BioCreative VII-Track 1: A BERT-based System for Relation Extraction in Biomedical Text

Darshini Mahendran[1]*, Sudhanshu Ranjan[2], Jiawei Tang[2], Mai H. Nguyen[3,] and Bridget T. McInnes[1]

[1]Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA, [2]Department of Computer Science & Engineering, University of California San Diego, La Jolla, CA, USA, [3]San Diego Supercomputer Center, University of California San Diego, La Jolla, CA, USA

***Abstract*—This paper describes our team's participation in Track 1 of the BioCreative VII challenge to automatically detect relations between chemical compounds/drugs and genes/proteins. Here, we discuss the three contextualized language-based models with different input representations: two general Bidirectional Encoder Representations from Transformers (BERT)-based models and a BioBERT-based model. Our best model for this task achieved an overall Precision of 0.55, Recall of 0.52, and an F_1 score of 0.54 on the test set.**

Keywords—Natural Language Processing (NLP); Relation Extraction (RE); Biomedical text; Contextualized language model; BERT.

I. Introduction

Biomedical literature connects several types of users, including biomedical researchers, clinicians, and database curators, as they share their findings in articles, patents, or reports. However, the exponential growth of the literature makes it difficult for users to retrieve information efficiently in a timely manner. Therefore, there is an increasing need to develop Natural Language Processing (NLP) systems to automatically extract relevant information for users, reducing the time it takes them to identify and extract the information manually (1). NLP is an area of research focused on developing algorithms to allow the computer to process and analyze unstructured language. One such area is Relation Extraction (RE), which identifies relationships between entities in a text.

A considerable amount of existing systems focus on recognizing mentions of genes/proteins and chemicals in text automatically, but a limited number of approaches focus on extracting interactions between them (2). Therefore, it is necessary to study the different types of relationships of drugs and chemical compounds with certain biomedical entities, particularly genes and proteins, and their systematic extraction to analyze and explore key biomedical properties in biomedical applications (3).

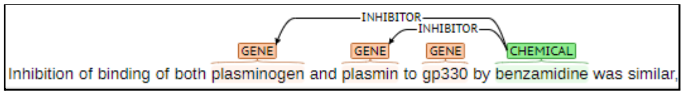
In this paper, we describe our participation in the Biocreative VII Track 1 (3), whose task is to automatically identify the relationship between chemical compounds with genes in biomedical literature. We explored three variations of the Bidirectional Encoder Representations from Transformers (BERT) architectures (5).

II. Related Work

BioCreative VI Task 5 (2) introduced a similar task to automatically detect relations between chemical compounds/drugs and genes/proteins in PubMed[1] abstracts, and they released a manually annotated corpus, the CHEMPROT (2). Peng, et al. (11) developed an ensemble of three systems: Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN). The output is combined using a decision based on majority voting or stacking. Antunes, et al. (12) used a CNN and a Bidirectional Long Short-term Memory (Bi-LSTM) together with a very narrow representation of the relation instances, using a few words from the shortest dependency path and the respective dependency edges. Yuksel, et al. (13) presented a CNN model and used word-embeddings and distance embeddings to represent a potential relation. Sun, et al. (14) proposed a novel Deep-contextualized stacked Bi-LSTM model (DS-LSTM), which consists of deep contextualized word representations, the entity attention mechanism, and stacked Bi-LSTMs. Sun, et al. (15) proposed a novel hierarchical recurrent CNN (Hierarchical RCNN)-based approach to learn latent features from short context subsequences efficiently. Liu, et al. (16) used CNNs and attention-based RNNs, to extract chemical protein relationships. Hafiane, et al. (17) explored various BERT-based architectures and transfer learning strategies for biomedical RE.

III. Data

We evaluate our models on the Biocreative VII Track 1 DrugProt corpus (3). The training set contains chemical mentions (46274), gene/protein mentions (43255), and drug/chemical-protein/gene interactions (17288) from 3500 PubMed abstracts. The development and test set includes 750 and 10750 abstracts, respectively. Fig. 1. shows the Brat Rapid Annotation Tool (BRAT) annotation of the entities and relations of a sentence from the dataset.



Inhibition of binding of both plasminogen and plasmin to gp330 by benzamidine was similar,

Fig.1. An example of a BRAT annotated sentence from the training dataset

[1] https://www.ncbi.nlm.nih.gov/pubmed/

Table I shows the number of instances for each relation type in the training and development datasets.

TABLE I. Relation type statistics of DrugProt corpus

Annotated relations statistics		
	Training set	***Development set***
INDIRECT-DOWNREGULATOR	1330	332
INDIRECT-UPREGULATOR	1379	302
DIRECT-REGULATOR	2250	458
ACTIVATOR	1429	246
INHIBITOR	5392	1152
AGONIST	659	131
AGONIST-ACTIVATOR	29	10
AGONIST-INHIBITOR	13	2
ANTAGONIST	972	218
PRODUCT-OF	921	158
SUBSTRATE	2003	495
SUBSTRATE_PRODUCT-OF	25	3
PART-OF	886	258
TOTAL	17288	3765

IV. Methods

In this section, we describe the three models we developed for chemical-gene RE. Fig. 2. shows the architecture of our overall system.

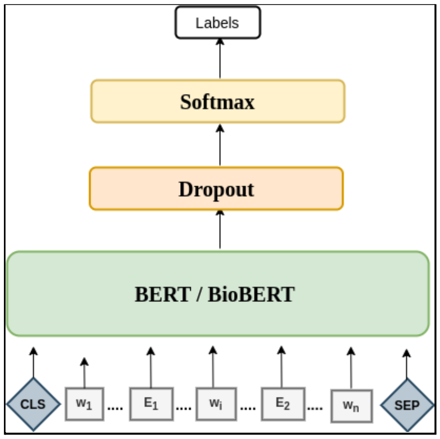


Fig.2. Architecture that represents our overall system

BERT is an NLP model introduced by Google in 2018 (4). BERT is a transformer (8) that utilizes attention mechanisms to learn the contextualized semantic relations between words of a text. The encoder reads the input as the sum of token, segmentation, and position embeddings. BERT is the first deep bidirectionally trained language model that learns the representation of a word based on its context. The general BERT models are trained on a large corpus of English data: Book-Corpus (800M words) and Wikipedia (2,500M words) in a self-supervised manner to serve as a general-purpose language representation model. In this work, we also explore BioBERT, which is general BERT further pre-trained over a corpus of biomedical research articles from PubMed abstracts and article full texts for biological text mining tasks. There are two BioBERT models: BioBERT-Base and BioBERT-Large. BioBERT-Large is based on BERT-Large and has twice as many layers as BERT-base.

To determine the relation between a chemical entity and a gene entity, we first locate the sentence where the entity pair is located. Next, we develop a representation specifically for that entity pair as multiple entity pairs can be located in the same sentence. We explore two different representations. Fig. 3. describes the two representations for the entity pair *benzamidine-plasminogen* (T1-T14) in (A). Representation B shows the input representation where the non-targeted entity pairs (genes *plasmin* and *gp330*) are removed from the input representation. Representation C shows the input representation where the entity pair is replaced with its semantic type: *benzamidine* and *plasminogen* are replaced with *@Chemical#* and *@Gene#*, respectively.

For our *Model-1* and *Model-2* (general BERT-based models), we explore using general BERT-cased embeddings into a simple feed-forward neural network. The key difference between the *Model-1* and *Model-2* is the input sentence representation. For *Model-1*, we remove the other entity pairs in the input sentence except for the targeted entity pair. For *Model-2,* to represent the entity pair in an input sentence, we use the semantic type of an entity to replace the entity itself. The modified input representation is passed through the pre-trained general BERT-cased model. The output is fed into a dropout layer and then a softmax layer for multi-class classification (6). When there is no relation between a chemical and gene/protein in a sentence, we treat it as an instance of a *'No-Relation'* class during the training. For our BioBERT-based model (*Model-3*), we explore using BioBERT embeddings into a feed-forward network for multi-class classification. Like *Model-2,* we represent an entity pair in a sentence by replacing the entities with the semantic types (Fig. 3.C). The maximum input sequence length for the *Model-3* is 128. We trim the sentence from both ends if a sentence is longer than the maximum sequence length. We perform this by taking the midpoint between the two entities and extending it by 64 tokens in both directions. We pass the input into the BioBERT-Large model, and embeddings of the [CLS] token are fed into a top model, consisting of a dropout and softmax layers.

V. Experimental Details

Here, we describe our experimental details.

Framework: Our code is open source and freely available at:

- https://github.com/synbioks/Text-Mining-NLP/tree/master/relation-extraction/biobert_RE/models/pt,
- https://github.com/NLPatVCU/BioCreative-VII-Track1

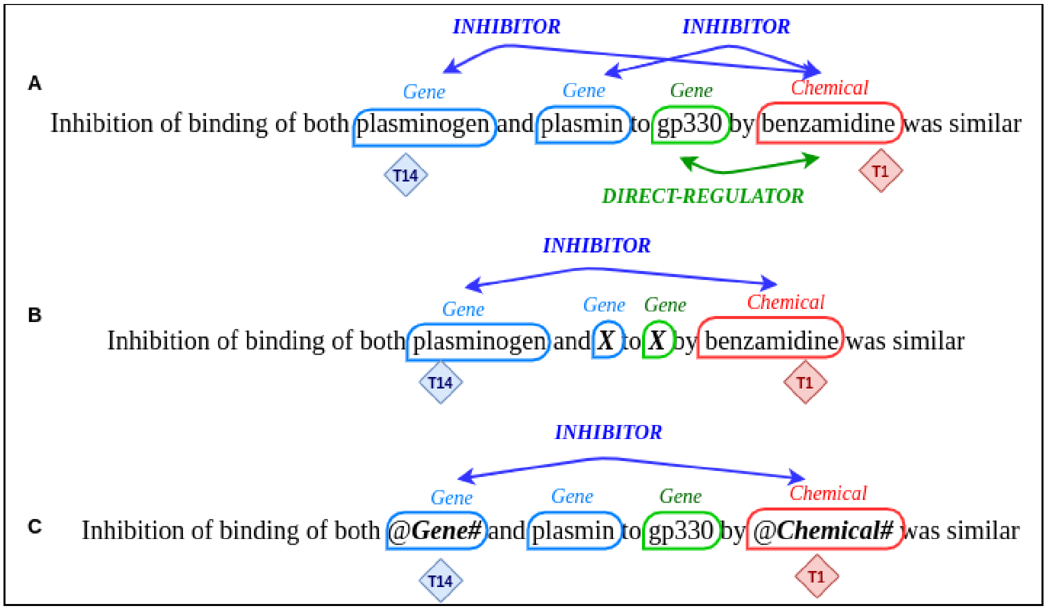

A
INHIBITOR
INHIBITOR
Gene
Gene
Gene
Chemical
Inhibition of binding of both plasminogen and plasmin to gp330 by benzamidine was similar
T14
T1
DIRECT-REGULATOR
B
INHIBITOR
Gene
Gene
Gene
Chemical
Inhibition of binding of both plasminogen and X to X by benzamidine was similar
T14
T1
C
INHIBITOR
Gene
Gene
Gene
Chemical
Inhibition of binding of both @Gene# and plasmin to gp330 by @Chemical# was similar
T14
T1

Fig. 3. Different representations of the input sentence used in our models. Model-1 utilizes the input representation B and the Models 2 & 3 utilize the input representation C

Tokenization: We used spaCy[2] and Scipy[3] to extract input sentences and the BERT and BioBERT tokenizers to convert the sentence into tokens.

Training parameters: We used a learning rate of 2e-5 (*Model-1&2)* and 3e-5 (*Model-3*) and a linear learning rate schedule with 1/10th of the total training steps as a warm-up. We used a batch size of 12 for the training in all models. We applied early stopping to the training for both BioBERT-based models (six epochs) and the BERT-based models (15 epochs).

Downsampling: We downsampled the class that denotes no relation between the entity pairs by 75% to overcome the heavy class imbalance during the training in the BERT-based models.

VI. Evaluation criteria

We evaluated our system using the DrugProt evaluation library provided by the organizers. Our approach was evaluated using micro-averaged Precision (P), Recall (R), and F_1 score (F). Precision calculates how many instances are predicted correctly out of all instances, and Recall calculates out of all the correct instances that should have been predicted how many instances are correctly predicted. F_1 score is the harmonic mean of Precision and Recall.

VII. Results and Discussion

Here, we discuss the results of our three models on the development and training sets. Table II shows the Precision, Recall, and F_1 scores for our three models on the development and test sets. The bold terms indicate the best F score of each class for development and test sets.

A. Results over the development set

We utilized the development set results to obtain the best set of weights for our model. The results show that *Model-3* (BioBERT-based model) outperformed the other two models (general BERT-based models) except for two classes. Also, we can see a decrease in performance when the number of class instances decreases, especially the three classes AGONIST-ACTIVATOR, AGONIST-INHIBITOR, and SUBSTRATE_PRODUCT-OF, which have the lowest number of instances. This is mainly because these classes do not have enough instances to be differentiated from other classes during training.

Compared to *Models 1 & 3*, *Model-2* could predict instances for the classes AGONIST-ACTIVATOR, AGONIST-INHIBITOR despite fewer training instances. We believe this is because we downsampled the *'No-Relation*' (entity pairs with no relation between the entities) due to the heavy class imbalance during training. Downsampling and the input representation of the *Model-2* improved the performance of the classes with few instances.

Overall performance of *Model-2* is higher than *Model-1*, but the Recall of *Model-1* is higher than *Model-2* for most classes. We assume this is due to the difference in the input representation of the models. Since *Model-1* eliminates the entities except for the targeted entities, the Recall is high. We experimented with both general BERT-cased and BERT-uncased, and we found that comparatively, BERT-cased performed better.

[2] https://spacy.io/
[3] https://www.scipy.org/

TABLE II. PRECISION (P), RECALL (R), AND F_1 SCORE (F) RESULTS FOR ALL MODELS OVER THE DEVELOPMENT AND TEST DATA.

	Development set									**Test set**								
	Model 1			***Model 2***			***Model 3***			***Model 1***			***Model 2***			***Model 3***		
	P	***R***	***F***	***P***	***R***	***F***	***P***	***R***	***F***	***P***	***R***	***F***	***P***	***R***	***F***	***P***	***R***	***F***
INDIRECT-DOWN REGULATOR	0.48	0.69	0.57	0.62	0.67	0.64	0.75	0.73	**0.74**	0.44	0.57	0.50	0.50	0.72	0.59	0.67	0.72	**0.70**
INDIRECT-UPREG ULATOR	0.33	0.64	0.44	0.58	0.66	0.62	0.76	0.78	**0.77**	0.34	0.63	0.44	0.49	0.68	0.57	0.68	0.75	**0.71**
DIRECT-REGULAT OR	0.35	0.67	0.46	0.48	0.63	0.54	0.72	0.52	**0.61**	0.34	0.55	0.42	0.41	0.6	0.48	0.70	0.57	**0.63**
ACTIVATOR	0.32	0.61	0.42	0.53	0.63	0.58	0.78	0.75	**0.77**	0.47	0.61	0.53	0.56	0.74	0.63	0.79	0.70	**0.74**
INHIBITOR	0.50	0.82	0.62	0.65	0.83	0.73	0.86	0.83	**0.85**	0.53	0.75	0.62	0.61	0.78	0.69	0.81	0.79	**0.80**
AGONIST	0.43	0.63	0.51	0.67	0.68	0.67	0.74	0.75	**0.74**	0.49	0.67	0.57	0.58	0.63	0.61	0.73	0.65	**0.69**
AGONIST-ACTIVA TOR	0.0	0.0	0.0	0.75	0.3	**0.43**	0.0	0.0	0.0	0.0	0..0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AGONIST-INHIBIT OR	0.0	0.0	0.0	0.25	0.5	**0.33**	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.33	**0.50**	0.0	0.0	0.0
ANTAGONIST	0.43	0.76	0.55	0.68	0.76	0.72	0.91	0.90	**0.90**	0.54	0.80	0.64	0.65	0.88	0.74	0.86	0.85	**0.86**
PRODUCT-OF	0.25	0.47	0.33	0.38	0.53	0.44	0.61	0.58	**0.60**	0.33	0.43	0.38	0.42	0.63	0.50	0.61	0.59	**0.60**
SUBSTRATE	0.31	0.69	0.43	0.44	0.69	0.54	0.72	0.76	**0.74**	0.42	0.44	0.43	0.38	0.53	0.44	0.61	0.55	**0.58**
SUBSTRATE_PRO DUCT-OF	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0..0	0.0	0.0	0..0	0.0	0.0	0.0	0.0
PART-OF	0.31	0.45	0.37	0.46	0.41	0.44	0.76	0.68	**0.72**	0.40	0.38	0.39	0.39	0.49	0.43	0.71	0.61	**0.66**
	0.39	0.69	0.50	0.56	0.69	0.62	0.78	0.74	**0.76**	0.33	0.45	0.38	0.46	0.54	0.48	0.55	0.52	**0.54**

Therefore, we assume the difference in the casing of the words in the dataset played a role in determining the context of the words. Also, we experimented with BioBERT-Base and BioBERT-Large and found that BioBERT-Large provided a performance improvement of 1.6%. Again, we assume this is because BioBERT-Large is based on BERT-Large, which has twice as many layers as BERT-base and is trained over a more extensive biomedical-based vocabulary.

B. *Results over the test set*

The observations from the results of the test set are similar to the development set. Overall, *Model-3 (*BioBERT-based model) outperformed the other two models except for one class. However, the overall results of the test set are lower compared to the development set. Here, also we can see a decrease in performance when the number of class instances decreases. However, *Model-2* could predict all the positive instances correctly (Precision-1.0) for the class AGONIST-INHIBITOR.

From the results of both the development and test sets, *Model-2* performed better than *Model-1*. Therefore, it is safe to assume that replacing the entities with their semantic types is an efficient way of representation than training with the actual entity tokens. Furthermore, since the BioBERT is pre-trained on biomedical articles, it gives more efficient contextualized embeddings than the BERT trained on general English. We believe this is why *Model-3 (*BioBERT-based model) outperforms the other two models (general BERT-based models).

VIII. CONCLUSION

This paper presented three contextualized language-based models, a BioBERT-based and two general BERT-based models, to automatically detect relations between chemical compounds/drugs and genes/proteins. We evaluated our models on the DrugProt dataset and found that the BioBERT-based model outperformed the other models. From the results of both the development and the test set, we can conclude that BioBERT embeddings represent the tokens effectively when used on biomedical data. Also, replacing the entities with their semantic types is an effective unique representation of the input sentence.

Here, we use a simple neural network on the output of the contextualized embeddings. In the future, we plan to explore more complex deep neural networks with contextualized embeddings, for example, Graph Convolutional Networks (GCNs) (9). Traditional neural networks perform well on euclidean data; however, they do not handle non-euclidean data representations within language well because the model considers the positional information of the words. Therefore, utilizing GCN with contextualized embeddings provides the flexibility of language when expressing relationships between entities. Also, we plan to explore Joint Learning for RE in the future. Named entities are essential to extract relations, and named entity recognition (NER) helps identify the entities in the text (10). Therefore, simultaneously learning NER and RE can be beneficial to capture such two different types of information in the learning process.

Acknowledgment

This work was funded in part by NSF award numbers 1939885 and 1939951, and utilized computer resources provided in part by NSF award numbers 1730158, 2100237, 2120019.

References

1. Sousa, D., Lamurias, A., & Couto, F. M. (2021). Using neural networks for relation extraction from biomedical literature. In Artificial Neural Networks, pp. 289-305, Humana, New York, NY.
2. Krallinger, M., Rabal, O., Akhondi, S. A., Pérez, M. P., Santamaría, J., Rodríguez, G. P., ... & Intxaurrondo, A. (2017, October). Overview of the BioCreative VI chemical-protein interaction Track. In Proceedings of the sixth BioCreative challenge evaluation workshop, Vol. 1, pp. 141-146.
3. 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.
4. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
5. 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.
6. Mahendran, D., & McInnes, B. T. (2021). Extracting Adverse Drug Events from Clinical Notes. arXiv preprint arXiv:2104.10791.
7. Wei, Q., Ji, Z., Si, Y., Du, J., Wang, J., Tiryaki, F., ... & Xu, H. (2019). Relation extraction from clinical narratives using pre-trained language models. In AMIA Annual Symposium Proceedings, Vol. 2019, p. 1236. American Medical Informatics Association.
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems, pp. 5998-6008.
9. Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.
10. Qiuyan, X., & Fang, L. (1978). Joint Learning of Named Entity Recognition and Relation Extraction. In 2011 International Conference on Computer Science and Network Technology, Vol. 1982).
11. Peng, Y., Rios, A., Kavuluru, R., & Lu, Z. (2018). Chemical-protein relation extraction with ensembles of SVM, CNN, and RNN models. arXiv preprint arXiv:1802.01255.
12. Antunes, R., & Matos, S. (2019). Extraction of chemical–protein interactions from the literature using neural networks and narrow instance representation. Database, 2019.
13. Yüksel, A., Öztürk, H., Ozkirimli, E., & Özgür, A. (2017, October). CNN-based chemical–protein interactions classification. In Proceedings of the BioCreative VI Workshop (pp. 184-186).
14. Sun, C., Yang, Z., Luo, L., Wang, L., Zhang, Y., Lin, H., & Wang, J. (2019). A deep learning approach with deep contextualized word representations for chemical–protein interaction extraction from biomedical literature. IEEE Access, 7, 151034-151046.
15. Sun, C., Yang, Z., Wang, L., Zhang, Y., Lin, H., Wang, J., ... & Zhang, Y. (2018, December). Hierarchical Recurrent Convolutional Neural Network for Chemical-protein Relation Extraction from Biomedical Literature. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 765-766). IEEE.
16. Liu, S., Shen, F., Wang, Y., Rastegar-Mojarad, M., Elayavilli, R. K., Chaudhary, V., & Liu, H. (2017). Attention-based neural networks for chemical protein relation extraction. Training, 1020(25.247), 4157.
17. Hafiane, W., Legrand, J., Toussaint, Y., & Coulet, A. (2020). Experiments on transfer learning architectures for biomedical relation extraction. arXiv preprint arXiv:2011.12380.