Using Knowledge-Based Pretrained Language Model for Mining Drug and Chemical-Protein Interactions

Qingyao Li[1], Ying Xiong[1], Jingyue Hu[1], Buzhou Tang[1,2*]
[1]Department of Computer Science, Harbin Institute of Technology (Shenzhen), Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China

***Abstract*—In this paper, we describe our systems for the DrugProt task of BioCreative VII. This task is to automatically detect in relations between chemical compounds/drug and genes/proteins. First, we use KeBioLM pretrained language model as text encoders and replace the cross-entropy function with focal loss to alleviate the imbalance in relation samples. Then we run five times with different seeds to obtain our ensemble model. Experimental results on the test set demonstrate our ensemble model achieves the F1-score of 0.7419, which outperforms the mean results of this track by 0.1222.**

Keywords—biomedical relation extraction, multitask learning, fine-grained markers, ensemble learning

I. Introduction

The BioCreative VII launches DrugProt track[1] on automatic detection of drug/chemical interactions with genes, proteins and miRNAs, which is similar to ChemProt track of the BioCreative VI(1). These tasks are actually relation extraction (RE) task. Relation extraction is an important process to construct knowledge graph and aims to extract the semantic relation given entities. Traditional relation extraction includes rule-based methods (2) and feature-based engineering methods (3). Many researchers have recently proposed deep learning methods. Zeng et al. (4) first introduce entity position information into relation extraction. Multi-Level Attention CNNs (5) is proposed to use the attention in the input and used pooling layers to capture key information. Sorokin et al. (6) propose a contextual aware approach as other relations in the same sentence affect the judgment of given entity pair. The superiority of pre-trained language model has brought subversive changes to the improvement of the field of natural language processing. The output of BERT (7) is directly used to represent the word embedding, which can be fine-tuned or fixed according to the specific tasks. The BERT model has variants in the biomedical domain, such as BioBERT (8), SciBERT (9), BlueBERT (10), and PubMedBERT (11), which are trained based on different pre-training data. PubMedBERT proposes a new paradigm for domain-specific pre-training, using PubMed summaries to start training from scratch. KeBioLM(12) explicitly uses knowledge in UMLS[2] and absorbs more biomedical information, outperforming other language models on named entity recognition and relation extraction of BLURB benchmark.

In this paper, we employ BioBERT or KeBioLM as model encoder and define the input and output of model encoder. We propose some strategies to enhance the model, such as multitask learning and relation attention. To alleviate the imbalance of different relations, we apply focal loss(13). Since the DrugProt track provides fine-grained gene entities, we proposed a simple and effective way to replace coarse-grained entity markers. This is an alternative approach to multitask learning, releasing from the tedious adjustment of hyperparameters. We run five times with different seeds and vote them as our ensemble model, which achieves precision, recall and F1-score of 0.7671, 0.7183, 0.7419. Our ensemble model improves about 12.41% (precision), 8.92% (recall), 12.22% (F1-score) compared with the mean results of this track.

II. Analysis of the Dataset

A. Preliminary Statistics

We conduct preliminary statistics on the dataset of DrugProt track(14). TABLE I. presents the number of 13 types of interactions in the dataset. Surprisingly, we found an imbalance in the proportion of category instances. In the training set, the interactions with the largest number of instances have 5,392 instances, while the interactions with the least instances have only13 instances. In addition, we also counted the interactions between CHEMICAL and GENE-Y/N. Note that GENE-Y and GENE-N are unified as GENE in the development set and test set. This detail will be applied to our model in the next section.

TABLE I. ENTITY TYPE PAIR ON THE TRAINING SET

Relations	**Entity Pair**	
	CHEMICAL-GENE-Y	**CHEMICAL-GENE-N**
PRODUCT-OF	677	244
ANTAGONIST	687	285
SUBSTRATE	1370	633
ACTIVATOR	788	641
INHIBITOR	3423	1969
INDIRECT-DOWNREGULATOR	1048	282
INDIRECT-UPREGULATOR	1052	327
AGONIST	495	164
PART-OF	617	269
DIRECT-REGULATOR	1583	667
AGONIST-ACTIVATOR	28	1
AGONIST-INHIBITOR	6	7
SUBSTRATE_PRODUCT-OF	21	4

[1] https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-1/
[2] https://www.nlm.nih.gov/research/umls/index.html

B. *Construction of Negative Samples*

In general, many constructions of negative sample methods combine two entities with no interacting facts. However, too many negative samples will affect the distribution of samples. To make the best use of the negative samples, we revise the interval across the number of GENE entities for CHEMICAL entities. The interval represents the number of GENE crossed from the current CHEMICAL to GENE. Our experiments show that the interval [-10,16] performs best on the development set.

III. Methods

In this section, we introduce our proposed methods in DrugProt task BioCreative VII. Meanwhile, we present the details of our proposed fine-grained entity markers replacement approach (FGEMR). Compared with multitask learning, FGEMR does not require tedious work on adjustment of hyperparameters.

We employ BIOBERT and KEBIOLM as the encoder to obtain contextualized embeddings of instances. The relation statement $r = (s, e1, e2)$ contains the sequence of tokens s and the entity span identifiers e1 and e2. Similar to PURE (15), we define the input encoding and the output relation representation. We introduce $[S{:}\, e_1^{type}]$, $[/S{:}\, e_1^{type}]$, $[\ O{:}\, e_2^{type}]$ and $[/O{:}\, e_2^{type}]$ and insert them on both sides of input entities. Accordingly, our relation representation is the concatenation of two output representations that correspond $[S{:}\, e_1^{type}]$ and $[\ O{:}\, e_2^{type}]$. The representational learning framework is illustrated in Fig. 1.

A. *System1: Relation attention with BioBERT*

Relation attention mechanism focuses on the relation that is more relevant to a given sentence. In Pre-Processing phase, we obtain the relation definition and process as tokens sequences. For each relation, we put its definition into the encoder and take the embedding corresponding to [CLS][1] as the representation of the relation. The attention mechanism is adopted to get the relation-enhanced representation, this process can be written as:

$$s_j^r = \sum_{j \in Data} a_j^T \cdot R, \tag{1}$$

$$a_j = softmax(R \cdot h_j^T), \tag{2}$$

where s_j^r is the relation representation based on relation attention of the j-th instance in dataset. R is the matrix of the embeddings of the relations. h_j^T is the relation representation of the j-th instance.

B. *System2: Multitask learning with BioBERT*

Multitask Learning proposed by Collobert et al. (16) aims to reduce the risk of over-fitting caused by noise through parameter sharing. Besides, the auxiliary tasks help the model focus on those features that are more important. We first choose entity classification as an auxiliary task because of its high correlation with relation extraction task, which promotes feature interaction between the two tasks. Concretely, relation extraction and entity classification shared the same encoder. Then we use the start maker as the entity representation and modify the objective function of our model,

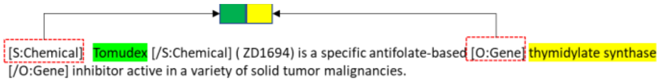
[S:Chemical] Tomudex [/S:Chemical] (ZD1694) is a specific antifolate-based [O:Gene] thymidylate synthase [/O:Gene] inhibitor active in a variety of solid tumor malignancies.

Fig. 1. The representational learning framework of relation.

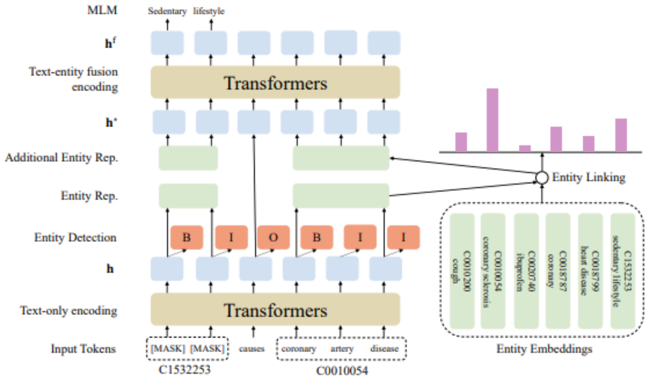

Fig. 2. The framework of KeBioLM(12).

$$\mathcal{L}_{all} = \lambda\mathcal{L}_{FL} + (1-\lambda)\mathcal{L}_{en}, \tag{3}$$

where λ is a hyperparameter to balance the two terms. $\mathcal{L}_{FL}$ is the focal loss function, and $\mathcal{L}_{en}$ is the loss function of the entity type classification task.

C. *System 3: Fine-Grained Entity Markers Replacement*

In section II, we learn that GENE-Y and GENE-N are unified as GENE in the development and test sets. GENE is a coarse-grained entity type, while GENE-Y and GENE-N are fine-grained entity types. We need to be consistent in training and testing whether we use coarse-grained entity types or fine-grained entity types.

Intuitively, fine-grained entity types are better at capturing the details of relation, so we use fine-grained entity types as the input marker. To obtain fine-grained entity types, we add an independent encoder for classifying GENE-Y or GENE-N. For entity type encoder, the input is a sequence of tokens with coarse-grained entity type markers. After that, we can easily acquire the fine-grained classification for GENE.

Finally, we replace all coarse-grained entity types with fine-grained entity types, and the replaced sequence is the input of relation extraction encoder. It is worth noting that the objective function of entity type encoder is cross-entropy function since there has no obvious imbalance between the instances of GENE-Y and GENE-N.

[1] The sentence representation of BERT.

TABLE II. RESULT OF MODELS ON BIOCREATIVE VII TRACK 1 DEVELOPMENT TEST S

Encoder	Model	Development set			Test set		
		P	R	F1	P	R	F1
BioBERT	Baseline	0.752	0.77	0.761	NA	NA	NA
	Relation attention	0.757	0.769	0.763	NA	NA	NA
	Mutitask learning	0.773	0.76	0.766	0.7468	0.7065	0.7261
	FGEMR	0.766	0.766	0.766	0.7475	0.70	0.7229
	Ensemble	**0.81**	0.757	0.779	**0.7782**	0.6936	0.7335
KeBioLM	Baseline	0.781	0.756	0.768	NA	NA	NA
	Ensemble	0.792	**0.776**	**0.784**	0.7671	**0.7183**	**0.7419**

D. System 4: Ensemble learning with KeBioLM

KeBioLM extracts entities from PubMed abstracts and linked with UMLS. It applies the plain text coding layer to learn entity representation and the text-entity fusion coding to aggregate entity representation, and adds the loss of name entity detection and entity linking. Finally, we run five times on KeBioLM with different seeds to ensemble our systems. The framework of KebioLM is shown in Fig. 2.

IV. EXPERIMENTS

A. Implementation Details

We evaluate our model on DrugProt dataset(14), and take the BioBERTv1.1[1] and KeBioLM[2] as the encoder, where the maximum length is 256, and the dimension of embedding is 768. The model applies AdamW optimizer (17) to perform gradient descent, trains for 10 epochs, and evaluate every 0.5 epoch. The learning rate is set to 2e-5. The best checkpoint on the development will be saved and used for the testing phase

In order to solve the problem of serious imbalance of positive and negative samples, focal loss (13) reduces the weight of a large number of simple negative samples in training. Through analysis of the dataset, there is a large gap in the number of category instances. We use focal loss instead of cross-entropy to alleviate the phenomenon of sample imbalance. The focal loss is defined as follows:

$$\mathcal{L}_{FL} = -(1-p_r)^{\gamma}\log(p_r), \qquad (4)$$

where γ is a hyperparameter to adjust the weight between simple samples and hard samples. p_r is the probability distribution for relations.

Model ensemble is to improve the generalization ability of models by fusing multiple models. The relation predictions use hard voting methods on five models, and our model further achieves better performance.

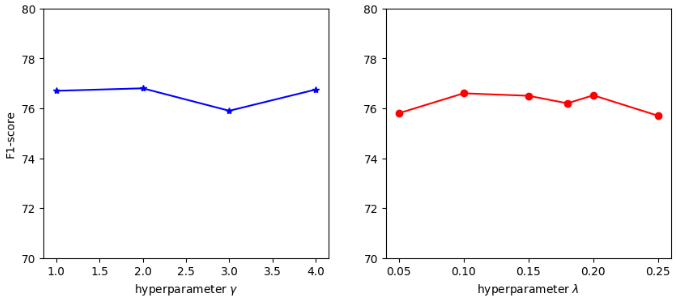

Fig. 3. The analysis of hyperparameters.

B. Results

TABLE II. shows the experimental results on the development set and test set. The first score is the result of the development set, and the second score represents the result on the test set. Our baseline introduces focal loss instead of cross-entropy. FGEMR is the method that uses fine-grained entity markers replacement.

Our ensemble model achieves F1-score 0.7419 on the test set and improves about 12.22% compared with the mean results of this track. The relation attention performs slightly better than the baseline and indicates that the attention mechanism can capture informative and subtle features which relate to gold relation. The F1-score of multitask learning improves 0.5% compared with the baseline on the development set as an auxiliary task with the entity type classification to learn more complicated features. Similarly, FGEMR leveraging fine-grained entity types gains competitive performance. Multitask learning is greatly influenced by hyperparameters, while FGEMR saves time and is free from manual tuning for hyperparameters. We believe that if it could take advantage of more fine-grained entity types, the performance of FGEMR would be better.

We can see that BioBERT performs worse than KeBioLM for our baseline because KeBioLM explicitly uses knowledge in UMLS and absorbs more biomedical information. It implies the effectiveness of using knowledge graph during the training phase, especially for the biomedical domain.

[1] https://huggingface.co/monologg/biobert_v1.1_pubmed
[2] https://github.com/GanjinZero/KeBioLM

C. *Hyper-Parameter Analysis*

Fig. 3 shows the F1-score among different γ and λ values, respectively. The hyperparameter γ adjusts the weight of simple samples. The focal loss degenerates to cross-entropy loss when γ is set to 0, and the experiment illustrates 2 is an optimal solution. λ is a trade-off between relation extraction and entity type classification, and the model achieves the best performance when λ is set to 0.1.

V. Conclusion

In this paper, we have attempted some meaningful experiments for the DrugProt task of BioCreative VII. We apply KeBioLM pretrained language model as text encoders and use focal loss instead of cross-entropy loss to alleviate the effect of imbalance classes. Using model ensemble further improves the performance. Experimental results on the test set demonstrate our ensemble model achieves the F1-score 0.7419, which outperforms the mean results of this track by a large margin of 12.22%.

Acknowledgment

This paper is supported in part by grants: National Natural Science Foundations of China (U1813215 and 61876052), National Natural Science Foundations of Guangdong, China (2019A1515011158), Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20180306172232154 and 20200821174109001).

References

1. 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).
2. Sohn, S., Wu, S. & Chute, C. G. Dependency parser-based negation detection in clinical narratives. *AMIA Summits on Translational Science Proceedings* **2012**, 1 (2012).
3. Kambhatla, N. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. in *Proceedings of the ACL Interactive Poster and Demonstration Sessions* 178–181 (2004).
4. Zeng, D., Liu, K., Lai, S., Zhou, G. & Zhao, J. Relation classification via convolutional deep neural network. in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* 2335–2344 (2014).
5. Wang, L., Cao, Z., De Melo, G. & Liu, Z. Relation classification via multi-level attention cnns. in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 1298–1307 (2016).
6. Sorokin, D. & Gurevych, I. Context-aware representations for knowledge base relation extraction. in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* 1784–1789 (2017).
7. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* 4171–4186 (Association for Computational Linguistics, 2019). doi:10.18653/v1/N19-1423.
8. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
9. Beltagy, I., Lo, K. & Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* 3615–3620 (2019).
10. Peng, Y., Chen, Q. & Lu, Z. An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining. in *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing* 205–214 (2020).
11. Gu, Y. *et al.* Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779* (2020).
12. Yuan, Z., Liu, Y., Tan, C., Huang, S. & Huang, F. Improving Biomedical Pretrained Language Models with Knowledge. in *Proceedings of the 20th Workshop on Biomedical Language Processing* 180–190 (Association for Computational Linguistics, 2021). doi:10.18653/v1/2021.bionlp-1.20.
13. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. in *Proceedings of the IEEE international conference on computer vision* 2980–2988 (2017).
14. 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.
15. Zhong, Z. & Chen, D. A Frustratingly Easy Approach for Entity and Relation Extraction. in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 50–61 (2021).
16. Collobert, R. & Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. in *Proceedings of the 25th international conference on Machine learning* 160–167 (2008).
17. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. in *International Conference on Learning Representations* (2018).