Biomedical Relation Extraction via Domain Knowledge and Prompt Learning

Jianyuan Yuan[1,†], Wei Du[1,†], Xiaoxia Liu[2] and Yijia Zhang[1,*]

[1]Dalian Maritime University, Dalian 116024, Liaoning, China
[2]Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA 94305, USA

Abstract

Biomedical relation extraction plays a crucial role in extracting crucial biomedical information from extensive literature, thereby supporting disease treatment and the construction of biomedical knowledge bases. However, biomedical texts contain highly technical language and domain-specific terminology, which makes it difficult for models to fully understand their semantics. Furthermore, imbalances in the distribution of samples across different categories in biomedical datasets result in reduced classification accuracy for categories with limited training samples. In this study, we propose a biomedical relation extraction model based on domain knowledge and prompt learning. The prompt template guides the model to focus on key features and information, so that more knowledge can be obtained from limited data. Utilizing domain knowledge to acquire refined entity representations, thereby mitigating the challenges posed by technical language and domain-specific terminology. The model is evaluated on the DDI Extraction 2013 dataset and the ChemProt dataset, and the experimental results demonstrate that our model can achieve state-of-the-art performance.

Keywords

biomedical relation extraction, prompt learning, biomedical literature, domain knowledge

1. Introduction

With the rapid development of the biomedical field, the amount of biomedical literature has exploded, which contains a wealth of biomedical information [1]. Biomedical relation extraction is a natural language processing technology whose purpose is to extract the relation between entities from biomedical text data [2]. This technology can help researchers quickly extract important biomedical information from literature, and provide important support for drug development and disease treatment [3].

The highly technical language and domain-specific terminology used in biomedical texts complicates this task, and traditional approaches often struggle to achieve high performance [4]. Moreover, there are differences in the number of samples of each category in the biomedical data set, resulting in low classification accuracy for categories with fewer training samples. Meanwhile, biomedical relation extraction usually requires a large amount of labeled data to effectively train the model. However, due to the huge amount of data, the cost of manual labeling is very high, and how to obtain more knowledge from limited data becomes very important [5].

The application of pre-trained language models in biomedical texts has received widespread attention and exploration [6]. Most of the current biomedical relation extraction methods mainly rely on pre-trained language models. Although the pre-trained language model has the ability to learn the general representation of language, there is a significant difference between the pre-training target and the downstream task fine-tuning, which has a very important impact on the performance of the model in the downstream task. As shown in Fig.1, since the target of unsupervised prediction of the input text sequence of the pre-trained model is inconsistent with the supervised classification operation of the downstream task, the model cannot fully apply its prior knowledge to the downstream task.

We propose a biomedical relation extraction model based on domain knowledge and prompt learning. Domain knowledge can provide entities with richer feature representations, which can better reflect the essence of entities and improve the effect of entity representation. Prompt learning is a method that can effectively bridge the gap between pre-training and fine-tuning on downstream tasks. The core idea of this method is to transform the traditional classification task into a cloze problem. By designing a prompt template, replace a word or a continuous short sentence (usually represented by [MASK]) in the input text with the corresponding label words, and ask the model to predict the label words. This approach makes the model need to consider more contextual information when predicting, so as to better understand the semantics of the input text. Overall, the contributions of this paper are as follows:

1) We propose a biomedical relation extraction model based on prompt learning, which can guide the model to focus on key features and information by constructing multiple task-related prompt. By introducing prompt learning, more knowledge can be obtained from limited data, which effectively alleviates the problem of insufficient knowledge that the model can learn when the amount of data is small.
2) The model obtains detailed information of biomedical entities through domain knowledge and obtains enhanced entity representation. In addition, special tokens are embedded around entities, enabling entities to better integrate domain knowledge, thereby reducing the impact of high-tech language and domain-specific terminology in biomedical texts on model performance.
3) The model is experimented on the ChemProt dataset and the DDI Extraction 2013 dataset. Experimental results demonstrate that the proposed model outperforms existing methods and achieves state-of-the-art performance in biomedical relation extraction.

*Corresponding author.
†These authors contributed equally.
jianyuany@dlmu.edu.cn (J. Yuan); duwei@dlmu.edu.cn (W. Du); xxliu@stanford.edu (X. Liu); zhangyijia@dlmu.edu.cn (Y. Zhang)
0000-0002-5843-4675 (Y. Zhang)

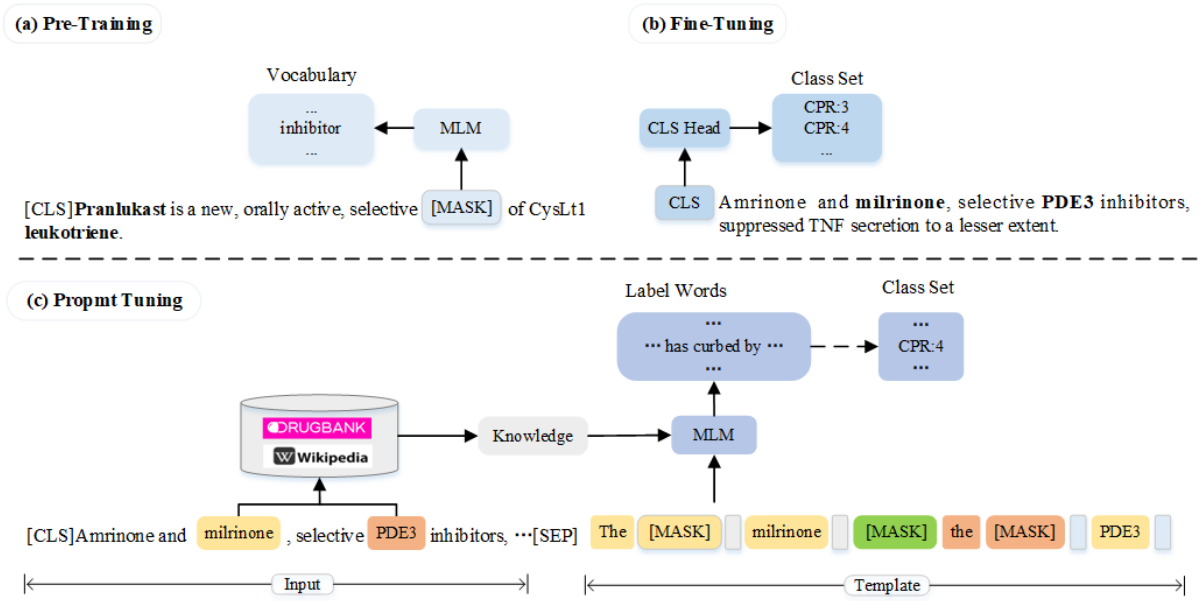

Figure 1: The instances of pre-training, fine-tuning, and prompt-tuning for relation extraction in the biomedical domain.

2. Related Work

Recently, various neural network-based approaches have demonstrated commendable outcomes in diverse relation extraction tasks and have been extensively employed in biomedical research. Liu et al. [7] utilized a convolutional neural network (CNN) model for biomedical relation extraction, demonstrating its effectiveness in achieving high performance. In this model, the words in the sentences of the biomedical dataset serve as inputs to the CNN, which can effectively capture local features. Liu et al. [8] introduced a model for biomedical relation extraction tasks,which is the dependency convolutional neural network (DCNN) model. By utilizing the dependency parse tree, the DCNN model can effectively capture the interdependency between words. Sasaki et al. [9] applied an attention-based CNN model to biomedical relation extraction tasks. Each word in a biomedical sentence has a varying impact on the final classification outcome in relation extraction. Kavuluru et al. [10] proposed a method that employs recurrent neural networks (RNNs) at the word and character levels to extract drug-drug interaction relations. Lim et al. [11] proposed a method using recurrent neural networks to automatically extract drug interactions in the literature. This method decomposes the text into a syntax tree and uses RNN to recursively process the tree structure to extract drug-drug interaction information.

Sahu et al. [12] used Long Short-Term Memory Network (LSTM) to automatically extract drug interaction information from biomedical texts. Mostafapour et al. [13] proposed a model that uses Bi-directional Long-Short Term Memory (BiLSTM) to model context information in text sequences and uses a hierarchical structure to consider different levels of semantic information. Wang et al.[14] used dependency parsing to model the relation between drugs in text and used the LSTM network to capture contextual information in text sequences. Huang et al. [15] employed a hybrid model consisting of a support vector machine(SVM) and LSTM for extracting drug interaction information. Zheng et al. [16] proposed a BiLSTM model with an attention mechanism to extract the interaction relation between drugs in biomedical texts. Zhang et al. [17] utilizes the shortest dependency path to determine the grammatical relations within a sentence, and extracts keywords located between two entities.

Peng et al. [18] proposed a multi-model approach that combines a SVM, CNN, and RNN to improve the performance of biomedical relation extraction. Sun et al. [19] improved biomedical relation extraction by integrating attention and ELMo representations with bidirectional LSTM networks. A neural model for extracting CPI was proposed by Zhang et al. [20], which utilized depth context representation and a multi-head attention mechanism. Xiong et al. [21] presented a model that utilizes a combination of a Graph Convolutional Neural Network (GCNN) and a LSTM network for extracting biomedical relations. Park et al. [22] utilized attention-based GCN for the task of biomedical relation extraction.

Peng et al. [23] applied the BERT (Bidirectional Encoder Representation of Transformer) model to the task of biomedical relation extraction. Lee et al. [24] extended the BERT model by training it on a large-scale biomedical corpus, resulting in the BioBERT model. Huang et al. [25] proposed an EMSI-BERT method for drug-drug interaction extraction. This method utilizes an asymmetric entity masking strategy and a symbol insertion structure. Sun et al. [26] proposed a model that uses a combination of Gaussian probability distribution and external biomedical knowledge to extract CPI. Sun et al. [27] proposed a model(BERT Att capsule) that utilizes a BERT-based attention-guided capsule network to extract CPI. This method uses attention mechanisms to guide the extraction of interactions and capsule networks to capture the interactions' semantic features. Liu et al. [28] proposed a grammar-enhanced model and a category keyword-based approach. The model uses graph-based grammar to build a syntactic tree and uses type keywords to guide the model to extract specific types of relations. Su et al. Su and Vijay-Shanker [6] explore the approaches to improve the BERT model for relation extraction tasks in both the pre-training and fine-tuning stages.

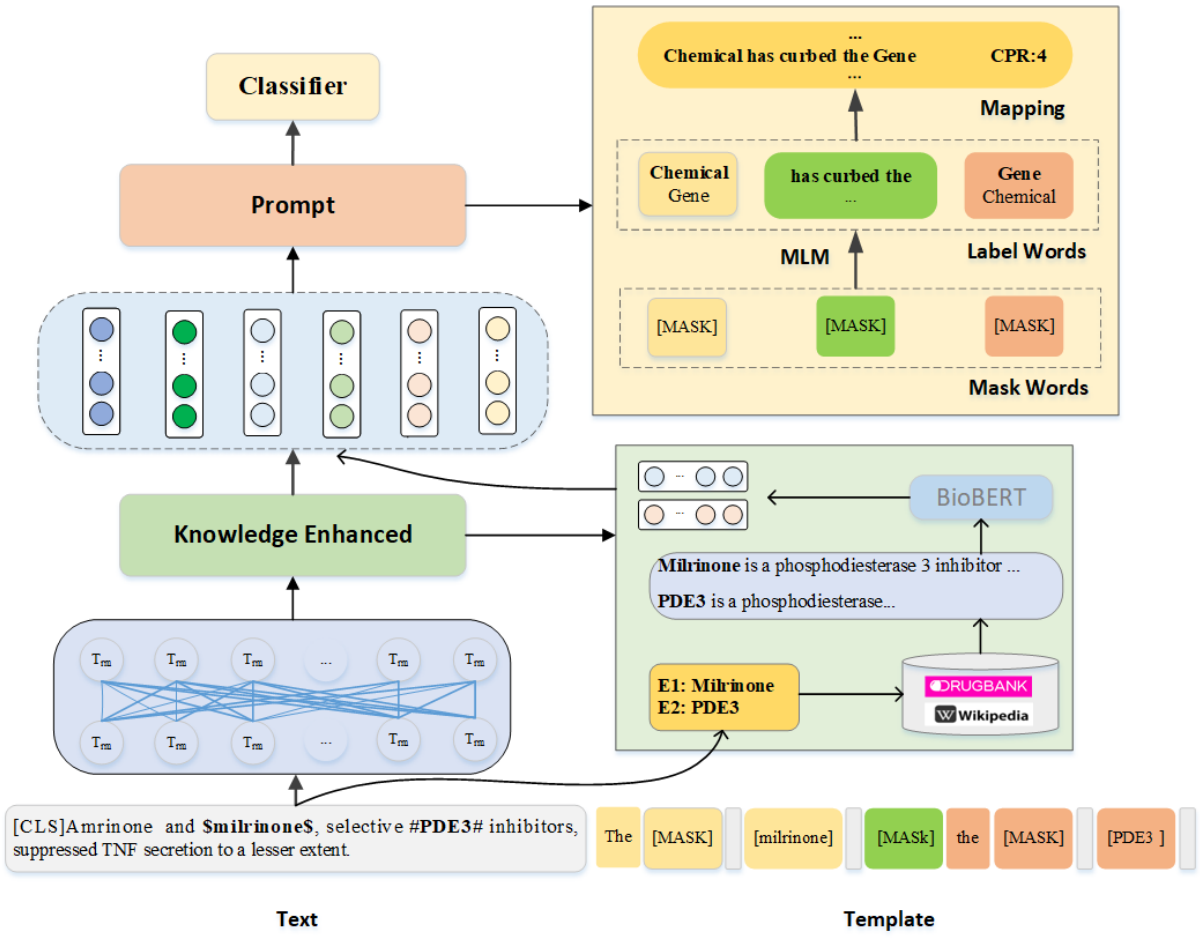

Figure 2: The schematic overview of the proposed model. The black arrows indicates the input stream.

3. Method

3.1. Problem Definition

Given a sentence sequence $S = \{c_1, c_2, \ldots, c_{n-1}, c_n\}$, where c is a word in sentence and n is the length of the sentence. The subject entity $e_1 = \{c_i, \ldots, c_j\}$ and the object entity $e_1 = \{c_x, \ldots, c_y\}$ are located in the same sentence. Biomedical relation extraction aims to identify the relation r between e_1 and e_2, where r is either selected from a predefined relation set R or NA.

3.2. Model Framework

Fig.2 shows the architecture of the biomedical relation extraction model based on domain knowledge and prompt learning. The model consists of four modules: input module, encoding module, knowledge enhancement module, and prompt learning module. We have designed three prompt templates, namely the prompt for biomedical entity $e1$, the prompt for biomedical entity relations, and the prompt for biomedical entity $e2$. Firstly, input biomedical text and prompt templates into the model for encoding. Then, the enhanced entity representation is obtained through knowledge enhancement. Finally, through the prompt module, the model can predict the label words at the [MASK] position and select their corresponding labels for classification.

3.3. Input Module

For the biomedical relation extraction task, it is represented as $T = X, Y$, where X represents the input text and Y represents the category label. The sentence in the biomedical dataset is represented as $x = \{x_1, \ldots, e_1, \ldots e_2, \ldots, x_n\}$, where e_1, e_2 represents two biomedical entities, respectively. A key part of prompt learning is to construct an appropriate template P and label word V. M: Y $\rightarrow$ V is a mapping that connects the task label with the label word V.

The model's input comprises two components, specifically the input text denoted as x and the prompt template denoted as $p(x)$. The sentence is subjected to tokenization, and each token is encoded using a vector of d dimensions. Moreover, an embedded "CLS" token is added at the beginning of each sentence sequence. To denote the boundaries of each biomedical entity, special symbols are introduced. The first entity is enclosed by "$" symbols on both sides, while the second entity is enclosed by "#" symbols on both sides.

In addition to retaining the original input in x, multiple [MASKs] need to be fed into the model. Three prompts are designed in the input prompt template, respectively, the prompt $p_{e1}(x)$ corresponding to the biomedical entity e_1, the prompt $p_r(x)$ corresponding to the biomedical entity relation and the prompt $p_{e2}(x)$ corresponding to the biomedical entity e_2. Denote the prompt template $p(x)$ corresponding to the input text x as:

$$p(x) = \{p_{e1}(x), p_r(x), p_{e2}(x)\} \quad (1)$$

The prompt $p_{e1}(x)$ corresponding to biomedical entity e_1 and the prompt $p_{e2}(x)$ corresponding to biomedical entity e_2 can be formalized as follows:

$$p_{e1}(x) = \{x, the[MASK]e_1\} \quad (2)$$

$$p_{e2}(x) = \{x, the[MASK]e_2\} \quad (3)$$

Then, the prompt $p_r(x)$ for the relation between biomedical entities is designed. For example, in the biomedical example sentence above, the relation type is CPR: 4, which means that the relation between entity e_1 and entity e_2 is "inhibition". The prompt template for the relation type is "e_1 [MASK] e_2", and the prompt label word is "has curved the". Prompt $p_r(x)$ for the relation corresponding to the input text x can be expressed as:

$$p_r(x) = \{x, e_1[MASK]e_2\} \quad (4)$$

The complete input composition can be formalized as follows:

$$Input_x = \{x, p(x)\} \quad (5)$$

3.4. Encode Module

BioBERT is a pre-trained model based on BERT, which is suitable for natural language processing tasks of biomedical texts. The BioBERT model is trained using a large corpus in the biomedical field, which can improve the text understanding and classification performance in the biomedical field, making BioBERT a model widely used in the biomedical natural language processing field.

The model's input consists of biomedical text and a prompt template, wherein [MASK] denotes the portion that requires completion by the model. Within the input sequence, [MASK] is substituted with a special token, signifying its prediction requirement. To ensure the model comprehends the word's position within the sentence, each word embedding vector is added to its corresponding position vector in the sequence. The Transformer architecture is employed to encode the sequence of embedding vectors and position vectors. This architecture comprises multiple layers, each containing a multi-head attention mechanism and a feed-forward neural network. Each layer encodes an input vector sequence to extract its representation. This encoding approach effectively captures both the semantic and syntactic information present in the input sequence, thereby enhancing the model's ability to predict the content to fill the [MASK].

3.5. Knowledge Enhancement Module

Biomedical entities are sourced from Wikipedia and DrugBank using crawler technology to obtain interpretation information in the biomedical domain. This interpretation information is denoted as S_e = { E_1,E_2,E_3,...,E_N }, where E_i represents the $i-th$ word and N represents the length of the sentence.

The vector e_1 for a biomedical entity is computed as the average of the hidden layer vectors from H_i to H_j in the model. Similarly, the vector e_2 for another biomedical entity is obtained as the average of the hidden layer vectors from H_k to H_m in the model. The calculation formulas for these vectors are as follows:

$$H_1' = W_1 \left[tanh \left(\frac{1}{j-i+1} \sum_{t=i}^{j} H_t \right) \right] + b_1 \quad (6)$$

$$H_2' = W_2 \left[tanh \left(\frac{1}{m-k+1} \sum_{t=k}^{m} H_t \right) \right] + b_2 \quad (7)$$

where $W_1 \in R^{d \times d}$, $W_2 \in R^{d \times d}$ denote weight matrices. b_1, b_2 denote bias vectors.

The semantic feature representation of domain knowledge is acquired by the model using BioBERT. This vector is then combined with entity interpretation information and the corresponding entity vector to generate an improved vector representation of biomedical entities. When a sentence S_e containing biomedical knowledge is successfully matched with entity e_1, the final hidden layer vector H_{e1} of "CLS" can be obtained from BioBERT. The acquired enhanced representation is integrated into the model, with the calculation formulas being as follows:

$$H_{E1} = W_4 \left[concat \left(H_1', W_3 \left(tanh \left(H_{e1} \right) \right) + b_3 \right) \right] + b_4 \quad (8)$$

$$H_{E2} = W_6 \left[concat \left(H_2', W_5 \left(tanh \left(H_{e2} \right) \right) + b_5 \right) \right] + b_6 \quad (9)$$

where W_3, W_4, W_5, W_6 denote weight matrices. b_3, b_4, b_5, b_6 denote bias vectors.

3.6. Prompt Learning Module

In the prompt module, multiple prompts are combined directly to form a complete prompt for a specific task. The complete prompt template is as follows:

$$\begin{aligned} p(x) = x \quad & the \quad [MASK]_1 \quad e_1 \\ & [MASK]_2 \quad the \quad [MASK]_3 \quad e_2 \end{aligned} \quad (10)$$

where $[MASK]_1$ is mask of entity, $[MASK]_2$ is mask of entity relation, and $[MASK]_3$ is mask of entity. The corresponding label words are as follows:

$$V_{[MASK_1]} = \{Chemical,\ Gene, Drug\} \quad (11)$$

$$V_{[MASK_2]} = \{has\,activated\,the,\ has\,curbed\,the,\ ...\} \quad (12)$$

$$V_{[MASK_3]} = \{Chemical,\ Gene, Drug\} \quad (13)$$

Due to the possibility that the aggregated template may contain multiple [MASK], all masked locations must be considered for prediction. In the sentence, each [MASK] is equivalent to a classification mapped to a label word. Each position will get a corresponding probability, and the probability of the entire sentence being predicted correctly is the cumulative multiplication of the probabilities of each position. The final probability calculation formula is as follows:

Table 1
The label words for the prompt of the CPI dataset

Class Label	Prompt1	Prompt2	Prompt3
False	Gene/Chemical	has nothing to	Chemical/Gene
CPR: 3	Gene/Chemical	was activated by	Chemical/Gene
CPR: 4	Gene/Chemical	has curbed by	Chemical/Gene
CPR: 5	Gene/Chemical	is agitation of	Chemical/Gene
CPR: 6	Gene/Chemical	is antagonist of	Chemical/Gene
CPR: 9	Gene/Chemical	is substrate of	Chemical/Gene

$$P\left(y|x\right)=\prod_{j=1}^{n}P\left(\left[MASK_j\right]=\phi_j\left(y\right)|p\left(x\right)\right) \quad (14)$$

where n is the number of mask positions in $p(x)$, and $\phi_j\left(y\right)$ is the label word set $V_{\left[MASK_j\right]}$ that maps class y to the $j-th$ mask position $\left[MASK\right]_j$.

During the training process, the model will predict the [MASK] part of the input sequence through the masked language model (MLM) according to the information in the context, which makes the goal of the model consistent with the task goal of the MLM, thus effectively reducing the pre-training and downstream task gap.

In our model, label words are critical to accurately classify the relation between biomedical entities. We design a set of label words for each relation type, and further verify their effectiveness by using them for model training and testing. Entity label words refer to words that describe biomedical entity types, such as Chemical or Gene. Label words can help the model better understand entity types and thus correctly predict the relation between entities.

Relational label words are key short sentences describing the relation types of biomedical entities, which are very important for the classification results of biomedical entities. During the learning process of the model, fill in the prediction result of [MASK] and the closest set of label words in the label word set, and the relation label words can make the model better understand the relation between biomedical entities. Table 1 and Table 2 show the details of biomedical entity label words and relation label words in CPI dataset and DDI dataset respectively.

4. Experiments and Discussion

4.1. Datasets and Evaluation Metrics

The performance of the model is evaluated by the DDI Extraction 2013 dataset [29] and the ChemProt dataset [30].

DDI Extraction 2013 Dataset

The DDI Extraction 2013 dataset is a dataset for extracting drug-drug interaction relations. This dataset contains medical texts from multiple sources such as DrugBank and MedLine. DrugBank provides drug names, chemical formulas, and pharmacological information, while MedLine provides abstracts and full-text articles containing DDI information. All drug pairs in the text are annotated as having or not having interactions, with a total of four types of interactions, namely Advice, Effect, Mechanism, and Int. The quantity statistics of the dataset are shown in Table 3.

ChemProt Dataset

ChemProt dataset is a benchmark dataset used for extracting chemical-protein interactions (CPIs) from biomedical literature. The dataset consists of documents from PubMed and PubMed Central, which are annotated with different types of CPIs, such as inhibition, activating. The dataset was originally created for the BioCreative IV challenge in 2013, and has since become a widely used benchmark dataset in the field of biomedical natural language processing. Detailed statistics are shown in Table 4.

Evaluation Metrics

To assess the efficacy of the proposed model, its performance is measured using precision, recall, micro-F1 and macro-F1 metrics. In particular, the micro-averaged metrics are employed to derive an average metric by amalgamating the contributions of all classes. The macro-F1 score is more effective in accurately reflecting the superior performance of the model in classes with fewer samples.

Table 2
The label words for the prompt of the DDI dataset

Class Label	Prompt1	Prompt2	Prompt3
False	DRUG	has nothing to	DRUG
Advice	DRUG	need advice with	DRUG
Mechanism	DRUG	generate mechanisms with	DRUG
Effect	DRUG	make effect with	DRUG
Int	DRUG	will interact with	DRUG

Table 3
Statistics of the DDI corpus

Relation type	Train set		Test set	
	DrugBank	MEDLINE	DrugBank	MEDLINE
Advice	818	8	214	7
Mechanism	1257	62	278	24
Effect	1535	152	298	62
Int	178	10	94	2
Negative	22217	1555	4381	401
Total	26005	1787	5265	496

Table 4
Statistics of the CPI corpus.

Relation type	Training set	Development set	Test set
CPR:3	768	550	665
CPR:4	2251	1094	1661
CPR:5	173	116	195
CPR:6	235	199	293
CPR:9	727	457	644
False	15306	9404	13485
Total	19460	11820	16943

Table 5
The setting of hyper-parameters parameter.

Parameter Name	Value
Sentence feature dimension	768
Max sentence length	512
Number of hidden layers of BioBERT	12
Batch size	8
Learning rate	2e-5
Epoch	10
Dropout rate	0.1
Weight decay	1e-5

Table 6
Performance comparison on the DDI dataset.

Model	F1-score on each type				P	R	Micro-F1	Macro-F1
	Advice	Mechanism	Effect	Int				
CNN	77.7	70.2	69.3	46.4	75.7	64.7	69.8	65.9
DCNN	78.2	70.6	69.9	46.4	77.2	64.4	70.2	66.3
ACNN	-	-	-	-	76.3	63.3	69.1	-
RNN	-	-	-	-	78.6	63.8	72.1	-
LSTM	80.3	72.3	65.5	44.1	74.5	65.0	69.4	65.5
Two-stage LSTM	-	-	-	-	-	-	69.0	-
ASDP-LSTM	80.3	74.0	71.8	54.3	74.1	71.8	72.9	70.1
ATT-BLSTM	85.1	77.5	76.6	57.7	78.4	76.2	77.3	74.2
AGCN	86.2	78.7	74.2	52.6	78.2	75.6	76.9	72.9
BERT	-	-	-	-	-	-	78.8	-
BioBERT	-	-	-	-	79.9	78.1	79.0	-
EMSI-BERT	**86.8**	**86.6**	80.7	56.0	-	-	82.0	**77.5**
SECK[28]	-	-	-	-	83.0	81.1	82.0	-
Our model	84.3	78.4	**86.3**	**58.2**	**84.2**	**83.4**	**83.8**	76.8

Table 7
Performance comparison on the CPI dataset.

Model	F1-score on each type					P	R	Micro-F1	Macro-F1
	CPR:3	CPR:4	CPR:5	CPR:6	CPR:9				
[31]	49.8	66.5	56.5	69.6	28.3	63.5	51.2	56.7	54.1
LSTM	-	-	-	-	-	59.1	67.8	63.1	-
GA-BGRU [32]	-	-	-	-	-	65.4	64.8	65.1	-
[20]	59.4	71.8	65.7	72.5	50.1	70.6	61.8	65.9	63.9
Bi-LSTM	64.7	75.3	68.1	79.3	55.7	67.0	72.0	69.4	68.6
BERT	-	-	-	-	-	74.5	70.6	72.5	-
Sun et al [26]	71.5	81.3	70.9	79.9	**69.9**	77.1	76.1	76.6	74.7
BERT-Att-Capsule	72.9	78.6	72.7	77.9	64.4	77.8	71.7	74.7	73.3
BioBERT	-	-	-	-	-	77.0	75.9	76.5	-
Our model	**74.3**	**81.4**	**77.7**	**82.3**	69.4	**80.0**	**81.1**	**80.5**	**77.1**

Table 8
Ablation study of the model.

Model	DDI 2013			ChemProt		
	P	R	Micro-F1	P	R	Micro-F1
Our Model (DMPL)	**84.2**	**83.4**	**83.8**	**80.0**	**81.1**	**80.5**
DMPL w/o DK	82.9	82.5	82.7	77.9	79.9	78.9
DMPL w/o PL	82.8	81.0	81.9	78.0	77.6	77.8
DMPL w/o DK w/o PL	81.8	80.7	81.3	77.9	76.9	77.4
BioBERT	79.9	78.1	79.0	77.0	75.9	76.5

4.2. Experimental Settings

Implement the model proposed in this article through the Python programming language and PyTorch development framework. The Python language has good compatibility with existing deep learning frameworks. Set the batch size to 8. During the training process, an Adam optimizer was used to optimize the parameters that affect model training and output. Set the maximum sentence length to 512 and the learning rate of the model to 2e-5. The experimental parameter settings are detailed in Table 5.

4.3. Experimental Results

Comparison with Other Models The CPI dataset and the DDI dataset were employed to evaluate the performance of the model. Table 6 presents the experimental results of the model and other approaches on the DDI dataset. Precision, recall, Micro-F1 and Macro-F1 scores were used to assess the model's performance. The Micro-F1 and Macro-F1 scores provide a comprehensive evaluation of the model's performance, with higher values indicating better performance. The model achieved P, R, Micro-F1 and Macro-F1 scores of 84.2%, 83.4%, 83.8% and 76.8%, respectively, better than achieved baselines on the DDI dataset .

Furthermore, the model achieved F1-scores of 84.3%, 78.4%, 86.3%, and 58.2% in the Advice, Mechanism, Effect, and Int categories, respectively. Notably, the F1-scores in the Int type, which has limited data, surpassed those of other methods. Comparison results with alternative models suggest that the model proposed in this study effectively enhances biomedical relation extraction performance.

Table 7 exhibits the comparison results between this model and other approaches on the CPI dataset. The model achieved P, R, Micro-F1 and Macro-F1 scores of 80.0%, 81.1%, 80.5% and 77.1%, respectively, representing a 4% improvement in Micro-F1 score compared to the BioBERT model. Moreover, the model obtained F1-scores of 74.3%, 81.4%, 77.7%, 82.3%, and 69.4% in the CPR:3, CPR:4, CPR:5, CPR:6, and CPR:9 types, respectively. The comparison results with other models demonstrate that the model proposed in this paper effectively enhances the classification performance of types with limited data.

Ablation Study The ablation studies were conducted to assess the individual contributions of each module in the model towards the overall performance. The outcomes of these studies are presented in Table 8. After eliminat-

Sentence instances of biomedical dataset	Prediction results	Prediction results
Case1: Glutathione-independent prostaglandin D synthase [<e1>prostaglandin-H2 D-isomerase</e1>; (5Z,13E)-(15S)-9 alpha,11 alpha-epidioxy-15-hydroxyprosta-5,13-dienoate D-isomerase, EC 5.3.99.2] is an enzyme responsible for biosynthesis of <e2>prostaglandin D2</e2> in the central nervous system.	BioBERT: Negative	Our model: CPR:9
Case2: Drugs that induce hepatic enzymes such as phenobarbital, <e1>phenytoin</e1>, and rifampin may increase the clearance of methylprednisolone and may require increased in <e2>methylprednisolone</e2> dose to achieve the desired response.	BioBERT: Mechanism	Our model: Advice
Case3: Use with Anticholinergics: Because of their mechanism of action, <e1>cholinesterase inhibitors</e1> have the potential to interfere with the activity of <e2>anticholinergic medications</e2>.	BioBERT: Mechanism	Our model: Int

Figure 3: Examples of extraction results by different methods.

Table 9
Comparative experimental results of low resource biomedical relation extraction.

Dataset	Model	8-shot	16-shot	32-shot
CPI	BERT	10.37	16.85	25.01
	BioBERT	17.41	24.17	31.02
	SCIBERT	18.26	23.59	30.76
	Our Model	**29.67**	**35.58**	**41.15**
DDI	BERT	12.76	20.45	29.58
	BioBERT	21.57	27.14	34.82
	SCIBERT	22.01	26.32	34.17
	Our Model	**33.90**	**38.26**	**43.29**

ing domain knowledge from the model, the Micro-F1 score decreases by 1.1% and 1.6% in the DDI and CPI datasets, respectively. The experimental findings indicate that domain knowledge plays a moderating role in mitigating the impact of domain-specific terminology on model performance.

When prompt learning is removed from the model, the Micro-F1 score experiences a decline of 1.9% and 2.7% in the DDI dataset and the CPI dataset, respectively. We hypothesize that prompt learning can narrow the gap between pre-training and downstream tasks, enabling the model to acquire more knowledge from limited data and thereby enhancing the effectiveness of biomedical relation extraction. Upon removing both domain knowledge and prompt learning from the proposed model, the Micro-F1 score exhibits a decrease of 2.5% and 3.1% in the DDI dataset and CPI dataset, respectively. The experimental results demonstrate that domain knowledge and prompt learning are crucial components of the model, contributing significantly to the improvement of biomedical relation extraction performance.

4.4. Low-resource Results

The dataset in the relation extraction task usually requires manual annotation of a large amount of high-quality data, which requires the participation of domain experts. However, the cost of collecting these data is high, especially in the biomedical field. Therefore, in the context of resource scarcity, how to make the model fully utilize existing data to achieve better performance has become a highly concerned issue.

The relation extraction performance of the model is evaluated by simulating low-resource relation extraction when biomedical data is scarce. The K-shot support set is constructed using the training set of the biomedical dataset, where each entity type contains K samples. To simulate low-resource biomedical relation extraction, 8, 16, and 32 samples are sampled for each entity type, and each relation type is sampled at least once. Table 9 shows the comparison of biomedical relation extraction performance of our model and other pre-trained models under low-resources.

According to the comparative findings presented in Table 9, it is evident that our model exhibits commendable performance in scenarios characterized by limited resources. In such instances, our model surpasses other pre-trained models in terms of efficacy. Notably, even when working with a relatively modest data volume at K=8, our model manages to attain desirable outcomes. Even upon increasing K to 16, the F1 score of our model remains superior to

that of other models. As K is further elevated to 32, the discrepancy between our model and other pre-trained models gradually diminishes alongside the expansion of the sample size. Nevertheless, our model's performance continues to outshine that of other models. Empirical evidence substantiates the notion that our model effectively enhances the accuracy of biomedical relation extraction when confronted with limited resources.

4.5. Case Study

As shown in Figure 3, we selected some examples from the biomedical dataset for detailed analysis. We compare the prediction results of BioBERT with our model. According to Case 1, the result of the BioBERT model is Negative, indicating that the prediction is incorrect, while our model is CPR: 9, indicating that the prediction is correct. The sentence contains multiple biomedical entities, which makes it difficult for the model to fully learn the Semantic information of biomedical text. Upon integrating the biomedical entities with the expertise found in the knowledge base, the model is fortified to represent the said entities, facilitating a better understanding of the textual information. The prediction results show that our model can obtain enhanced text representation after integrating domain knowledge, and improve the classification effect in sentences containing complex biomedical entities.

According to Case 2, there are multiple biomedical entities in the sentence, which makes it difficult for the model to fully learn the Semantic information of biomedical text, and BioBERT model makes wrong predictions. Our model fused domain knowledge and made correct predictions. According to Case 3, the BioBERT model incorrectly predicts Int type text as Mechanism. The small number of Int type training samples makes it difficult for the BioBERT model to fully learn its class characteristics. Our model can obtain more knowledge from limited data by introducing prompt learning, effectively alleviating the problem of insufficient learning knowledge when the data volume is small. Therefore, our model made the correct prediction.

5. Conclusion

In this study, we propose a biomedical relation extraction model based on domain knowledge and prompt learning. The model can enhance entity representation by integrating domain knowledge, thus reducing the impact of highly technical languages and domain specific terms in biomedical texts on model performance. By introducing prompt learning, more knowledge can be obtained from limited data, effectively alleviating the problem of insufficient knowledge that models can learn when the data volume is small, thereby improving the classification effect of biomedical relation. The experimental results show that the model can effectively improve the accuracy of biomedical relation extraction by introducing domain knowledge and prompt learning.

In the future, we will continue to explore the potential of prompt learning, try different prompt methods, and apply our model to document-level relation extraction.

Acknowledgments

This work is supported by grant from the Natural Science Foundation of China (No. 62106034)

References

[1] S. Zhao, C. Su, Z. Lu, F. Wang, Recent advances in biomedical literature mining, Briefings in Bioinformatics 22 (2021) bbaa057.

[2] T. Zhang, J. Leng, Y. Liu, Deep learning for drug–drug interaction extraction from the literature: a review, Briefings in bioinformatics 21 (2020) 1609–1627.

[3] Y. Zhang, H. Lin, Z. Yang, J. Wang, Y. Sun, B. Xu, Z. Zhao, Neural network-based approaches for biomedical relation classification: a review, Journal of biomedical informatics 99 (2019) 103294.

[4] Y. Qiu, Y. Zhang, Y. Deng, S. Liu, W. Zhang, A comprehensive review of computational methods for drug-drug interaction detection, IEEE/ACM transactions on computational biology and bioinformatics 19 (2021) 1968–1985.

[5] Q. Zhao, D. Xu, J. Li, L. Zhao, F. A. Rajput, Knowledge guided distance supervision for biomedical relation extraction in chinese electronic medical records, Expert Systems with Applications 204 (2022) 117606.

[6] P. Su, K. Vijay-Shanker, Investigation of improving the pre-training and fine-tuning of bert model for biomedical relation extraction, BMC bioinformatics 23 (2022) 120.

[7] S. Liu, B. Tang, Q. Chen, X. Wang, et al., Drug-drug interaction extraction via convolutional neural networks, Computational and mathematical methods in medicine 2016 (2016).

[8] S. Liu, K. Chen, Q. Chen, B. Tang, Dependency-based convolutional neural network for drug-drug interaction extraction, in: 2016 IEEE international conference on bioinformatics and biomedicine (BIBM), IEEE, 2016, pp. 1074–1080.

[9] M. Asada, M. Miwa, Y. Sasaki, Extracting drug-drug interactions with attention cnns, in: BioNLP 2017, 2017, pp. 9–18.

[10] R. Kavuluru, A. Rios, T. Tran, Extracting drug-drug interactions with word and character-level recurrent neural networks, in: 2017 IEEE International Conference on Healthcare Informatics (ICHI), IEEE, 2017, pp. 5–12.

[11] S. Lim, K. Lee, J. Kang, Drug drug interaction extraction from the literature using a recursive neural network, PloS one 13 (2018) e0190926.

[12] S. K. Sahu, A. Anand, Drug-drug interaction extraction from biomedical texts using long short-term memory network, Journal of biomedical informatics 86 (2018) 15–24.

[13] V. Mostafapour, O. Dikenelli, Attention-wrapped hierarchical blstms for ddi extraction, arXiv preprint arXiv:1907.13561 (2019).

[14] W. Wang, X. Yang, C. Yang, X. Guo, X. Zhang, C. Wu, Dependency-based long short term memory network for drug-drug interaction extraction, BMC bioinformatics 18 (2017) 99–109.

[15] D. Huang, Z. Jiang, L. Zou, L. Li, Drug–drug interaction extraction from biomedical literature using support vector machine and long short term memory networks, Information sciences 415 (2017) 100–109.

[16] W. Zheng, H. Lin, L. Luo, Z. Zhao, Z. Li, Y. Zhang, Z. Yang, J. Wang, An attention-based effective neural model for drug-drug interactions extraction, BMC bioinformatics 18 (2017) 1–11.

[17] Y. Zhang, W. Zheng, H. Lin, J. Wang, Z. Yang, M. Du-

montier, Drug–drug interaction extraction via hierarchical rnns on sequence and shortest dependency paths, Bioinformatics 34 (2018) 828–835.
[18] Y. Peng, A. Rios, R. Kavuluru, Z. Lu, Extracting chemical–protein relations with ensembles of svm and deep learning models, Database 2018 (2018) bay073.
[19] C. Sun, Z. Yang, L. Luo, L. Wang, Y. Zhang, H. Lin, J. Wang, A deep learning approach with deep contextualized word representations for chemical–protein interaction extraction from biomedical literature, IEEE Access 7 (2019) 151034–151046.
[20] Y. Zhang, H. Lin, Z. Yang, J. Wang, Y. Sun, Chemical–protein interaction extraction via contextualized word representations and multihead attention, Database 2019 (2019) baz054.
[21] W. Xiong, F. Li, H. Yu, D. Ji, Extracting drug-drug interactions with a dependency-based graph convolution neural network, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2019, pp. 755–759.
[22] C. Park, J. Park, S. Park, Agcn: Attention-based graph convolutional networks for drug-drug interaction extraction, Expert Systems with Applications 159 (2020) 113538.
[23] Y. Peng, S. Yan, Z. Lu, Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets, arXiv preprint arXiv:1906.05474 (2019).
[24] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.
[25] Z. Huang, N. An, J. Liu, F. Ren, Emsi-bert: Asymmetrical entity-mask strategy and symbol-insert structure for drug–drug interaction extraction based on bert, Symmetry 15 (2023) 398.
[26] C. Sun, Z. Yang, L. Su, L. Wang, Y. Zhang, H. Lin, J. Wang, Chemical–protein interaction extraction via gaussian probability distribution and external biomedical knowledge, Bioinformatics 36 (2020) 4323–4330.
[27] C. Sun, Z. Yang, L. Wang, Y. Zhang, H. Lin, J. Wang, Attention guided capsule networks for chemical-protein interaction extraction, Journal of Biomedical Informatics 103 (2020) 103392.
[28] X. Liu, J. Tan, J. Fan, K. Tan, J. Hu, S. Dong, A syntax-enhanced model based on category keywords for biomedical relation extraction, Journal of Biomedical Informatics 132 (2022) 104135.
[29] M. Herrero-Zazo, I. Segura-Bedmar, P. Martínez, T. Declerck, The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions, Journal of biomedical informatics 46 (2013) 914–920.
[30] J. Kringelum, S. K. Kjaerulff, S. Brunak, O. Lund, T. I. Oprea, O. Taboureau, Chemprot-3.0: a global chemical biology diseases mapping, Database 2016 (2016) bav123.
[31] P.-Y. Lung, T. Zhao, Z. He, J. Zhang, Extracting chemical protein interactions from literature, in: Proceedings of the BioCreative VI Workshop, 2017, pp. 159–162.
[32] H. Lu, L. Li, X. He, Y. Liu, A. Zhou, Extracting chemical-protein interactions from biomedical literature via granular attention based recurrent neural networks, Computer methods and programs in biomedicine 176 (2019) 61–68.