Identifying Drug/chemical-protein Interactions in Biomedical Literature using the BERT-based Ensemble Learning Approach for the BioCreative 2021 DrugProt Track

Ting-Wei Chang[1], Tzu-Yi Li[2], Yu-Wen Chiu[3], Sheng-Jie Lin[3], Panchanit Boonyarat[3], Wen-Chao Yeh[4], Neha Warikoo[5], Yung-Chun Chang[3,*]

[1]Taipei Medical University Hospital, Taipei, Taiwan

[2]School of Health Care Administration, Taipei Medical University, Taipei, Taiwan

[3]Graduate Institute of Data Science, Taipei Medical University, Taipei, Taiwan

[4]Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu, Taiwan

[5]Helmholtz Centre for Infection Research, Braunschweig, Germany

***Abstract*—Accurate biomedical entity-relation prediction is essential in numerous biomedical and physiology research fields. Machine Learning (ML) and Natural Language Processing (NLP) methods can provide efficient tools to detect crucial information regarding chemical entities and gene entities in a great deal of biomedical and physiology literature. Our goal is to examine the prediction result of the chemical gene relations from different models and evaluate their performances. In particular, experiments are conducted with the state-of-the-art language model, BERT, on the dataset provided by the BioCreative track 1: DrugProt datasets.**

Keywords—drug-protein interaction; chemical-protein interaction; BERT; ensemble learning; log-likelihood ratio

I. INTRODUCTION

The current global pandemic triggered abundant studies that provided new knowledge and methods, including biomedical research. As biomedical literature accumulates at a record-breaking rate, it has become increasingly challenging to organize and analyze drug-related information described in the scientific literature efficiently. A key aspect of drugs and chemical compounds is their relationship with certain biomedical entities, in particular, genes and proteins.

Chemical Natural Language Processing (ChemNLP) and text mining technologies, such as chemical text mining, is crucial to improve the access and integration of unstructured data sources, including patents and scientific literature. A great number of data scientists are dedicated to studying protein-protein interactions (PPIs) extraction. Protein relationship prediction provides a significant data research foundation for biomedical research with the development of natural language processing approaches (1–4). As a result, the number of biological entities worthy of our attention is increasing rapidly, which include not only proteins, but also genetic, chemical, and cellular entities. Besides, more annotated databases are available, which include information on genetic relationships between entities based on their occurrence in text and stored as structured data. They are of key relevance not only for biological but also for pharmacological and clinical research. Nevertheless, extracting meaningful relationships is a tedious and time-consuming process.

Various tasks and processes have been proposed to predict the relationship between gene entities using the databases. Moreover, the interactions between chemicals and proteins/genes can be classified into a range of categories, including metabolic interactions (e.g., substrates, products), inhibition, binding, and induction (5-8). To cover all important biomedical relationships, granular annotation was applied to the DrugProt track. For the BioCreative VII DrugProt track, 13 types of interactions and none interaction will be included.

To address these issues, we propose a BERT-based ensemble learning method with concatenating additional linguistic features. In the first step, the dataset was preprocessed to prepare them for text representation and feature extraction. Second, before the data is input into the BERT model, two features are applied which are the Log-likelihood ratio (LLR) and distance feature. The LLR will identify the keywords that strongly relate to these 13 relations. The key component of this method is that it can use each word within a text segment to determine its LLR value or weight. Correspondingly, the distance feature is a helpful feature that displays the connection of Gene and Chemical pairs from both perspectives, Gene and Chemical. Having this feature will assist the model in recognizing relation possibilities that can occur between Gene and Chemical. As a final step, we finetune BERT-based models by repeating the experiment five times, each combined with the previously mentioned features to enhance performance.

II. Dataset and Method

A. Dataset

We evaluated our method with BioCreative VII track 1 which is publicly available on DrugProt corpus that contains chemical and genetic entities annotations. DrugProt were constructed specifically for drug and chemical-protein interactions text mining. This standard corpus for training and testing chemical and genetic entities' relationships is a public corpus of very novel topics. In particular, DrugProt database contains a set of 3500 titles and abstracts from PubMed, which were identified as containing chemical and genetic entities, and 750 titles and development abstracts from PubMed that were annotated manually by domain experts to identify chemical and genetic interactions in the titles and abstracts (5-8). Additionally, certain sentences that contain chemical and genetic entities relationships were classified into 13 types of relationships. Precisely, it contains three files of datasets, including PubMed abstracts, annotated chemical compound and gene mentions, and annotated chemical compound-gene relations. In these data, the distribution of the dataset is described in Table I.

TABLE I. The distribution of training and development dataset of DrugProt corpus

	Article	Entities	Tokens	Relation
Training	3500	89529	1001168	17288
Development	750	18858	199620	3765

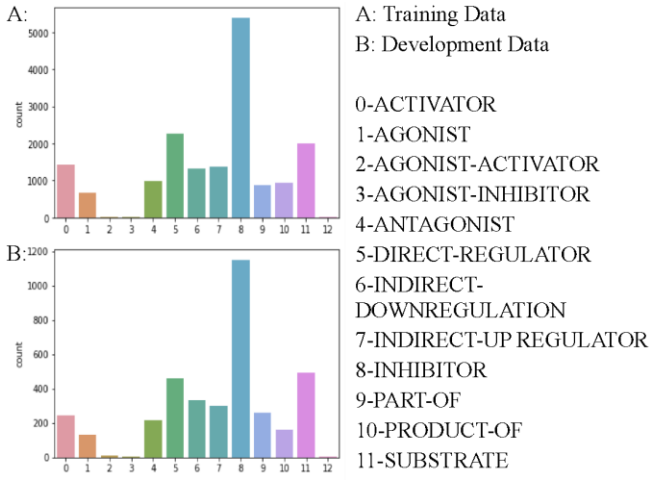

Fig. 1. Distribution of the entities pairs relations.

B. Methods

A visualization of the distribution of relation types in the DrugProt corpus is shown in Figure 1. To investigate the impact on model performance from a larger training dataset, we further merge training and development data. The corpora are utilized to create distance features and a Log-Likelihood Ratio-based scoring scheme to generate the input linguistic features of sentences and output of the relations tagging (9).

For the distance features engineering, based on the original data set, we calculated the relation distances for each genetic and chemical entities pair, and used it as our novel DIST mechanism. Distance feature is a useful feature that presents the relationship between Gene and Chemical pair from both aspects, namely, Gene and Chemical. The feature helps the model learn the correlation of the distance between Gene and Chemical entities in a sentence and their semantic relation. It is later shown in our experiments that this feature greatly improves our performances.

We utilize LLR keyword features for keyword extraction. Due to the size of the dataset, not every word of every sentence is succinct and comprehensive. Extracting keywords is another essential technique to define important features. We adopted the log-likelihood ratio (LLR) (10) to extract those keywords in each relation which are considered as task's classes. LLR will find the keywords which are strongly associated with those relations. In addition, it also can be applied to determine LLR values for every word in the text. Let w be a word and S a relation. Note that S could be 'no relation'. And, k represents the number of abstracts that contain w and S. At the same time, it represents the number of abstracts that include w but not S. Define m as the number of abstracts that contain S but do not contain the word w, and n as the number of abstracts that contain S but do not contain the word w. Furthermore, the greater the LLR value of the term, the more closely connected to the task's relations it is.

$$LLR(w,S) = 2\log\left[\frac{p(w|S)^k(1-p(w|S))^m\, p(w|\neg S)^l(1-p(w|\neg S))^n}{p(w)^{k+l}(1-p(w))^{m+n}}\right]$$

For our implementation, we use BERT with biomedical language representation (BioBERT) to develop the relation extraction model. BERT uses a document-level corpus rather than a shuffled sentence-level corpus to extract long contiguous sequences (7). BioBERT further improves its robustness by including biomedical pre-training data, and is shown to be powerful in biomedical text mining tasks such as protein-protein interaction (PPI) (1).

- *Submission 98*: to find out the proper pre-trained BERT model for the task, we applied hundreds of biomedical-related BERT pre-trained models. We conducted experiments to evaluate the performance under a five-fold cross-validation scheme and found that PubMedBERT is comparable with BioBERT, and BlueBERT tops others by a small margin. Finally, to take advantage of different BERT models, our submission adopted these three models' predictions as an ensemble.
- *Submission 99*: in this submission, we introduce distance feature engineering. It has been shown to further improve the model. To make the predictions more stable, we further compare them with the model without the DIST mechanism. Ultimately, we applied PubMedBERT to accomplish the overall model architecture.
- *Submission 102*: in this experiment, we integrate different batch size settings of Sultan models using ensemble learning with a majority voting

TABLE II. BIOCREATIVE VII DRUGPROT TRACK PERFORMANCE

Submission ID	Testset Micro-avg.	Large Submit No.	Large scale Text Mining sub-track set Micro-avg.
	Precision / Recall / F1-score		Precision / Recall / F1-score
98	0.4697 / 0.1045 / 0.1710	#1	0.4324 / 0.8481 / 0.5728
99	0.5678 / 0.1223 / 0.2013	#2	0.4501 / 0.8286 / 0.5834
102	0.4749 / 0.1005 / 0.1659	#3	0.4372 / 0.7994 / 0.5652
111	0.0138 / 0.0009 / 0.0016		

mechanism. Moreover, we join the prediction results, which are from Sultan, PubMedBERT, and BioBERT by the voting mechanism.

- *Submission 111*: for this experiment, we propose a new BERT model, named Log-Likelihood Ratio (LLR)-based BioBERT. This model not only can be jointly optimized with the contextual recognition pre-trained model via biomedical literature but also can calculate the utterances of chemical and gene entities relations in the function intuitively by embedding an LLR in the sequence. In this submission, we investigate LLR Scoring with PubMedBERT for genetic and chemical entities extraction and prediction.
- *Large Submit No. 1*: the first experiment to predict large scale data, we employed a pre-trained model, PubMedBERT, for the large data prediction. In the data preprocessing, we deploy sentence selection to increase efficiency of entity extraction prediction.
- *Large Submit No. 2*: for this submission, we also build upon the PubMedBERT model. In order to obtain better performances, we further incorporate distance feature engineering to achieve the DIST mechanism.
- *Large Submit No. 3*: in this experiment, we constructed a model with biomedical language representation, the BioBERT model, for the large data prediction.

III. RESULT AND DISCUSSION

We made 5 submissions to Track 1, but the fourth, i.e., submission ID 110, encountered some systematic error. Therefore, we will not discuss this submission below; in addition, the performance of Interaction Identification for our first four submissions ranges from 0.16~20.13% in terms of the F_1-score on aspects of the small test set. This performance is far from what we have observed in our experiments. We speculate that this is due to some undetected errors in our program.

The large scale experiments posed new challenges for our team. First of all, the size of the large-scale dataset is huge. Even with the most basic BioBERT model that we have already trained, it takes 40 hours to make predictions. Moreover, if a BioBERT model with a more complex design structure is applied, it takes about 100 hours to make predictions. Due to hardware limitations, we can produce and upload only 3 predictions from the simpler BERT models. The performances of Interaction Identification for these three submissions range from 56.52 to 58.34% F_1-score on aspects of a large-scale test set. Among them, #2, 'disbert_microsoft', is the best; #3, 'disbert_dmisv1_1', performs the worst. As for the comparison of the model itself, the 'disbert_dmisv1_1' model was pre-trained based on BERT-large-cased (custom 30k vocabulary) and integrated with our novel DIST mechanism, while the 'disbert_microsoft' model was pre-trained by using abstracts from PubMed as well as full-text articles from PubMedCentral, and also with our proposed DIST mechanism, where PubMedCentral consists of 7.3 million articles. The training corpus of the latter is larger than the former. Moreover, the model pre-trained by Microsoft was now on the Biomedical Language Understanding and Reasoning Benchmark. The reason for the better performance could be because the original training data set was very large and not scattered. We found that such a large number of text training models used in the professional field, here are biomedical-related texts and medium tasks, performs better. This phenomenon is especially notable in larger test sets. In the end, we obtained a 45.01% precision, an outstanding 82.86% recall, and 58.34% F_1-score.

IV. CONCLUSION

This work identifies important features for chemical-gene interactions and chemical-protein interactions mentioned in biomedical literature. In addition to biomedical text, we note that the distance feature also provides substantial support of the existence of interaction. In the presence of increasing amounts of real-world data, an efficient interaction extraction model using NLP technology as well as feature engineering is bound to bring out useful information hidden in a huge amount of text. It is possible to even employ machines to learn self-labelling on its own to assist scholars conduct more in-depth research.

ACKNOWLEDGEMENT

This research was supported by the Ministry of Science and Technology of Taiwan under grant MOST 109-2410-H-038 -012 -MY2.

REFERENCES

1. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234-1240.

2. Y.-C. C. N.-W. C. a. W.-L. H. Yu-Lun Hsieh, "Identifying Protein-protein Interactions in Biomedical Literature using Recurrent Neural Networks with Long Short-Term Memory," ACL, vol. 2, p. 240–245, 2017.

3. S. Y. C. N. H. W. Chang YC., "An Interaction Pattern Kernel Approach for Protein-Protein Interaction Extraction from Biomedical Literature. In: Cheng SM., Day MY. (eds) Technologies and Applications of Artificial Intelligence.," TAAI, Lecture Notes in Computer Science, vol. 8916, 2014.

4. C. H. C. Y. C. S. C. C. C. a. W. L. H. Y. C. Chang, "PIPE: a protein-protein interaction passage extraction module for BioCreative challenge.," Database (Oxford), 2016.

5. M. e. a. Krallinger, "Overview of the protein-protein interaction annotation extraction task of BioCreative II.," Genome biology, vol. 9.2, pp. 1-19, 2008.

6. M. e. a. Krallinger, "CHEMDNER: The drugs and chemical names extraction challenge.," Journal of cheminformatics 7.1, pp. 1-11, 2015.

7. M. e. a. Krallinger, "Overview of the BioCreative VI chemical-protein interaction Track.," Proceedings of the sixth BioCreative challenge evaluation workshop, vol. 1, 2017.

8. Antonio Miranda, Farrokh Mehryary, Jouni Luoma, Sampo Pyysalo, Alfonso Valencia and Martin Krallinger. Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations. Proceedings of the seventh BioCreative challenge evaluation workshop. 2021.

9. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Association for computational linguistics. Proceedings of the 2019 Conference of the North, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171–4186.

10. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval (1 edition ed.). New York: Cambridge University Pre.