MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231043

Extracting Drug-Protein Relation from Literature Using Ensembles of Biomedical Transformers

Avisha DAS[a], Zhao LI[a], Qiang WEI[a], Jianfu LI[a], Liang-chin HUANG[a], Yan HU[a], Rongbin LI[a], Wenjin Jim ZHENG[a,1] and Hua XU[a]
[a] *School of Biomedical Informatics, University of Texas Health Science Center at Houston;Yale University; Melax Technologies, Houston*

Abstract. Automatic extraction of relations between drugs/chemicals and proteins from ever-growing biomedical literature is required to build up-to-date knowledge bases in biomedicine. To promote the development of automated methods, BioCreative-VII organized a shared task – the DrugProt track, to recognize drug-protein entity relations from PubMed abstracts. We participated in the shared task and leveraged deep learning-based transformer models pre-trained on biomedical data to build ensemble approaches to automatically extract drug-protein relation from biomedical literature. On the main corpora of 10,750 abstracts, our best system obtained an F1-score of 77.60% (ranked 4th among 30 participating teams), and on the large-scale corpus of 2.4M documents, our system achieved micro-averaged F1-score of 77.32% (ranked 2nd among 9 system submissions). This demonstrates the effectiveness of domain-specific transformer models and ensemble approaches for automatic relation extraction from biomedical literature.

Keywords. Deep Learning, Drug-protein relation extraction, BERT, Ensemble Learning, Pubmed abstracts.

1. Introduction

Automatic identification of biomedical entities and relations from literature have gained a lot of attention from the research community. These diverse entities include names of chemical compounds, drugs, gene products (including genes, proteins, and miRNAs), etc. Mining relations between drugs and proteins from existing biomedical knowledge is useful for many applications such as drug discovery, combination, and repurposing studies [1,2]. However, the volume of biomedical literature has grown exponentially, but manually mining relevant information from biomedical literature for downstream applications, is tedious and time-consuming – therefore making automated methods indispensable for mining and processing documents from databases like PubMed. Such systems make use of natural language processing (NLP) techniques to extract and normalize content into computable information. Tasks like drug-protein entity recognition and relation extraction from biomedical text previously used methods like parsing [3,4], diverse set of features and deep learning models [5,6-9]. Bio Creative-VII Track 1[10] provides the Drug Prot corpus in an aim to promote the development and evaluation of systems that can automatically detect relations between drugs and proteins

[1] Corresponding author: Wenjin Jim Zheng, email:Wenjin.J.Zheng@uth.tmc.edu.

from PubMed abstracts. The DrugProt track focused on the evaluation of automatic systems able to extract 13 different types of drug-genes/protein relations used to understand gene regulatory and pharmacological mechanisms. This competition has two sub-tracks – (a) Main track, with 10,750 test instances, and (b) Large-scale track, of ~2.4M records provided as test set. A total of 30 teams participated in the main track, of them 9 teams submitted their system predictions for the Large-scale track [10]. We participated in the DrugProt track and leveraged deep transformer models with attention masking and trained on biomedical literature for the relation extraction. In addition to fine-tuned domain specific BERT-based models, we combine the model predictions using another layer of ensemble-based learning using two schemes: majority-based voting and stacking. In this paper, we show how ensemble learning with majority voting-based prediction scoring outperformed the baseline domain specific BERT.

2. Methods

2.1. Dataset

The organizers of BioCreative-VII provided participants with two different datasets for the main track and the large-scale subtrack of DrugProt corpus [10]. The corpus provided during the main track consisted of 15,000 PubMed abstracts, titles, and corresponding article PubMedIDs of articles published between years 2005 and 2014. The statistics of the DrugProt have been shown in Table 1. For a more detailed description, we refer the readers to the task overview paper by the organizers [10].

Based on the number of training instances/examples available and the consistency in the manual annotation process, 12 relation types were annotated. The classes with highest number of relations – Part-Of and Inhibitor, classes Agonist-Inhibitor and Direct-Regulator had the least. We performed the following steps for preprocessing the dataset prior to training our models – (a) *Preprocessing*: We used the tool called CLAMP [11] for sentence boundary detection; (b) *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 2 candidate relation pairs are generated in total.

Table 1. Statistics of the DrugProt Dataset. M: Manual, A: Automated.

Set (Annotation Type)	# of abstracts	# of entities		# of relations
		Gene	Chemical	
Training (M)	3,500	43,255	46,274	17,274
Development (M)	750	9,005	9,853	3,761
Test (M)	750	9,515	9,434	3,491
Background (A)	10,000	1,57,523	1,34,333	--
Large Scale (A)	23,66,081	3,35,78,479	2,04,15,123	--

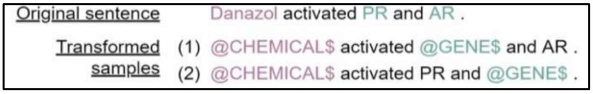
Original sentence
Danazol activated PR and AR .
Transformed samples
(1) @CHEMICAL$ activated @GENE$ and AR .
(2) @CHEMICAL$ activated PR and @GENE$.

Figure 1. Data representation.

2.2. Biomedical BERT-based Models and Ensemble Learning

<u>BERT-based Models:</u> We use Bidirectional Encoder Representations from Transformers (BERT)-based models [11] for our system implementation. We use the domain-specific BERT models that have been trained on biomedical literature, specifically PubMed and

PMC articles – BioM-BERT model; variations of pretrained BioM-ALBERT$_{\text{xxlarge-PMC}}$ model was fine-tuned on two different masked input files: one masked file differentiated the overlapped chemical and gene entities (BioM-ALBERT-1); and another mask file ignored the overlapped chemical and gene entities (BioM-ALBERT-2); pre-trained BioBERT, finetuned on PubMed abstracts and PMC full-text literature; finally, PubMedBERT, training the BERT architecture from scratch on biomedical literature. Since the training and development datasets provided by the organizers are smaller in size with 4,250 instances and a much larger test data of 10,750 abstracts, we use a cross-validation technique to train our baseline models to train the models using as many annotated instances as possible. We pooled the training and development sets (4,250 abstracts) and then randomly split them into 10 folds. Scaling our models to get the desirable set of results on the Large Scale subtrack was challenging.

Ensemble Learning: We combined the 50 prediction results from the five BERT-based models trained by ten different training sets of the main track, and further developed two sets of ensemble learners: voting and stacking. For *Majority voting,* we developed three strategies: “fold-first”, “model-first”, and “overall”, based on the order of combining the results. *Weighted majority voting* uses the same combination strategies as the majority voting, but each vote was given a different weight based on the performance of its relation type in different training sets. Finally, *Stacking* uses 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 with default settings for each training set.

3. Results

For model evaluation, organizers' metrics and library are utilized [10]. Teams are ranked based on micro-averaged precision, recall, and balanced micro F1-score. Among 30 main track teams, we ranked 4th (F1: 77.6%). In the large-scale subtrack, among 9 teams, we ranked 2nd (F1: 77.3%). Table 2 presents F1 scores for five BERT-based models and three ensemble learners across ten development sets from training data. All ensemble learners outperform individual BERT models, indicating improved predictions through ensemble learning. BioM-ALBERT 1 excels among BERT models. For the test set submission, five models are chosen: Weighted majority voting (fold-first and model-first), overall majority voting, stacking-based ensemble, and BioM-ALBERT.

Table 3 displays top 5 models' results in DrugProt track's main and large-scale tasks. Majority voting (MV) algorithms lead in both cases, with model-first MV scoring highest on main track (77.6%), and fold-first MV topping large-scale subtrack (77.3%). The performance achieved is notable, since the test datasets for the main and large-scale subtracks consist of 10,750 and 2,366,081 abstracts, while the models were trained on mentions from 4,200 abstracts. Subsequently, we also report the results by the relation level granularity in Table 4 for the main track and large-scale subtracks. Evaluating the systems on a per-class basis gives an in-depth view of how the top 5 classifiers performed for each relation extraction.

Table 2. Performance of the models (using F1-score) on the development sets during training.

Model Type	Model	Development sets										
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	**Overall**
Deep Learning (BERT-based)	BioBERT	0.766	0.76	0.72	0.803	0.71	0.771	0.759	0.74	0.724	0.756	0.753
	BioMALBERT 1	**0.775**	0.784	0.711	0.828	0.72	0.806	0.786	0.777	0.739	**0.810**	**0.777**
	BioMALBERT 2	0.768	0.776	0.718	0.828	0.7	0.773	**0.845**	0.767	0.742	0.763	0.769
	BioM-BERT	0.76	0.776	0.731	0.828	0.69	0.767	0.794	0.785	0.754	0.785	0.769
	PubMedBERT	0.761	0.778	0.708	0.815	0.71	0.766	0.796	0.797	0.72	0.776	0.765
Ensemble Learning	Majority voting	0.771	**0.806**	0.757	**0.844**	0.73	**0.808**	0.818	0.787	**0.763**	0.805	**0.791**
	Weighted majority voting	0.774	**0.806**	**0.759**	**0.844**	**0.73**	**0.808**	0.818	0.789	**0.763**	0.806	**0.792**
	Stacking	0.767	0.797	0.735	0.838	0.72	0.762	0.789	**0.81**	0.75	0.797	**0.779**

Table 3. Model Performance on the Main Track Test Set and the Large-Scale Test Set.

Run_ID	Main Track Test Set				Large-Scale Test Set			
	Run Name	**Precision**	**Recall**	**F1**	**Run Name**	**Precision**	**Recall**	**F1**
1	Voting/FM	0.795	0.75	0.772	BioM-AB1	0.764	0.714	0.738
2	Voting/MF	**0.804**	**0.75**	**0.776**	Stacking	0.776	0.747	0.761
3	Voting	0.8	0.746	0.772	Voting/FM	0.795	0.753	0.773
4	Stacking	0.8	0.733	0.765	Voting/MF	**0.801**	**0.746**	**0.773**
5	BioM-AB 1	0.797	0.753	0.775	Voting	0.797	0.749	0.772

Table 4. 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	0.7	0.78	0.75	0.78	0.79	0.78	0.78	0.77	0.79
AGONIST-INHIBITOR	1	1	1	1	0	0	1	1	1	1
ANTAGONIST	0.91	0.84	0.9	0.87	0.9	0.9	0.9	0.89	0.9	0.9
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	0.7	0.7	0.7	0.71
PRODUCT-OF	0.7	0.63	0.67	0.6	0.69	0.67	0.68	0.67	0.69	0.68
SUBSTRATE	0.65	0.6	0.65	0.65	0.65	0.66	0.67	0.66	0.64	0.66
SUBSTRATE_PRODUCT-OF	0	0	0	0	0	0	0	0	0	0

4. Discussion

Biomedical BERT models show promising results in drug-protein extraction with majority voting and ensemble learning enhancing performance on a limited training dataset. Achieving over 77% F1-score on larger test data is notable. Model-first and Fold-first Majority Voting excel with 77.6% and 77.3% F1-scores. Notably, AGONIST-INHIBITOR class lacks test instances, impacting evaluation. ANTAGONIST and INHIBITOR classes yield strong 87% and 85% average F1-scores, contributing to overall performance due to their prevalence in training data. However, zero F1-score for SUBSTRATE_PRODUCT-OF class suggests potential overfitting on specific classes like ANTAGONIST or INHIBITOR, highlighting the need for contextual understanding. Future work involves increased training data, exploring entity features, and leveraging expert feedback for better relation extraction. Integrating additional resources, knowledge bases, and biomedical abstracts holds promise for model enhancement.

5. Conclusions

We propose a computational pipeline by leveraging domain-specific transformer-based deep neural architectures to extract relations between drugs and diseases from large corpora of medical literature provided by the organizers of the BioCreative-VII DrugProt Track. The main contribution of this work is the identification of fine-tuned BERT-based models for relation classification, followed by their combination through ensemble learning for better model performance. Second, the models were trained on a limited set of instances (from 4,200 abstracts) but performed notably well, with a micro-averaged F1-score of 77.3% on the test set of the main sub-track (10,750 abstracts) and 77.6% on the large-scale sub-track (~2.4 million abstracts). This demonstrates the high predictability power of fine-tuned BERT-based models and their increased performance when combined through ensembling methods.

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). Dr. Avisha Das is a CPRIT Postdoctoral Fellow in the Biomedical Informatics, Genomics, and Translational Cancer Research Training Program (BIG-TCR) funded by the Cancer Prevention & Research Institute of Texas (CPRIT RP210045).

References

[1] Anderson E, Havener TM, Zorn KM, Foil DH, Lane TR, Capuzzi SJ, Morris D, Hickey AJ, Drewry DH, Ekins S. Synergistic drug combinations and machine learning for drug repurposing in chordoma. Sci Rep. 2020 Jul;10(1):1-0.

[2] Xu R, Wang Q, Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC Bioinform. 2013 Dec;14(1):1-1, doi: 10.1186/1471-2105-14-181.

[3] Fundel K, Küffner R, Zimmer R. RelEx—Relation extraction using dependency parse trees. Bioinformatics. 2007 Feb;23(3):365-71, doi: 10.1093/bioinformatics/btl616.

[4] Koike A, Niwa Y, Takagi T. Automatic extraction of gene/protein biological functions from biomedical text. Bioinformatics. 2005 Apr;21(7):1227-36, doi: 10.1093/bioinformatics/bti084.

[5] Peng Y, Rios A, Kavuluru R, Lu Z. Extracting chemical–protein relations with ensembles of SVM and deep learning models. Database. 2018 Jan;2018, doi: 10.1093/database/bay073.

[6] Luo L, Lai PT, Wei CH, Lu Z. A sequence labeling framework for extracting drug-protein relations from biomedical literature. Database. 2022 Jul;2022, doi: 10.1093/database/baac058.

[7] Luo L, Lai PT, Wei CH, Lu Z. Extracting Drug-Protein Interaction using an Ensemble of Biomedical Pre-trained Language Models through Sequence Labeling and Text Classification Techniques. in Proceedings of the BioCreative VII challenge evaluation workshop. 2021; 26-30.

[8] Weber L, Sänger M, Garda S, Barth F, Alt C, Leser U. Humboldt@ drugprot: Chemical-protein relation extraction with pretrained transformers and entity descriptions. in Proceedings of the BioCreative VII challenge evaluation workshop. 2021.

[9] Yoon W, Yi S, Jackson R, Kim H, Kim S, Kang J. Using knowledge base to refine data augmentation for biomedical relation extraction. in Proceedings of the BioCreative VII challenge evaluation workshop. 2021; 31-35.

[10] Miranda A, Mehryary F, Luoma J, Pyysalo S, Valencia A, Krallinger M. Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations. in Proceedings of the seventh BioCreative challenge evaluation workshop. 2021; 11-21.

[11] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018 Oct, doi: 10.48550/arXiv.1810.04805.