates the workflow.

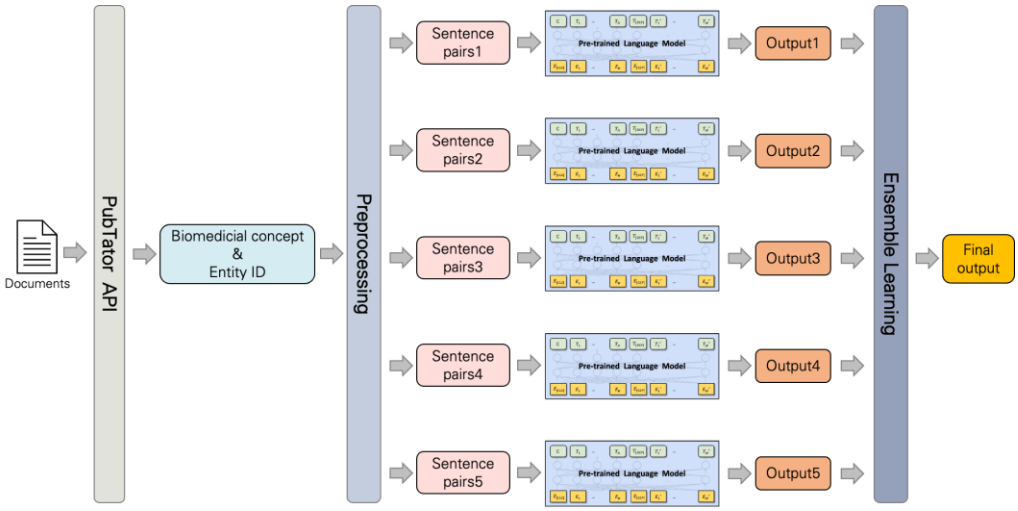

Figure 2: Workflow of our approach.

Method

Table 1 shows several preprocessing sentence pairs we did to improve model performance. In each instance, we entered two sentences. The first sentence (sentence1) is the generated prompt question or entity ID pair to provide semantic similarity to the model for its corresponding pair. A second sentence (sentence2) describes a co-occurrence context in which entity span is replaced by entity ID, and two boundary tags are inserted at the beginning and the end of the entities (e.g., <GeneOrGeneProduct> and </GeneOrGeneProduct> for genes). To ensure that these tags are not separated into multiple tokens, we also added those tags to the Pre-train language model's vocabulary. Additionally, in order to adhere to the standard practice of using pre-trained BERT models for classification tasks, we also include the special tokens "[CLS]" at the beginning of the instance, and "[SEP]" in the middle between sentence1 and sentence2.

This study uses PubMedBERT (5), which is a state-of-the-art pre-trained language model with specialization in the medical domain, to fine-tune a text classification model on the BioRED dataset. In our fine-tuning process, two distinct aspects are classified: relation type and novelty, each of which is treated as a separate classification task. The relation type classifier and the novelty classifier are both trained to provide confidence scores for each class within their respective tasks. As a result of these confidence scores, we can measure the model's confidence in the predicted class for each instance.

The Max Rule ensemble learning mechanism is used to enhance the quality and reliability of our predictions. By using this mechanism, confidence scores are aggregated from multiple models, each trained using a different preprocessing input. Max Rule ensures that the final prediction is based on the class with the highest probability score, thereby improving the robustness and accuracy of the classification results. Using the predictions made by the relation type classifier, we generate the submission file. In situations where an instance indicates a relationship but is categorized as 'None' for novelty, we classify the novelty as 'Novel,' which indicates the presence of new findings in the medical text.

Table 1: The different preprocessing sentence pair inputs.

Sentence Pairs	Sentence1		Sentence2
	Relation Type Task	Novelty Task	Relation Type & Novelty Tasks
pairs1	<entity_type1> entity_id1 </entity_type1> <entity_type2> entity_id2 </entity_type2>	<entity_type1> entity_id1 </entity_type1> <entity_type2> entity_id2 </entity_type2>	Association between promoter -1607 polymorphism of <entity_type1> entity_id1 </entity_type1> and <entity_type2> entity_id2 </entity_type2> in Southern Chinese …
pairs2	What is the relation type between <entity_type1> entity_id1 </entity_type1> and <entity_type2> entity_id2 </entity_type2>?	The relation between <entity_type1> entity_id1 </entity_type1> and <entity_type2> entity_id2 </entity_type2> is novel findings?	Association between promoter -1607 polymorphism of <entity_type1> entity_id1 </entity_type1> and <entity_type2> entity_id2 </entity_type2> in Southern Chinese …
pairs3	What is the relation type between entity_id1 and entity_id2?	The relation between entity_id1 and entity_id2 is novel findings?	Association between promoter -1607 polymorphism of <entity_type> entity_id1 </entity_type1> and <entity_type2> entity_id2 </entity_type2> in Southern Chinese …
pairs4	What is the relation type between entity_id1 and entity_id2?	What is the novelty type between entity_id1 and entity_id2?	Association between promoter -1607 polymorphism of <entity_type1> entity_id1 </entity_type1> and <entity_type2> entity_id2 </entity_type2> in Southern Chinese …
pairs5	<entity_type1> entity_id1 </entity_tpye1> and <entity_type2> entity_id2 </entity_type2>	<entity_type1> entity_id1 </entity_tpye1> and <entity_type2> entity_id2 </entity_type2>	Association between promoter -1607 polymorphism of <entity_type> entity_id1 </entity_type1> and <entity_type2> entity_id2 </entity_type2> in Southern Chinese …

Results and Discussion

Using the BioCretive VIII BioRED Track CodaLab, the evaluation score is calculated. Subtask 1 has three benchmark schemas: (i) entity pair: extract the concept identifiers within the relation, (ii) entity pair + relation type: recognize the specific relation type for the extracted pairs, and (iii) entity pair + relation type + novelty: label the novelty for the extracted pairs. Subtask 2 has two

more benchmark schemas: (i) biological concepts (NER): to recognize the biomedical named entity, and (ii) ID: to extract the entity ID.

As shown in Table 2, various model input formats can capture a variety of aspects and excel at certain evaluation metrics. To improve the model performance, we selected sentence pairs that demonstrated remarkable F1 scores for both relation classification and novelty classification. These top-performing pairs were subsequently subjected to an ensemble learning technique. Based on the results in Table 3, ensemble learning significantly improved the performance of our model. We have found that this approach has proved to be a valuable tool for achiveing better results, emphasizing the importance of utilizing ensemble techniques for text classification.

Table 2: Performance on the validation dataset of subtask 1. P is precession, R is recall, and F is F1 score.

Sentence Pairs	Entity Pair			+ Relation Type			+ Novelty		
	P	R	F	P	R	F	P	R	F
pairs1	0.8058	0.7352	0.7689	0.6569	0.5993	0.6268	0.5580	0.5090	**0.5324**
pairs2	0.8209	0.7567	0.7875	0.6744	0.6217	**0.6470**	0.5429	0.5004	0.5208
pairs3	0.8158	0.7730	0.7938	0.6661	0.6311	**0.6481**	0.5581	0.5288	**0.5430**
pairs4	0.8256	0.7369	0.7787	0.6763	0.6036	**0.6379**	0.5568	0.4970	0.5252
pairs5	0.8411	0.7326	0.7831	0.6851	0.5967	**0.6379**	0.5735	0.4996	**0.5340**

Table 3: Performance of ensemble learning mechanism on the validation dataset of subtask 1. All numbers are F1 scores.

ensemble	Sentence Pairs		Entity Pair	+ Relation Type	+ Novelty
	Relation Type Task	Novelty Task			
ensemble	pairs2 + pairs4 + pairs5	pairs1 + pairs3 + pairs5	0.7991	0.6731	0.5793

Key point for each run

For subtask1;

- Run1: Ensemble
- Run2: Preprocessed input from sentence pairs3

For subtask2;

- Run1: PubTator + ensemble
- Run2: PubTator + preprocessed input from sentence pairs3

Table 4 presents the results of our test submission for subtask 1. Our best submission (Run1) achieved a significantly higher F1 score than the median and average F1 scores of other participants. Subtask 2 also showed promising results. Our F1 score was higher than the median and average F1 scores of the participants, with the exception of the NER benchmark. As shown in Table 5, our F1 score was slightly lower than the median F1 score of participants. As compared to the broader pool of submissions, these results show how effective and competitive our approach is in both subtask 1 and subtask 2.

Table 4: Performance on the test submission of subtask 1. All numbers are F1 scores.

Run	Entity Pair	+ Relation Type	+ Novelty
Run1	0.7538	0.5593	0.4304
Run2	0.7383	0.5319	0.4056
Median of participants runs	0.7356	0.5317	0.4073
Average of participants runs	0.6703	0.4774	0.3522

Table 5: Performance on the test submission of subtask 2. All numbers are F1 scores.

Run	NER	ID	Entity Pair	+ Relation Type	+ Novelty
Run1	0.7830	0.7598	0.3945	0.2976	0.2280
Run2	0.7830	0.7598	0.3931	0.2889	0.2194
Median of participants runs	0.7858	0.6681	0.3447	0.2540	0.1979
Average of participants runs	0.7687	0.6336	0.2862	0.2139	0.1625

Limitation and Future Work

In order to predict novelty, we rely on the prediction of relation type. If the prediction of the relation type is "None" (no relation as a negative example), the novelty prediction will be ignored, resulting in a false negative outcome. The novelty prediction that is predicted as "None" (negative example), while the prediction of relation type is not "None", we will classify that novelty as "Novel" (novel findings), which will affect the performance. As a part of our future research, we intend to apply some other techniques, such as the Hierarchical Bayesian approach, to help solve the false negative issue.

References

1. Arighi, C. (n.d.-a). BioCreative - BioCreative VIII challenge and workshop. https://biocreative.bioinformatics.udel.edu/events/biocreative-viii/biocreative-viii-challenge/

2. Arighi, C. (n.d.-c). BioCreative - Track 1: BioRED (Biomedical Relation Extraction Dataset) Track. https://biocreative.bioinformatics.udel.edu/tasks/biocreative-viii/track-1/

3. Luo, L., Lai, P., Wei, C., Arighi, C. N., & Lu, Z. (2022b). BioRED: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics, 23*(5). https://doi.org/10.1093/bib/bbac282

4. U.S. National Library of Medicine. (n.d.). Pubtator Central API - NCBI - NLM - NIH. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/research/pubtator/api.html

5. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-Specific Language Model Pretraining for biomedical natural language

processing. *ACM Transactions on Computing for Healthcare, 3*(1), 1–23. https://doi.org/10.1145/3458754