Biomedical Relation Extraction Based on Deep Transfer Learning and Prompt Learning

Rongrong Ma[*] MRR@STU.CQUT.EDU.CN
School of Computer Science and Engineering, Chongqing University of Technology

Huan Liu HUANLIU_UESTC@163.COM
Chongqing Institute of Artificial Intelligence, Shanghai Jiao Tong University

Editors: Nianyin Zeng and Ram Bilas Pachori

Abstract

Biomedical Relationship Extraction (BioRE) is an important task for electronic health record mining and biomedical knowledge base construction. Due to the complexity of biomedical information in the literature, it is difficult to realize the construction of large-scale datasets. Although previous models showed good results in a fully supervised environment, they did not generalize well to the low resource situation common in this field. We propose a biomedical relationship extraction model based on deep transfer learning and prompt learning. We use deep neural network for transfer learning, using large-scale biomedical field data set (BioREx) training model. The trained model was applied to low-resource datasets in biomedical field, which can make the model to learn more rich knowledge, so as to alleviate the problem of insufficient training data. In addition, we used a prompt learning method by appending a prompt sentence describing the desired relationship label to the beginning of the input sentence, which can reduce the gap between the pre-trained language model (PLM) and downstream tasks. Experiments show that using deep transfer learning and prompt learning can effectively improve the prediction results. Experiments on three BioRE benchmark datasets DrugProt, DDI and BC5CDR, F1 values are significantly improved compared to previous models.

Keywords: Biomedical relationship extraction, Deep transfer learning, Prompt learning, Pre-trained language models, Biomedical datasets

1. Introduction

Using text-mining techniques, valuable information can be extracted from the massive biomedical literature and provided to biomedical experts for further research. As the core task of biomedical information extraction, biomedical named entity identification and biomedical entity relationship extraction have been attracting the attention of researchers. After identifying biomedical entities (genes, drugs, diseases, etc.) from a vast amount of biomedical literature, the interaction relationships between these entities can be further explored. Identifying the interaction relationship between entities is the process of judging whether two entities appear in a piece of text. Biomedical relationship extraction can not only help to build knowledge maps within the field, but also promote new drug development and precision medicine research. The relationship between biomedical entities is rich in information. It is of great significance for researchers to excavate the relationship between biomedical entities from the literature to carry out systematic biomedical research, which can effectively promote the development of the field of life science.

In recent years, the development of biomedical entity relationship extraction methods has made remarkable progress. However, there are still obstacles in training a reliable biomedical RE model.

Biomedical RE often suffers from inadequate and imperfect annotation problems, because the annotation process is very challenging. The inadequacy and imperfection of annotation inevitably lead to existing state-of-the-art (SOTA) biomedical regeneration systems that, despite showing satisfactory results in fully supervised settings, lead to poor generalization of the low-resource environments more common in the field. And the model is difficult to learn a clear decision boundary, Incorrect prediction models may not be abandoned, especially in high-risk areas such as biomedicine, incorrect predictions may have serious direct consequences for patients.

For the above problems, this study proposed a biomedical relationship extraction model based on deep transfer learning and prompt learning, which performed well even in low-resource situations. In recent years, most relationship extraction based on model migration has been combined with a deep neural network, trained by adding a domain adaptation layer to the neural network, and then trained by joint feature-based migration. We use the deep neural network for transfer learning, and use the large-scale biomedical field merged data set (BioREx) (Lai et al., 2023) as the training data to train the model, so that the model can learn more abundant domain knowledge, which can effectively alleviate the problem of insufficient training data. Experiments are performed on low-resource data sets common in this field, and the results show that transfer learning can effectively improve the experimental results. Furthermore, we propose a prompt learning approach that preceding input sentences that require predicted relationship labels, shifting cue information from large PLM to downstream tasks, reducing the gap between PLM and downstream tasks, which significantly outperforms baseline in low-resource situations. Experiments show that using deep transfer learning and prompt learning can effectively improve the prediction results. There are three main contributions to this work:

- A deep transfer learning method is proposed for training models using large-scale biomedical domain merged datasets, enabling the model to learn more abundant domain knowledge with good generalization ability in low resource environment;
- Introduce a prompt learning method, adding prompt sentences in front of input sentences, transferring prompt information from large PLM to downstream tasks, reducing the gap between PLM and downstream tasks, especially in the case of low resources and annotation is scarce;
- We propose a novel ranking-based loss that penalizes overconfident unexpected instances while encouraging expected instances, to implicit abstention calibration that handles abstinent relations in the dataset to learn a fine-grained, instance-aware decision boundary.

Through extensive experiments on three commonly used biomedical datasets, DrugProt (Miranda-Escalada et al., 2021), DDI (Herrero-Zazo et al., 2013), and BC5CDR, we verified the advantages of the model, showing that deep transfer learning and prompt learning can alleviate the problem of corpus insufficiency and have good generalization ability in low-resource environments.

2. Method

Here, we introduce the model presented in this paper. We formulate the problem in 2.1, introduce to large-scale merged datasets in 2.2, illustrate the fine-tuning approach based on prompt learning in 2.3, and describe the overall framework of the model in 2.4.

2.1. Problem Definition

The relationship extraction model takes one sentence x and two entities (e_1 and e_2) as input to predict the relationships between e_1 and e_2 from a label space Y containing all possible relationships y. The label space Y includes the expected instance ($y = \perp$) and the unexpected instance ($y = \{Y - \perp\}$). We treat RE as a multi-classification task designed to classify instances as a predefined RE relationship type or no relationship ("None"). A successful relationship extraction model should avoid unintended instances and accurately predict the expected instances. We will treat it as a separate relational label that represents the relational label of the expected instance. X represents the sample space, which is the input data that needs to be classified.

In this paper, we use the deep transfer learning method and build three training sets: Source domain training set D_{source}, target domain training set D_{target} and training set D_{train}, which combines the source domain training set D_{source} and the target domain training set D_{target}.

$$D_{source} = \{x_{Si}, y_{Si}\}_{i=1}^{n}, D_{target} = \{x_{Ti}, y_{Ti}\}_{i=1}^{m}, D_{test} = \{x_{Ti}\}_{i=1}^{k} \tag{1}$$

$$D_{train} = \{D_{source}, D_{target}\} \tag{2}$$

The goal of transfer learning is to help train with the source data set D_{source}, and to obtain the observed value h_T on the target data set D_{target}, so that the observed value is as close to the target value as possible.

2.2. Datasets Construction

In this paper, we prove that the pre-trained model built on the combined dataset is robust and generalizable. However, most relationship extraction datasets contain only one relationship type, and these benchmark datasets are collected from different ranges of annotations. Most systems are developed and validated on a single dataset, which severely limits the development of generalized biomedical RE systems. To this end, we cite a new RE dataset, BioREx, which includes multiple relationships (e.g., Gene-Disease) among some of the most important biomedical concepts (e.g., Genes, Disease and Chemistry). As shown in Figure 1, BioREx Datasets coordinate the annotation differences between the data sources and integrate them to build a rich dataset with sufficient quantity. It systematically addresses the data heterogeneity of a single dataset and combines them into a large dataset, achieving higher performance compared to models trained on a single dataset. Using these methods, deep transfer learning aggregates the external data, eliminating the sparsity problem for small tasks.

In total, eight relevant datasets were selected for the BioREx dataset, including corpora and repositories. A text corpus is a collection of text with manually managed entities and the relationships between them. In contrast, the repository contains a list of document-level relationships. Document-level dataset, such as BC5CDR, and sentence level dataset, such as DDI, provide different objectives and provide different advantages. Sentence-level dataset provide precise context, including the location of entity pairs and supporting evidence. Document-level dataset do not provide explicit relationship sentences between pairs, but rather provide rich contextual information. By merging datasets from document-level and sentence-level datasets, the BioREx dataset contains different types of contexts, making it more versatile.To integrate this datasets into one dataset, the BioREx dataset adjusts each dataset according to its own annotated features.

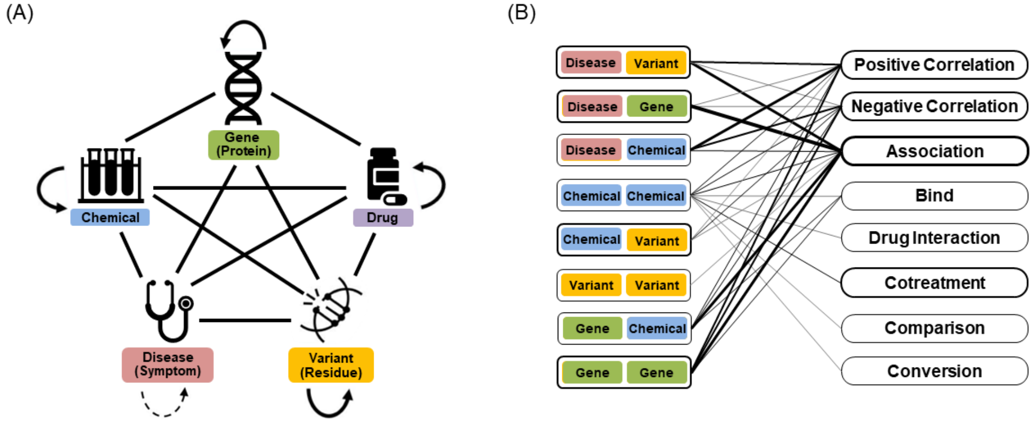


Figure 1: Profile in the BioREx dataset.(A) Relationships between the entity types.(B) Mapping between entity pairs and relationship types. The thickness of lines between entitys and pair-relationship types indicate the frequency of occurrence.

2.3. Fine-Tuning of Prompt Learning

For each sentence, two boundary labels are inserted at the beginning and end of the entity requiring a predicted relationship (e. g. <D> and </ D> at the beginning and end of the disease entity). We also added these markers to the PLM vocabulary to ensure that these markers were not segregated into multiple markers.

We further constructed a prompt question, providing contextual guidance for the model to perform the relation extraction task. The prompt question is attached at the beginning of the input sentence to emphasize the two entities in a pair (e_1 and e_2) and the RE task. The input instance of the model consists of the following

$$Inst = [CLS]\ Prompt\ [SEP]\ Sent. \tag{3}$$

Specifically, [SEP] is a special marker for splitting the prompt sentence and the original sentence.

$$\begin{aligned} Sent = w_1, \ldots, w_{a-2}, [Entity^1]_{a-1}, e^1_{a:b}, [/Entity^1]_{b+1}, w_{b+2}, \ldots, \\ w_{c-2}, [Entity^2]_{c-1}, e^2_{c:d}, [/Entity^2]_{d+1}, w_{d+2}, \ldots, w_t. \end{aligned} \tag{4}$$

where [Entity] and [/ Entity] represent the beginning and end of the entity token (e_1 and e_2), respectively, and the subscript indicates the position of the word.

Prompt is a cued template with boundary labels inserted at the start and end of two entities (e_1 and e_2), with [Corpus] indicating the possible relationship type between e_1 and e_2. [Corpus] is a task-specific name for various tasks, such as “DDI”, “BC5CDR”, etc.

$$Prompt = What\ is\ [Corpus]\ between\ [Entity^1]e^1[/Entity^1]\ and\ [Entity^2]e^2[/Entity^2]? \tag{5}$$

Prompt is a description of the relationship labels that need to be predicted. We introduced the symbol “?” in the prompt sentence, the sequence intended to indicate the prompt sentence is a question and the prediction of [Corpus] will be treated as an answer. Thus, the entire process of prompt tuning simulates a masking language modeling task to reduce the gap between the PLM and the downstream tasks.

2.4. Overview of Model

In this paper, we use transfer learning (TL) to pre-train the model on the large datasets and fine-tune the model in the downstream tasks. Thus, downstream tasks benefit from a pre-trained language model (PLM) that learns domain-specific neural networks and can focus on tuning neural parameters for a specific task. We prove that the pre-trained language models based on the merged dataset are robust and generalizable.

In Figure 2, we show the concrete workflow of processing a test input. We chose BioLink-BERT as our default PLM. The prediction of the relationship type is based on the class with the highest probability. The confidence score of the classification output depends on the state of the PLM's [CLS] marker.

To obtain a confidence score for each class, a fixed-length representation of the [CLS] marker was entered into a linear layer, followed by a softmax activation. A score of each relationship type y_i is provided for the model's output. The type with the highest score becomes the the predicted output.

$$R(x, e1, e2) = arg \max_{y \in Y} s(y_i) \tag{6}$$

In experiments, we observe that the model is easily misled by a large number of unintended instances in the dataset, leading to a performance deterioration. To mitigate this unintended imbalance problem, we introduce We propose a novel ranking-based loss that penalizes overconfident unexpected instances while encouraging expected instances. Specifically, if the relationship is unexpected instances, we calibrate the scores to are suppressed; otherwise, we control the ranking over the other relationships.

$$L_{AC} = \textstyle\sum_{(x,y) \in D} l_{AC}(x, y) \tag{7}$$

$$l_{AC} = \begin{cases} \sum_{i=1}^{n} l_{rank}(s(y), s(y_i); \gamma), if\ y = \perp \\ l_{rank}(s(y), s(\perp); \gamma), otherwise \end{cases} \tag{8}$$

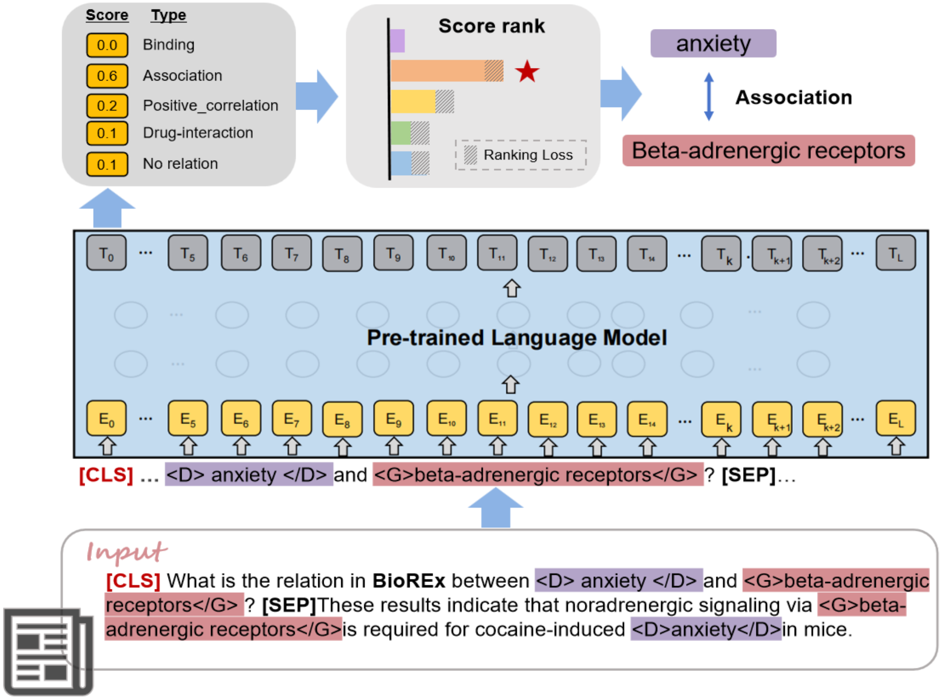
where, γ is a non-negative constant where the ranking loss $l_{rank}(x_1, x_2; \gamma)$ learning predicts x_1 higher than x_2 for γ. By training to achieve this goal, the present model can be seen as combining an implicit waiver calibrator, as a learnable instance-aware threshold.

3. Experiment

In this section, we discuss our experimental setup (3.1) and experimental outcome evaluation (3.2).

3.1. Experimental Setting

In this paper, we use transfer learning (TL) to pre-train the model on the large datasets and fine-tune the model in the downstream tasks. Thus, downstream tasks benefit from a pre-trained language model (PLM) that learns domain-specific neural networks and can focus on tuning neural parameters for a specific task. We prove that the pre-trained language models based on the merged dataset are robust and generalizable.


Score Type
0.0 Binding
0.6 Association
0.2 Positive_correlation
0.1 Drug-interaction
0.1 No relation
Score rank
Ranking Loss
anxiety
Association
Beta-adrenergic receptors
T0 ... T5 T6 T7 T8 T9 T10 T11 T12 T13 T14 ... Tk Tk+1 Tk+2 ... TL
Pre-trained Language Model
E0 ... E5 E6 E7 E8 E9 E10 E11 E12 E13 E14 ... Ek Ek+1 Ek+2 ... EL
[CLS] ... <D> anxiety </D> and <G>beta-adrenergic receptors</G> ? [SEP]...
Input

[CLS] What is the relation in **BioREx** between <D> anxiety </D> and <G>beta-adrenergic receptors</G> ? **[SEP]**These results indicate that noradrenergic signaling via <G>beta-adrenergic receptors</G>is required for cocaine-induced <D>anxiety</D>in mice.

Figure 2: Architecture of the model. We first use a merged dataset to train the model, and then we add a prompt learning approach to the input sentence.

3.1.1. DATASETS

We conduct experiments on three widely used biomedical relation extraction datasets: DrugProt,DDI and BC5CDR, as shown in Table 1. DrugProt is a dataset and evaluation benchmark for drug-protein interactions (Drug-Protein Interaction). The DDI (Drug-Drug Interaction) dataset is a dataset used for drug interaction studies, recording information on interactions between different drugs. The BC5CDR (BioCreative V Chemical-Disease Relations) dataset is a biomedical text dataset for the identification of chemical substances and disease relationships. The BC5CDR dataset consists of abstracts and full-text articles in PubMed, covering the literature related to chemicals and diseases.

Table 1: Dataset statistics.

Task	Relation	SEN/DOC Level	Train	Dev	Test
DrugProt (Miranda-Escalada et al., 2021)	Drug-Protein	SEN	17274	3761	3491
DDI (Herrero-Zazo et al., 2013)	Drug-Drug	SEN	25296	2496	5716
BC5CDR	Disease- Chemical	DOC	1038	1012	1066

3.1.2. EVALUATION INDICATORS

To evaluate whether the model is valid, use the evaluation index macro-averaged F1 value in the official documents in the dataset. To calculate macro-averaged F1 values, the accuracy (Precision), recall (Recall) and F1 values must be calculated for each category, and finally the average F1 value for the entire dataset.

3.1.3. EXPERIMENTAL METHOD

We chose BioLink-BERT (Yasunaga et al., 2022) as a pre-trained language model because of its performance advantages on tasks in various biomedical fields. We trained the models using different training sets, observing the experimental results using different training sets.

D_{target}: We only used the target domain dataset D_{target} to train the model, and explore the model performance without using the deep transfer learning method.

D_{source}: We used the deep transfer learning method and used the merged dataset BioREx to train the model.

D_{train}: We used the deep transfer learning method and used the merged dataset BioREx and the target training set D_{target} to train the model.

3.2. Experimental Result

We report the comparison between the model and the baseline presented in this paper.

Table 2: Performance of the model on the datasets.

Model	DrugProt	DDI	BC5CDR
Sci-BERT (Beltagy et al., 2019)	-	81.32	62.89
Bio-BERT (Lee et al., 2020)	-	80.33	61.42
PubMedBERT (Tinn et al., 2023)	76.50	82.36	61.13
CK-RET(2023) (Sousa and Couto, 2023)	-	87.19	64.28
$\text{Ours}_{\text{D}_{\text{target}}}$	80.43	85.99	65.81
$\text{Ours}_{\text{D}_{\text{source}}}$	80.06	85.69	67.95
$\text{Ours}_{\text{D}_{\text{train}}}$	**80.82**	**87.37**	**69.95**

As shown in Table 2, $\text{Ours}_{\text{D}_{\text{train}}}$ achieves the SOTA performance on all three datasets. Powerful performance improvements validate that the use of deep transfer learning can improve the effectiveness of biomedical RE. It shows that in the entity relationship extraction task in the biomedical field, using deep transfer learning can effectively alleviate the problem of insufficient annotation corpus to a certain extent, and obtain good performance.

4. Conclusion

We propose a biomedical relationship extraction model based on deep transfer learning and prompt learning. By using the deep neural network for transfer learning, we use a large-scale biomedical field merged dataset to train the model, and apply the trained model to the common low-resource dataset in this field for experiments, so that the model can learn more abundant domain knowledge. So as to alleviate the problem of insufficient training data. Furthermore, we added a prompt learning approach to the input sentence that describes the predicted relationship labels to reduce the gap between PLM and downstream tasks. Experiments show that using deep transfer learning and prompt learning can effectively improve the prediction results. Moreover, the model in this paper adopts the ranking loss approach to improve the accuracy of prediction by optimizing the implied score of true relationships to make the ranking higher. Experiments were conducted on three widely used biomedical RE datasets, the experimental results show that the present model is effective in the low-resource biomedical field. Future work may further investigate the improvement of the performance of the biomedical relationship extraction model by providing prompt information through the GPT (Generative Pre-trained Transformer) model.

Acknowledgments

I would like to express my sincere thanks to my tutor and laboratory members for their careful guidance and cooperation. Thank your family and friends for their support and encouragement. Finally, thanks to all the contributors to the dataset and tools.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 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)*, pages 3615–3620, 2019.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920, 2013.

Po-Ting Lai, Chih-Hsuan Wei, Ling Luo, Qingyu Chen, and Zhiyong Lu. Biorex: improving biomedical relation extraction by leveraging heterogeneous datasets. *Journal of Biomedical Informatics*, 146:104487, 2023.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Antonio Miranda-Escalada, Farrokh Mehryary, Jouni Luoma, Darryl Estrada-Zavala, Luis Gasco, 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. In *Proceedings of the seventh BioCreative challenge evaluation workshop*, pages 11–21, 2021.

Diana F Sousa and Francisco M Couto. K-ret: knowledgeable biomedical relation extraction system. *Bioinformatics*, 39(4):btad174, 2023.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4), 2023.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*, 2022.