UTHealth@BioCreativeVII: Domain-specific Transformer Models for Drug-Protein Relation Extraction

Avisha Das, Zhao Li, Qiang Wei, Jianfu Li, Liang-Chin Huang, Yan Hu, Rongbin Li, W. Jim Zheng, Hua Xu
School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, US

***Abstract*— It is important to automatically extract the relations between drugs and proteins from ever-growing biomedical literature, to build up-to-date knowledge bases in biomedicine. Through the DRUGPROT track at BioCreative VII, we developed automated methods to recognize drug-protein entity relations from PubMed abstracts. In this short system description paper, we outline and describe our proposed system submissions that leverage multiple transformer models pre-trained on biomedical data. The outputs of some of the systems have been combined using a decision based on majority voting. Our best system obtained 80.44% in precision and 74.96% in recall for an F1-score of 77.60%, demonstrating the effectiveness of deep learning-based approaches for automatic relation extraction from biomedical literature for the main track. We also participated in the Large-Scale Track - the micro-averaged precision, recall and F1-score of our best system being 79.49%, 75.27% and 77.32% respectively. (*Abstract*)**

Keywords— Deep learning; Relation Extraction; BERT; Ensemble Learning

I. INTRODUCTION

In the past few decades, biomedical literature has grown exponentially and much new knowledge is embedded in narrative texts in biomedical articles, requiring automated methods such as natural language processing (NLP) to extract and normalize them into computable information. Among diverse biomedical entities, drugs and proteins, as well as their relations, are important for many applications such as drug repurposing (1) and drug combination (2, 3) studies. Therefore, recognition of drug-protein entities and relations from biomedical medical literature has received great attention in the past few years. Automated relation extraction between gene/protein entities from text previously used methods like parsing (4, 5), diverse set of features (6, 7) along with deep learning networks (8). BioCreativeVI Track 5 (14) was organized in 2017 to predict Chemical-Protein relations on the ChemProt Corpus.

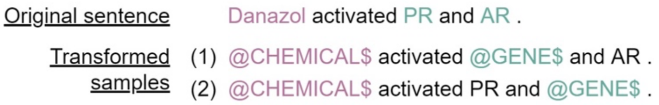
In 2021, the BioCreativeVII Track 1 (15) corpus aims to promote the development and evaluation of systems that are able to automatically detect relations between drugs and proteins from PubMed abstracts. This competition has two sub-tracks - the main track and a large-scale track. In this paper, we provide brief descriptions of our approaches and results for this task.

II. METHODOLOGY

A. Dataset

The organizers provided two different datasets - DrugProt corpus for the main track and an un-labelled test corpus for the large-scale track. In the main track, the DRUGPROT corpora consists of 15,000 PubMed abstracts and titles along with PMIDs, essentially published between 2005 and 2014. Of these, the data was split into three subsets - training (3,500 abstracts), development (750 abstracts) and test (10,750 abstracts). For the large scale sub-track, 23, 66081 records are provided with a total of 53,993,602 entity annotations. This does not have any entity relations labelled.

- *Preprocessing*: We used the tool called CLAMP [9] for sentence boundary detection.
- *Representation*: Each chemical and gene in a sentence will be made into a candidate relation pair for classifying. Also, text of entity will be replaced into its semantic type. For example, in Figure 1, there are 2 genes and 1 chemical, so totally 2 candidate relation pairs are generated.
- *Training sets*: Because the estimation of the parameters in NLP models is sensitive to the number of instances in the training corpus, we trained the models using as many annotated instances as possible. We pooled the 3,500 abstracts in the DrugProt training set and 750 abstracts in the development set and then randomly split them into ten folds. Each fold contains 425 abstracts and serves as a re-built development set once during 10-fold cross-validation.



Original sentence Danazol activated PR and AR .

Transformed samples (1) @CHEMICAL$ activated @GENE$ and AR .
(2) @CHEMICAL$ activated PR and @GENE$.

Figure 1. Data representation.

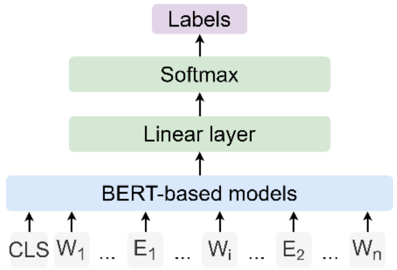


Figure 2. The architecture of the system.

B. Biomedical BERT-based models

Figure 2 shows the architecture of our system. First candidate relation pairs were generated as input (see representation in section A). Here BERT based models (BioM-BERT, BioM-ALBERT, BioBERT and PubMed-BERT) were used, and a linear classification layer was added on top to predict the label of a candidate pair, where the [CLS] vector from output of BERT was sent to a classification layer for prediction.

We aplied the pre-trained BioM-BERT model from (10), the implementation of ELECTRA model trained on the corpora of PMC and PubMed articles. We also applied the BioM-ALBERT model (10) on the DrugProt task. This model was firstly pre-trained on PubMed Abstracts only for 264K steps with a batch size of 8192 based on ALBERTxxlarge. Then it was continuously pretrained on PMC full articles for further 64K steps to investigate the influence of adding PMC articles on the language model. The pretrained BioM-ALBERTxxlarge-PMC model was fine-tuned on two different masked input files: one masked file differentiated the overlapped chemical and gene entities, resulted the fine-tuned model of BioM-ALBERT 1; another mask file ignored the overlapped chemical and gene entities, resulted the fine-tuned model of BioM-ALBERT 2.

The pre-trained biomedical language model, BioBERT (11), is applied to predict the relationship of a given candidate drugprotein pair. The BioBERT model has the same architecture and vocabulary set as BERT, and is initialized with the weights from BERT, a pretrained model on general domain. Then the model is finetuned on PubMed abstracts and PMC full-text literature. The performance gets a significant boost in various biomedical text mining tasks compared with BERT. In this challenge, we adopt the most recently pretrained BioBERT large model.

PubMedBERT (12), another pre-trained BERT model developed specifically for biomedical NLP tasks, was also applied in this challenge. Different from BioBERT, PubMedBERT model, though employing the similar BERT-base architecture, is trained from the scratch with the specifically constructed vocabulary sets. The result demonstrated PubMedBERT achieved consistent superior performance than the continual pre-training language model. Therefore, for this track, we employ five different BERT-based models - BioM-ALBERT 1, BioM-ALBERT 2, BioBERT, BioM-BERT and PubMedBERT.

C. Training Models

Scaling our models to get the desirable set of results on the Large Scale subtrack was challenging. We ran the BioM-BERT model on the A100-SXM4 GPU. Each fold ran for a total of 53.5 hours approximately.

D. Ensemble learning

Based on the 50 prediction results from the five BERT-based models trained by ten different training sets, we combined them and further developed two ensemble learners: voting and stacking.

•***Majority voting***: Based on the order of combining the results from different BERT models and training folds, we developed three strategies for majority voting: “fold-first”, “model-first”, and “overall”. In the fold-first majority voting, we first combined the results from the ten folds for each BERT model, kept the relations having no less than five votes, and then pooled these voting results from the five models and kept the relations having no less than three votes. While in the model-first majority voting, we first combined the results from the five models for each training fold, kept the relations having no less than three votes, and then pooled these voting results from the ten folds and kept the relations having no less than five votes. The overall majority voting is to pool the 50 prediction results from the five models for the ten folds and then kept the relations having no less than 25 votes.

•***Weighted majority voting***: We also developed the same strategies for weighted majority voting. We gave each vote a different weight according to the performance of its relation type in different training sets.

In the fold-first weighted majority voting, the weight is calculated as:

$$\omega_{mr} = \frac{\sum_{f=1}^{10} p_{mrf} F_{mrf}}{\sum_{f=1}^{10} F_{mrf}} \quad , \qquad (1)$$

where m represents the type of model, r is the relation type, f denotes the index of fold, and Fmrf represents the performance (F1 score) of predicting the relation r by the model m in the fold f. pmrf is 0 if the relation r between the chemical and the protein is predicted negative by the model m in the fold f, while pmrf is 1 if the relation is predicted positive. We kept the relation if its ωmr is no less than 0.5.

In the model-first weighted majority voting, the weight is calculated as:

$$\omega_{rf} = \frac{\sum_{m=1}^{5} p_{mrf} F_{mrf}}{\sum_{m=1}^{5} F_{mrf}} \quad , \qquad (2)$$

and we kept the relation if its ωrf is no less than 0.5. In the overall weighted majority voting, the weight is calculated as:

$$\omega_{r} = \frac{\sum_{m=1}^{5} \sum_{f=1}^{10} p_{mrf} F_{mrf}}{\sum_{m=1}^{5} \sum_{f=1}^{10} F_{mrf}} \quad , \qquad (3)$$

and we kept the relation if its ωrf is no less than 0.5.

•***Stacking***: Using the prediction results from the five BERT-based models as binary features (0 for negative and 1 for positive) for each chemical-protein combination, we trained a J48 decision tree by WEKA [13] with default settings for each training set. After implementing the stacking models to the DrugProt test set, we pooled the results from each training set and then kept the relations having no less than five votes.

III. Results and discussion

The prediction performances (F-1 scores) of the five BERT-based models and three ensemble learners for the ten development sets created from the training data, are shown in Table I. The results show that the overall performances are from 0.753 (BioBERT) to 0.792 (model-first weighted majority voting). Moreover, all three ensemble learners have better performances than every single BERT-based model, showing improved prediction performance using ensemble learning.

According to the performance shown in the development sets, we chose five models for the test set: fold-first weighted majority voting (61,850 relations), model-first weighted majority voting (62,398 relations), overall majority voting (61,723 relations), stacking (60,786 relations), and BioM-ALBERT 1 (62,512 relations). In total, there are 67,374 unique relations. The overlaps among the five results are shown in Figure 3. Although BioM-ALBERT 1 showed a lower prediction performance in development sets, Figure 3 shows that it provides the most unique predictions (2,311 relations) in the test set. The results on the main track and large-scale track of the BioCreative VII track are shown in Tables III and IV respectively.

We also provide the results of the top 5 submission runs on the main track and large-scale subtrack test sets in Tables II and III respectively. Based on the macro-averaged F1-score, we see that the Majority voting algorithms perform the best in each case. While the model-first Majority Voting has the highest score on the main track (77.6%), the fold-first MV algorithm was the top scorer (77.3%) on the large-scale subtrack.

Subsequently, we also report the results by the relation level granularity in Tables V and VI for the main track and large-scale subtrack. We report the F1-scores for each algorithm's performance in both cases and see the similar trend in the prediction values.

IV. CONCLUSION

We briefly outline the submitted systems for the BioCreative VII DRUGPROT Track. The results demonstrate domain-specific transformer models achieve reasonable performance on the drug-protein extraction task. Our ensemble system can further improve its performance with majority voting-based ensemble methods performing the best.

ACKNOWLEDGEMENTS

This work is partly supported by the Cancer Prevention and Research Institute of Texas through grant RP170668 (Dr. Wenjin J. Zheng), the National Institutes of Health (NIH) through grants 1UL1TR003167 and R01AG066749 (Dr. Wenjin J. Zheng), and NIA R56AG069880 (Dr. Hua Xu).

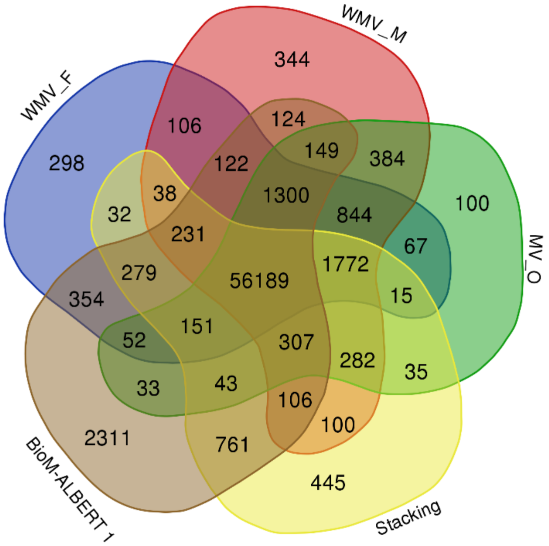


Figure 3. Overlaps among the five submissions for the test set. WMV_F: fold-first weighted majority voting; WMV_M: model-first weighted majority voting; MV_O: overall majority voting.

CONFLICT OF INTEREST STATEMENT

Dr. Hua Xu and the University of Texas Health Science Center at Houston have research financial interest at Melax Technologies Inc.

References

[1] E. Anderson, T. M. Havener, K. M. Zorn, D. H. Foil, T. R. Lane, S. J. Capuzzi, D. Morris, A. J. Hickey, D. H. Drewry, and S. Ekins, "Synergistic drug combinations and machine learning for drug repurposing in chordoma," *Scientific reports*, vol. 10, no. 1, pp. 1–10, 2020.

[2] R. Xu and Q. Wang, "Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing," *BMC bioinformatics*, vol. 14, no. 1, pp. 1–11, 2013.

[3] I. N. Dewi, S. Dong, and J. Hu, "Drug-drug interaction relation extraction with deep convolutional neural networks," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2017, pp. 1795–1802.

[4] K. Fundel, R. Kuffner, and R. Zimmer, "Relex—relation extraction using¨ dependency parse trees," *Bioinformatics*, vol. 23, no. 3, pp. 365–371, 2007.

[5] A. Koike, Y. Niwa, and T. Takagi, "Automatic extraction of gene/protein biological functions from biomedical text," *Bioinformatics*, vol. 21, no. 7, pp. 1227–1236, 2005. vol. 2018, 2018.

[6] Y. Peng, A. Rios, R. Kavuluru, and Z. Lu, "Extracting chemical–protein relations with ensembles of svm and deep learning models," *Database*,

[7] P.-Y. Lung, Z. He, T. Zhao, D. Yu, and J. Zhang, "Extracting chemical–protein interactions from literature using sentence structure analysis and feature engineering," *Database*, vol. 2019, 2019.

[8] Y. Peng, A. Rios, R. Kavuluru, and Z. Lu, "Chemical-protein relation extraction with ensembles of svm, cnn, and rnn models," *arXiv preprint arXiv:1802.01255*, 2018.

[9] E. Soysal, J. Wang, M. Jiang, Y. Wu, S. Pakhomov, H. Liu, and H. Xu, "Clamp–a toolkit for efficiently building customized clinical natural language processing pipelines," *Journal of the American Medical Informatics Association*, vol. 25, no. 3, pp. 331–336, 2018.

[10] S. Alrowili and K. Vijay-Shanker, "Biom-transformers: Building large biomedical language models with bert, albert and electra," in *Proceedings*

of the 20th Workshop on Biomedical Language Processing, 2021, pp. 221–227.

[11] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[12] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," 2020.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[14] Krallinger, M., Rabal, O., Akhondi, S.A., Pérez, M.P., Santamaría, J., Rodríguez, G.P., Tsatsaronis, G., Intxaurrondo, A., López, J.A., Nandal, U.K., Buel, E.M., Chandrasekhar, A., Rodenburg, M., Lægreid, A., Doornenbal, M.A., Oyarzábal, J., Lourenço, A., & Valencia, A. "Overview of the BioCreative VI chemical-protein interaction Track". 2017

[15] 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.

TABLE II MODEL PERFORMANCE ON THE MAIN TRACK TEST SET

Run ID	Run Name	Precision	Recall	F1
1	1-Voting w FM	0.795	0.750	0.77193
2	2-Voting w MF	0.804	0.750	0.776031
3	3-Voting	0.800	0.746	0.771763
4	4-Stacking	0.800	0.733	0.765181
5	5-BioM-ALBERT 1	0.797	0.753	0.774488

TABLE III MODEL PERFORMANCE ON THE LARGE-SCALE TEST SET

Run ID	Run Name	Precision	Recall	F1
1	1-BioM-ALBERT 1	0.763804	0.713467	0.737778
2	2-Stacking	0.77619	0.747278	0.76146
3	3-Voting w FM	0.794856	0.752722	0.773216
4	4-Voting w MF	0.800799	0.746418	0.772653
5	5-Voting	0.797194	0.748997	0.772345

TABLE IV PREDICTION PERFORMANCE IN F1-SCORE BY RELATIONS ON MAIN TRACK (MT) AND LARGE-SCALE (LS) TEST SETS

Relation-Type	**BioM- ALBERT 1**		**Stacking**		**Voting w FM**		**Voting w MF**		**Voting**	
	MT	**LS**	**MT**	**LS**	**MT**	**LS**	**MT**	**LS**	**MT**	**LS**
ACTIVATOR	0.81	0.74	0.79	0.81	0.82	0.81	0.81	0.81	0.82	0.81
AGONIST	0.78	0.70	0.78	0.75	0.78	0.79	0.78	0.78	0.77	0.79
AGONIST-INHIBITOR	1.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	1.00
ANTAGONIST	0.91	0.84	0.90	0.87	0.90	0.90	0.90	0.89	0.90	0.90
DIRECT-REGULATOR	0.67	0.65	0.66	0.69	0.68	0.69	0.68	0.69	0.68	0.68
INDIRECT-DOWNREGULATOR	0.76	0.74	0.75	0.76	0.77	0.77	0.78	0.77	0.78	0.77
INDIRECT-UPREGULATOR	0.77	0.74	0.76	0.74	0.76	0.76	0.77	0.76	0.76	0.76
INHIBITOR	0.87	0.84	0.85	0.85	0.85	0.85	0.86	0.85	0.86	0.85
PART-OF	0.68	0.67	0.69	0.69	0.69	0.71	0.70	0.70	0.70	0.71
PRODUCT-OF	0.70	0.63	0.67	0.60	0.69	0.67	0.68	0.67	0.69	0.68
SUBSTRATE	0.65	0.60	0.65	0.65	0.65	0.66	0.67	0.66	0.64	0.66
SUBSTRATE_PRODUCT-OF	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AGONIST-ACTIVATOR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

TABLE I PREDICTION PERFORMANCE ON THE DEVELOPMENT SETS

Model					**Development sets**						**Overall**
	1	2	2	4	5	6	7	8	9	10	
BioBERT	0.766	0.760	0.717	0.803	0.707	0.771	0.759	0.740	0.724	0.756	0.753
BioM-ALBERT 1	**0.775**	0.784	0.711	0.828	0.719	0.806	0.786	0.777	0.739	**0.811**	0.777
BioM-ALBERT 2	0.768	0.776	0.718	0.828	0.702	0.773	**0.845**	0.767	0.742	0.763	0.769
BioM-BERT	0.760	0.776	0.731	0.828	0.688	0.767	0.794	0.785	0.754	0.785	0.769
PubMedBERT	0.761	0.778	0.708	0.815	0.705	0.766	0.796	0.797	0.720	0.776	0.765
Model-first majority voting	0.771	**0.806**	0.757	**0.844**	0.728	**0.808**	0.818	0.787	**0.763**	0.805	0.791
Model-first weighted majority voting	0.774	**0.806**	**0.759**	**0.844**	**0.730**	**0.808**	0.818	0.789	**0.763**	0.806	**0.792**
Stacking	0.767	0.797	0.735	0.838	0.717	0.762	0.789	**0.811**	0.750	0.797	0.779