# PTDA: Improving Drug-Drug Interaction Extraction from Biomedical Literature Based on Prompt Tuning and Data Augmentation

Kailiang Wang, *Member, IAENG*, Xufeng Fu*, Yanping Liu, Weikun Chen, and Jun Chen

***Abstract*—Drug-drug interaction (DDI) extraction in biomedical literature is a particular entity relation classification task, which is significant for biopharmaceutical research. Current researchers mainly concentrated on fine-tuning the Pre-trained Language Model (PLM) on medical datasets to forecast the interactions between target drugs. Despite the above approaches having shown positive results, it is still confronted with some serious limitations. First of all, ordinary fine-tuning makes it difficult to take the best advantage of prior knowledge embedded in the PLM, due to the wide gap between objective forms of pre-training and downstream tasks. Secondly, the issue of limited training data or category imbalance also poses a great challenge for real-world DDI extraction tasks. In this work, we propose a novel model using Prompt Tuning and Data Augmentation (PTDA) to extract the DDI. This method can augment training data and alleviate the adverse effects of category imbalance by properly utilizing contextual word embedding substitution to generate examples for specific relation types. Furthermore, we have tried to introduce prompt tuning into the DDI extraction process, aiming to narrow the gap between pre-training and downstream tasks to leverage the prior knowledge in the PLM fully. We conduct a series of experiments on diverse biomedical datasets to verify the performance of our model. The results show that PTDA hugely outperforms existing DDI extraction methods, achieving F1-micro of 84.9%, 79.8%, and 73.1% on benchmark datasets, DDI 2013, ChemProt, and DTIs, respectively. Therefore, we believe that the PTDA model holds considerable potential for future practical applications.**

***Index Terms*—drug-drug interaction; relation extraction; pre-trained language model; data augmentation; prompt tuning**

## I. Introduction

DRUG-DRUG INTERACTION (DDI) describes the concurrent or sequential usage of two or more drugs, in which the magnitude, duration, or even the nature of one drug is significantly altered by the influence of other drugs or chemical substances [1]. Adverse drug-drug interactions can diminish the efficacy of drugs or even produce toxic substances that cause death in patients. For this reason, healthcare professionals must devote a lot of time to studying relevant literature and drug databases to comprehend the potential interactions between various medications to avoid harmful DDI. Currently, it has become a major challenge to identify the interaction between two drugs, due to the exponential growth of medical literature.

Initially, researchers relied on manual methods to collect DDI from documents to build medical databases [2]. However, these approaches rely heavily on manual intervention, which results in the creation and maintenance of databases that consume significant time and labor resources. This is coupled with slow and inefficient knowledge updating, as well as limited coverage. The identified flaws make it challenging to fulfill the actual requirements of drug-related research and clinical applications in the rapidly expanding data ecosystem, characterized by a dramatic increase in the size and complexity of data. Consequently, the automatic extraction of DDI from unstructured biomedical documents has emerged as a key research focus.

In comparison to manually established databases, machine learning-based methods can automatically extract DDI from large amounts of unstructured literature with high accuracy, which alleviates the consumption for manual construction of knowledge bases to a certain extent. However, these approaches rely heavily on specialized domain knowledge during the feature selection phase. Therefore, the automatic extraction of DDI using neural networks is becoming increasingly popular. Unlike traditional machine learning, neural networks can extract DDI between any two drugs from unstructured biomedical documents quickly and accurately, without the need for complex manual feature selection. This approach can substantially save time and labor costs.

In recent years, fine-tuning has started to be introduced into the field of DDI extraction with the rise of Pre-trained Language Models (PLMs), such as BERT [3] and GPT-2 [4]. Fine-tuning effectively leverages the extensive prior knowledge acquired by the model during the pre-training phase and has demonstrated excellent performance across various domains. Among them, the language models trained on the biomedical corpora, such as PubMedBERT [5] and BioBERT [6], are outstanding in biomedical-related tasks.

However, fine-tuning is still confronted with some serious limitations. On one hand, recent studies have revealed that the wide gap between the objective forms of pre-training and the downstream task severely limits the utilization of the

Manuscript received August 30, 2023; revised February 2, 2023. This work was supported by the National Natural Science Foundation of China Grant 61762063 and the Research Project of the Education Department of Jiangxi Province Grant GJJ170991.

Xuefeng Fu is an associate professor of Nanchang Institute of Technology (NIT), Nanchang 330099, China. (Corresponding author to provide phone: 86-18170936669; e-mail: fxf@nit.edu.cn).

Kailiang Wang is a graduate student of Nanchang Institute of Technology (NIT), Nanchang, China. (e-mail: WKLiang1995@163.com)

Yanping Liu is a graduate student of Nanchang Institute of Technology (NIT), Nanchang, China. (e-mail: 1796164455@qq.com)

Weikun Chen is a graduate student of Nanchang Institute of Technology (NIT), Nanchang, China. (e-mail: 1401619099@qq.com)

Jun Chen is a graduate student of Nanchang Institute of Technology (NIT), Nanchang, China. (e-mail: 1646428688@qq.com)

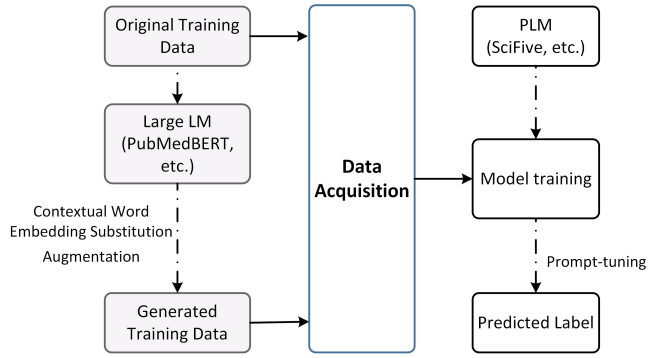


Fig. 1. The conceptual diagram of text augmentation and prompt tuning using large-scale language models.

prior knowledge embedded in PLMs, which in turn degrades the model's performance. On the other hand, most biomedical datasets used for training models often suffer from the issue of limited data or class imbalance due to the extreme amount of redundant information in biomedical documents, which causes a serious negative influence on model training.

Recently, it has been suggested by some researchers that generative models, such as GPT-2, could be used to generate samples to address the issue of limited training data [7]. The experimental results have proved that this approach can alleviate the negative effects of insufficient data and class imbalance to some extent. However, we have observed that the low semantic similarity between the generated samples and the original samples makes it hard to express the classification information clearly. The introduction of a large number of generated examples results in massive noise being added to the dataset, which can seriously interfere with model training.

To address the above challenges, we propose a novel DDI extraction approach based on prompt tuning and data augmentation (PTDA). Our approach leverages contextual word embedding substitution to generate instances based on the original sentences' contextual information. This approach mitigates the adverse effects of insufficient data and class imbalance while allowing generated examples to maintain a high relevance with the original samples in semantics. Simultaneously, we further employ prompt tuning instead of fine-tuning to narrow the gap between pre-training and downstream tasks. The conceptual diagram of the paper is shown in Figure 1.

The key contributions of our work can be summarized as follows:

(a) This paper proposes a novel data augmentation technique that leverages contextual word embedding substitution supported by large-scale language models to generate realistic text samples from a mixture of real samples. The approach not only mitigates the adverse effects of insufficient data and class imbalance but also ensures that semantic similarity between generated examples and original samples has existed forever, thus reducing the information interference in the dataset caused by introducing generated samples.

(b) We introduce prompt tuning into the domain of DDI extraction and make the appropriate adjustment to the relevant datasets and tasks so that the gap between the pre-training phase and downstream task can be narrowed to better unlock the potential of large-scale language models in the biomedical domain.

(c) Without using any external resources and manual annotation, our approach achieves a great improvement in performance on the benchmark datasets DDI 2013, ChemProt and DTIs, compared to the baseline models in the paper.

## II. Related Work

The goal of the entity relation extraction task is to identify the subject, predicate, and object from the literature to construct triples. As a special type in the relation classification task, DDI extraction treats target drugs as subjects and objects, while the predicate refers to drug-drug interactions. Over the past decade, researchers have developed many excellent DDI extraction models due to the emergence of numerous biomedical datasets. So far, these models can be classified into two main categories: traditional machine learning-based approaches [8-11] and deep learning-based approaches [12-15].

Feature-based approaches are the most common among traditional machine learning. Typical features in the method include but are not limited to, word features, contextual features, syntactic features, etc. In 2015, Kim et al [16] first proposed an abundant feature-based method for DDI extraction, which fused multiple features as a relation representation for classification and achieved quite excellent success at that time. Later, Huang et al [17] employed a feature-based classifier combined with Support Vector Machine (SVM) and Long Short-Term Memory (LSTM) to extract DDI, reaching the best performance at the time. Despite all the approaches mentioned above having made significant research contributions, there are still obvious limitations: manual feature extraction is inefficient and overly dependent on researcher expertise.

To cope with the challenges of traditional machine learning, deep learning techniques utilizing neural networks have emerged as a prominent research area in DDI extraction. In 2016, Liu et al [18] first achieved DDI extraction using a CNN model that transforms words into word vectors and combines them with location information as feature inputs, breaking the traditional method's over-reliance on domain expertise in the feature selection phase. Subsequently, Liu et al. [19] proposed a dependency-based CNN model to extract DDI, since the CNN model disregards sentence-level syntactic information as well as long-range word dependency. The model uses the structure of the Dependency Parsing Tree (DPT) that incorporates the syntax dependencies between two distant words in the text during model training. In addition to syntactic dependencies, the study of DDI extraction has gradually been extended to include techniques such as attention mechanisms [20-21], text augmentation [22-23], and so on. With the support of various deep learning techniques, the model performance has become increasingly powerful.

In 2018, Google launched BERT, which is a pre-trained language model that achieved the best performance in all 11 NLP tasks. As PLMs are becoming more prevalent, researchers are also focusing on how to extract DDI by fine-tuning PLMs. Soon after, specific language models in the biomedical domain were presented, such as BioBERT

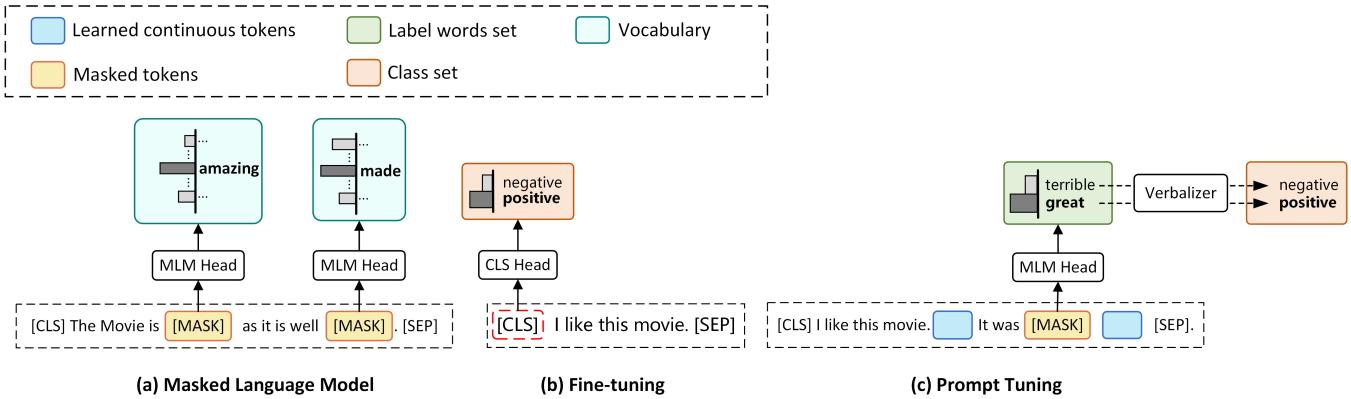


Fig. 2. The difference between the language masked model, ordinary fine-tuning and prompt tuning.

and BioGPT [24], which were trained on biomedical corpora and succeeded in DDI extraction tasks. However, with further study on the PLMs, researchers have found that there is a wide gap between the object form of pre-training and the downstream tasks, resulting in the ordinary fine-tuning not completely exploiting the power of the prior knowledge in the PLMs.

To address this issue, researchers [25-26] have proposed a novel prompt-tuning approach that aims to bridge the gap between the pre-training and the downstream tasks. Chapter III provides a detailed discussion of the differences between ordinary fine-tuning and prompt tuning.

Initially, many researchers tended to utilize manual prompts to direct pre-trained models toward downstream applications. Nevertheless, the experimental results have demonstrated that handcraft prompts are excessively reliant on large validation sets and are also inconsistent in performance. Sometimes, a change of a single word may lead to a drastic variance. Therefore, later researchers gradually started to focus on automatically constructing the prompt by employing continuous learnable parameters. On this basis, Liu et al. [27] proposed the P-Tuning model which achieved excellent outcomes in several natural language understanding tasks. Inspired by this, we make improvements to the construction of the prompts in P-Tuning. Considering the characteristics of the DDI extraction task, we further propose a DDI extraction model based on prompt tuning and data augmentation using contextual word embedding substitution, taking into account the data imbalance in the dataset.

## III. Background

The relation classification dataset can be represented as $D=\{X,Y\}$, where $X$ is the set of examples and $Y$ is the set of relation labels. For each example $x=\{w_1,w_{2,},...,w_s,...,w_o,...,w_n\}$, the goal of relation classification is to predict the relationship $y\in Y$ between the head entity $w_s$ and the tail entity $w_o$.

### A. Ordinary Fine-tuning on PLMs

In the previous fine-tuning process, it is necessary to convert the given example $x=\{w_1,w_{2,},...,w_n\}$ into an input sequence $\{[CLS],w_1,w_2,...,w_n,[SEP]\}$ with a special token. Subsequently, the given PLM $M$ encodes each token in the input sequence into corresponding hidden vectors $\{h_{[CLS]},h_1,h_2,...,h_n,h_{[SEP]}\}$. For the downstream classification task, a special task head "[CLS]" is utilized to compute the probability distribution over the class set $Y$ with the softmax function $p(\cdot|x)=Softmax(Wh_{[CLS]}+b)$, where $h_{[CLS]}$ is the hidden vector of "[CLS]", $W$ is a learnable matrix randomly initialized and $b$ is a learnable bias vector. The aim of fine-tuning PLM $M$ is to optimize the cross-entropy loss of the model.

### B. Prompt tuning for PLMs

Prompt tuning aims to bridge the gap between pre-training and the downstream task. The key is to construct a proper template $T(\cdot)$ and label words $V$. For an input sequence $x$, we first utilize the template to map $x$ to the prompt input $x_{prompt}=T(x)$. The template must determine the position of the $x$ and the number of special tokens inserted. The $V$ denotes the set of label words in the PLM $M$ and the role of the mapper $\psi:Y\to V$ is to correspond the class labels in $Y$ to the label words in $V$. In addition to retaining the original token in $x$, we also require at least one "[MASK]" to be added to $T(x)$ for the $M$ to predict the label word there. When PLM $M$ correctly predicts the masked position, we can depict the probability distribution of the masked token in $V$, that is, $p([MASK]=\psi(y)\,|\,x_{prompt})$.

Taking a simple binary sentiment classification task as an example, we can define the template $T(\cdot)=$"$\cdot$ *It was* $[MASK]$" and then map an input sequence $x$ to this template $T(x)=$"$x$ *It was* $[MASK]$". Next, utilizing the PLM $M$ to encode $x_{prompt}$, we can obtain the hidden vector of "[MASK]" and then compute a probability distribution $p([MASK]=\psi(y)\,|\,x_{prompt})$, which captures the most appropriate word from the set of label words $V$ to replace "[MASK]". Here, we ignore searching label words and just assign positive and negative sentiment labels, such as $\psi:(positive)\to$"*great*" or $\psi:(negative)\to$"*terrible*". In summary, only depending on the prediction of PLM $M$ for the masked word, we can accurately distinguish that the emotion expressed by instance $x$ is either "positive" or "negative".

## IV. Methodology

In this chapter, we present the details of the PTDA model. Sections A and B describe the relevant datasets and the corresponding data preprocessing operation. Section C shows how to apply contextual word embedding substitution to generate samples, mitigating the adverse effects of the data imbalance. The specific training program and optimization strategy are illustrated in sections D and F.

### *A. Related Dataset*

#### *1) DDI 2013 Dataset*

The DDI 2013 dataset consists of a manually annotated corpus of single sentences, mainly from 792 texts in DrugBank and 233 abstracts in Medline, containing 18,520 pharmacological substances and 5,028 DDIs [28].

Pharmacological substances in the dataset are divided into the following four categories: drug (generics), brand (trade drug), group (drug classes) and drug-N (unapproved active substances for human consumption). The drug-drug interactions in each sample are derived from the following four types: Advise (describing the DDI by recommendation or suggesting), Effect (describing the consequence of the pharmacological substance interaction), Mechanism (describing the way the interaction occurs), Int (not conveying any information about the DDI). In addition to this, a fake type "Negative" is also provided for neutral sentences.

The specific instance of each type is the following:

- *Advise*: These increases should be considered when selecting an oral **contraceptive** for a woman taking **atorvastatin**.
- *Effect*: Only **ibogaine** can enhance **cocaine**-induced increase in accumulating dopamine.
- *Mechanism*: Antacids increase the rate of absorption of **pseudoephedrine**, while **kaolin** decreases it.
- *Int*: Therefore, **linezolid** has the potential for interaction with **adrenergic** and serotonergic agents.
- *Negative*: Treatment of **toxin A** with [(14) C]-**diethyl** revealed concentration-dependent labeling of histidine residues on the toxin molecules.

The statistical results for the DDI 2013 dataset are presented in Table I. As the dataset contains only training and test datasets, we will randomly allocate 10% of the data from the training data as the validation dataset for model parameter adjustment. The statistics of the DDI 2013 dataset reveal a serious data imbalance, with a considerably larger number of negative examples than positive instances. Therefore, we decided to apply down-sampling on negative instances, following the previous work of Hong et al. [29] and Sahu et al. [30]. The down-sampling eliminates numerous negative instances and a few positive examples from the dataset. The statistics of the dataset after negative sample filtering are shown in Table II.

TABLE I
ORIGINAL STATISTICS OF DDI 2013 DATASET.

| Instances | Type | #Train | #Test |
|---|---|---|---|
| Positive | Advise | 826 | 221 |
| | Effect | 1,687 | 360 |
| | Mechanism | 1,319 | 302 |
| | Int | 188 | 96 |
| Negative | | 23,665 | 4,712 |
| Total | | 27,685 | 5,691 |

TABLE II
STATISTICS OF DDI 2013 DATASET WITH NEGATIVE SAMPLES FILTERING.

| Instances | Type | #Train | #Test |
|---|---|---|---|
| Positive | Advise | 814 | 221 |
| | Effect | 1,592 | 357 |
| | Mechanism | 1,260 | 301 |
| | Int | 188 | 92 |
| Negative | | 8,987 | 2,049 |
| Total | | 12,841 | 3,020 |

#### *2) ChemProt Dataset*

The ChemProt dataset comprises PubMed abstracts that annotate the interactions between chemical and protein entities [31]. In this study, we follow recommendations from the dataset developers to focus on classifying five high-level interactions: UPREGULATOR (CPR-3), DOWNREGULATOR (CPR-4), AGONIST (CPR-5), ANTAGONIST (CPR-6) and SUBSTRATE (CPR-9). The remaining types are marked FALSE and Table III provides statistics of the ChemProt dataset.

The following are the examples of the five types:

- *CPR-3*: Mutation of arginine 228 to **lysine** enhances the **glucosyltransferase** activity of bovine 1,4-galactosyl I.
- *CPR-4*: **PKC** isoforms did show different sensitivity and selectivity for down-regulation by I3A and phorbol 12-myristate 13-acetate (**PMA**) in Colo-205 cells.
- *CPR-5*: The selective **beta1AR** antagonists' atenolol and metoprolol blocked **isoproterenol**-induce enhancement with apparent K(b) values of 85 +/- 36.
- *CPR-6*: **Terfenadine** and astemizole are chemically unrelated to **histamine H1-receptor** antagonists.
- *CPR-9*: Total vitamin B6 is abnormally high in autism, consistent with previous reports of an impaired **pyridoxal-kinase** for the conversion of **pyridoxine** and pyridoxal to PLP.
- *FALSE*: Discovery and optimization of **Fc11a-2** as inhibitors of **NLRP3**: a structural basis for the reduction of albumin binding.

TABLE III
STATISTICS OF CHEMPROT DATASET.

| Instances | Type | #Train | #Dev | #Test |
|---|---|---|---|---|
| TRUE | CPR-3 | 756 | 546 | 663 |
| | CPR-4 | 2,227 | 1,091 | 1,655 |
| | CPR-5 | 173 | 115 | 178 |
| | CPR-6 | 229 | 199 | 292 |
| | CPR-9 | 727 | 457 | 642 |
| FALSE | | 13,923 | 8,860 | 12,315 |
| Total | | 18,035 | 11,268 | 15,745 |

*3) DTIs Dataset*

The DTIs dataset was constructed by Hong et al. [29]. It contains over 480k instances from nearly 20 million PubMed abstracts. Sentence labels are selected after aligning drug-target pairs based on DTIs facts in DrugBank. Sentence labels are divided into six types: substrate (the target acts upon the drug), inhibitor (a drug that binds to the target and impedes its function), agonist/antagonist (the drug that binds to the target and activates or blocks its biological response), unknown (the interaction of drug–target pair is exited, but the action mechanism is not reported in DrugBank), other (all the other types of interactions with fewer occurrences), and the fake type "NA".

Here are examples of the five types:

- *Substrate*: The objective of this study was to investigate the safety, pharmacokinetics and pharmacodynamics of **umeclidinium** in patients with normal and deficient **CYP2D6** metabolism.
- *Inhibitor*: Furthermore, sulfasalazine was found to be a potent inhibitor of **PCFT**, suggesting that it is a risk factor that would cause malabsorption of folate and also **MTX** when co-administered in the treatment of rheumatoid arthritis.
- *Agonist*: **Bromocriptine** was ten times more potent and pramipexole and ropinirole were ten times less potent at the dopamine D2 than at the **dopamine D3 receptor**, whereas pergolide was equipotent at the two receptors.
- *Unknown*: Expression of each cDNA individually yielded no detectable **prenyltransferase** activity; however, co-expression of the two together produced functional **geranyl-diphosphate** synthase.
- *Other*: Homeric studies showed that the COOH-terminal group of transmembrane helices (TMs), especially TM17, is responsible for the specificity of **nicorandil** for channels containing **SUR2**.
- *NA*: Surface tension measurements suggest that the mean time to minimum surface tension and the minimum surface tension were greater in **BAL** from mice exposed to **MMC** for 4 days.

To compare the performance of PTDA with the existing methods, the proportion of data in the training dataset, validation dataset, and test dataset is kept constant. Table IV shows the statistics of DTIs.

TABLE IV
STATISTICS OF DTIS DATASET.

| Instances | Type | #Train | #Test |
|---|---|---|---|
| Positive | Substrate | 1,810 | 18 |
| | Inhibitor | 2,622 | 26 |
| | Agonist | 925 | 11 |
| | Unknown | 2,835 | 28 |
| | Other | 616 | 10 |
| Negative | | 464K | 4,733 |
| Total | | 473K | 4,826 |

*B. Data Preprocessing*

This section illustrates our method for data preprocessing. The DDI 2013 dataset comprises instances that include a drug pair (e1, e2) and the corresponding interaction relation $R$. To collect more accurate information about the location and type of the entity, a pair of special tokens "<e1i>" and "</e1i>" is inserted around the first entity, where "1" denotes the first drug and "i" denotes the index of drug type. The correspondence between the index and the type of pharmacological substance is {1: drug, 2: brand, 3: group, 4: drug-N}. Similarly, a pair of special tokens "<e2i>" and "</e2i>" is added at the boundary of the second entity.

In addition, we further substitute "DRUG1" and "DRUG2" for the entities, which can enhance the model's generalization and avoid the influence of various drug word embedding. Take "**ZEBETA** should not be combined with other **beta-blocking-agents**." as an example, where "ZEBETA" and "beta-blocking-agents" are the target entity pairs, which after the above data preprocessing will become:

"<e11> DRUG1 </e11> should not be combined with other <e22> DRUG2 </e22>."

Since the ChemProt dataset focuses on the interactions between drugs and proteins, we replace specific drug and protein names with "DRUG" and "PROTEIN" for entity substitution. For example:

"Further, <e1> DRUG </e1> pretreatment blocked <e2> PROTEIN </e2>-induced increase in permeability of mouse-lung microvessels."

*C. Text Data Augmentation*

When dealing with classification problems, it is common to encounter issues with class imbalance and insufficient training data in a dataset. Taking the DDI 2013 dataset as an instance, we observe that filtering the negative samples in the dataset does not eliminate the adverse effects of the above challenges. Table II shows that the filtered dataset still contains approximately twice as many negative examples as positive instances. Simultaneously, a wide gap in numbers appears between positive types. As an illustration, the number of "Int" types is only 188, substantially less than the size of "Effect" and "Mechanisms", which directly leads to the trained model being much more sensitive to other types than to "Int" type. In other words, the probability of misidentifying "Int" rises dramatically.

In previous work, Papanikolaou et al [7] proposed utilizing GPT-2 to generate samples to address the challenges of class imbalance and insufficient training data. The method involved fine-tuning the CPT-2 model on a subset of different types and then using the fine-tuned model to generate numerous examples of that class. Experimental results have demonstrated that this approach can partially mitigate the negative effects. However, the low quality of the generated instances results in a significant amount of noise accumulating in the dataset as the number of generated samples continuously grows, leading to a sharp decline in model performance.

In this paper, we decided to abandon the generative models in favor of a pre-trained PubMedBERT, a language model

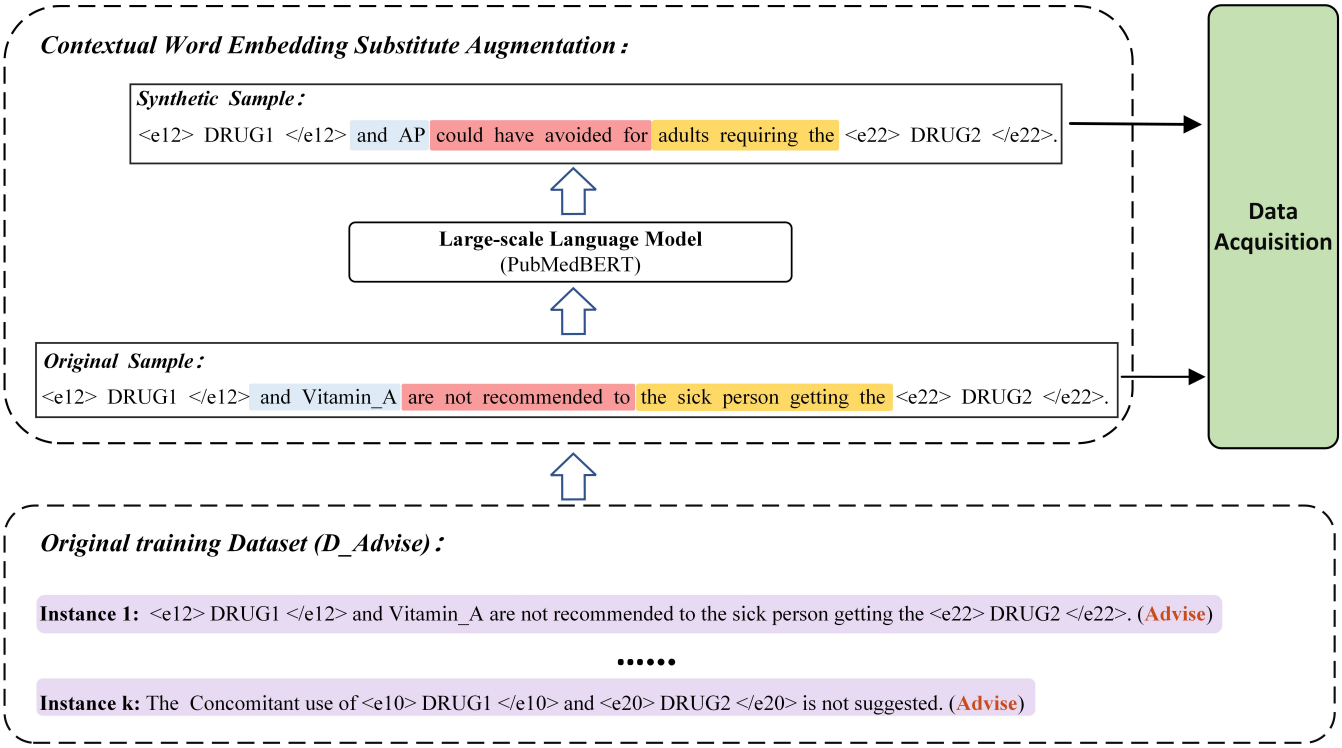


Fig. 3. An illustration of contextual word embedding augmentation. The Large-scale Language Model will replace the non-target entity part of the original instance to synthesize a new instance.

widely used in the biomedical domain, to synthesize the samples. All synthetic samples are constructed from the original sentences using contextual word embedding substitution, and the process is shown in Figure 3.

The following are the specific steps of text data augmentation:

- ***Step 1***: To synthesize training data, we split the training dataset $D$ into $n$ subsets, where each subset $D_n$ only contains examples belonging to the relation type $R_n$.
- ***Step 2***: After introducing the data augmentation library "nlpaug" [32], we choose PubMedBERT as the underlying access model for "nlpaug" to provide biomedical contextual information.
- ***Step 3***: We select an instance from the subset $D_n$ and fix the target entity in the original sentence. With the support of "nlpaug", it is easy to synthesize a new sample belonging to the relation type $R_n$ by substituting words besides the fixed target entities based on contextual information provided by PubMedBERT. The detailed process is shown in Figure 3.
- ***Step 4***: All generated samples are added to a new dataset $Dsynth$, which together with the original dataset $D$ serves as the training data.

The majority of the instances in the DDI 2013, ChemProt, and DTIs datasets are sourced from the PubMed corpus. Meanwhile, PubMedBERT was pre-trained on the PubMed corpus, making it better suited for providing contextual information than other pre-trained language models when synthesizing samples.

Table V displays the difference between the old sentence and the corresponding generated sample. Despite significant changes in words and sentence structures, the two remain remarkably similar in terms of semantic expression. This demonstrates that the new samples synthesized by using the contextual word embedding substitution, are successful in accurately conveying the original relational information between target entities.

In addition to the method of synthesizing samples, it is also crucial to consider the size of the generated data. Papanikolaou et al. [7] argue that the generated number of goal relation type $R_n$ should be equal to the size of the corresponding subset $D_n$ multiplied by the ratio $r$, i.e., $|Dsynth_n| = |D_n| * r$. From the experimental results, this approach can be of some purpose. However, in practice, maintaining the same ratio $r$ for different relation types can further worsen the effect of class imbalance, particularly when the value $r$ is large, due to the wide gap in the size of various relation types. To clarify this issue, we continue to treat the DDI 2013 dataset as an example.

For instance, there are 1,687 examples of the "Effect" type, but only 188 instances of the "Int" type. In this case, if we simply employ the same ratio (assuming $r = 1.0$) to generate data, it will further widen the gap between the number of "Effect" and "Int". This could result in the trained model being less sensitive to "Int" compared to the other types.

To address this challenge, we propose to limit the number of instances generated from different relation types by considering the ratio between the subset $D_n$ size and the smallest subset $D_{min}$ size. To be specific, we take the subset $|D_{min}|$ as the criterion, the generated sample number of one relation type is calculated as the product of the size of the subset $D_n$ and ratio $r$, multiplied by the ratio of the

TABLE V
COMPARISON BETWEEN ORIGINAL SENTENCES AND CORRESPONDING GENERATED INSTANCES.

| Dataset (Relation Type) | Original sentences/Generated sentences |
|---|---|
| DDI 2013 (Advise) | <e12> DRUG1 </e12> and Vitamin_A are not recommended to the sick person getting the <e22> DRUG2 </e22>. ***(Original)*** |
| | <e12> DRUG1 </e12> and AP could have avoided for adults requiring <e22> DRUG2 </e22>. ***(New)*** |
| ChemProt (CPR-3) | Mutation of arginine 228 to <e1> DRUG </e1> enhances the <e2> PROTEIN </e2> activity of bovine beta-1,4-galactosyltransferase I. ***(Original)*** |
| | Within coding sequence encoding <e1> DRUG </e1> improves its <e2> PROTEIN </e2> performance under aether-1,4-galactosyltransferase system. ***(New)*** |
| DTIs (Substrate) | The objective of this study was to investigate the safety, pharmacokinetics and pharmacodynamics of <e1> DRUG1 </e1> in patients with normal and deficient <e2> DRUG2 </e2> metabolism. ***(Original)*** |
| | In order to better study and investigate <e1> DRUG1 </e1> the chemical, physical properties of <e2> DRUG2 </e2> in different individuals. ***(New)*** |

number $D_{min}$ to $D_n$. Refer to Equation (1) for the specific formula.

$$|Dsynth_n| = |D_n| * r * \frac{|D_{\min}|}{|D_n|} = |D_{\min}| * r \quad (1)$$

From Equation (1), we can see that the way to calculate the number of new instances from each relation type is equal to the size of the smallest subset $|D_{min}|$ multiplied by the ratio $r$. Therefore, the sum of the generated data in a dataset is equal to:

$$|Dsynth| = |D_{\min}| * r * n \quad (2)$$

Now, we summarize the operational procedure of text data augmentation in Algorithm 1 to assist the reader in comprehending the entire process more clearly.

**Algorithm 1:** Text Data Augmentation

**Input:** dataset $D$, relation set $L$, ratio $r$;
**Output:** generated dataset $Dsynth$

**for** each DDI type $R_n \in L$ **do**
  subset $D_n = \{s \mid RE\_type(s) = R_n\}$;
  calculate generated data size $|Dsynth_n| = |D_{min}| * r$;
  set counter $i = 0$;
  **while** $i < |Dsynth_n|$ **do**
    Randomly select sentence $s \in D_n$;
    Use contextual word embedding substitution for $s$ to generate $s'$;
    $Dsynth_n$ append $s'$;
    $i++$;
  **end for**
  Obtain the $Dsynth = Dsynth_1 \cup ... \cup Dsynth_n$;
  **return** $Dsynth$;

### D. Constructing Prompt Template

In this section, we illustrate how to employ improved P-tuning in DDI extraction tasks. Given PLM $M$ and a sequence of discrete input tokens $x = \{w_1, w_2, ..., w_n\}$ with corresponding labels $y \in Y$, the role of the prompt $P$ is to organize the text input $x$, the label $y$ and itself into a template $T(x)$. Taking the sentence "<e11> DRUG1 </e11> should not be combined with other <e22> DRUG2 </e22>." as the example, a traditional discrete template in the DDI classification task is illustrated to:

$$T(x) = \{x; [P_{0:n}]; [MASK]\} = "x \text{ the relation of } DRUG1 \text{ and } DRUG2 \text{ is } [MASK]" \quad (3)$$

Where $x$ is the original input sequence, $[P_{0:n}]$ is prompt and "[MASK]" is the predicted target. The flexibility of prompt structure and location allows us to employ our linguistic intuition to design prompts and insert them anywhere in the sentence. However, traditional handcrafted prompt heavily relies on manual labor and their performance is also extremely unstable. Sometimes, a small change in word or token's location for prompt can cause a drastic variety in model performance.

Therefore, we have adopted the idea of P-tuning to design the prompts. Compared to traditional discrete prompts, we choose vector representations in continuous space as prompts and add some task-related anchor tokens (such as "drug1" and "drug2" in Figure 4), which convert manual prompt construction into continuous parameter optimization. The specific flow is shown in Figure 2. To begin with, we design the prompt template $T(x)$ with the details given in Equation (4):

$$T(x) = \{x; [P_{0:n}]; DRUG1; [MASK]; DRUG2; [P_{i+4:n}]\} \quad (4)$$

Where "drug1" and "drug2" are artificially inserted anchor tokens related to the task, facilitating the model to better understand the task itself. In contrast to traditional discrete prompts, where $[P_i]$ is just a static token in $V$ that refers to the vocabulary of a pre-trained language model $M$, but we view $[P_i]$ as a pseudo-token and apply the Prompt-Encoder to map the template $T(x)$ into:

$$e(T(x)) = \{e(x), h_0 ... h_i, e(DRUG1), e([MASK]), e(DRUG2), h_{i+4} ... h_n\} \quad (5)$$

where $h_i (0 < i < n)$ are trainable embedding tensors. This enables us to search for better continuous prompts with the assistance of the downstream loss function $Loss$, which may extend beyond the expression in the original vocab $V$ of the PLM $M$. Equation (6) illustrates the process of optimizing the continuous prompts $h_i (0 < i < n)$.

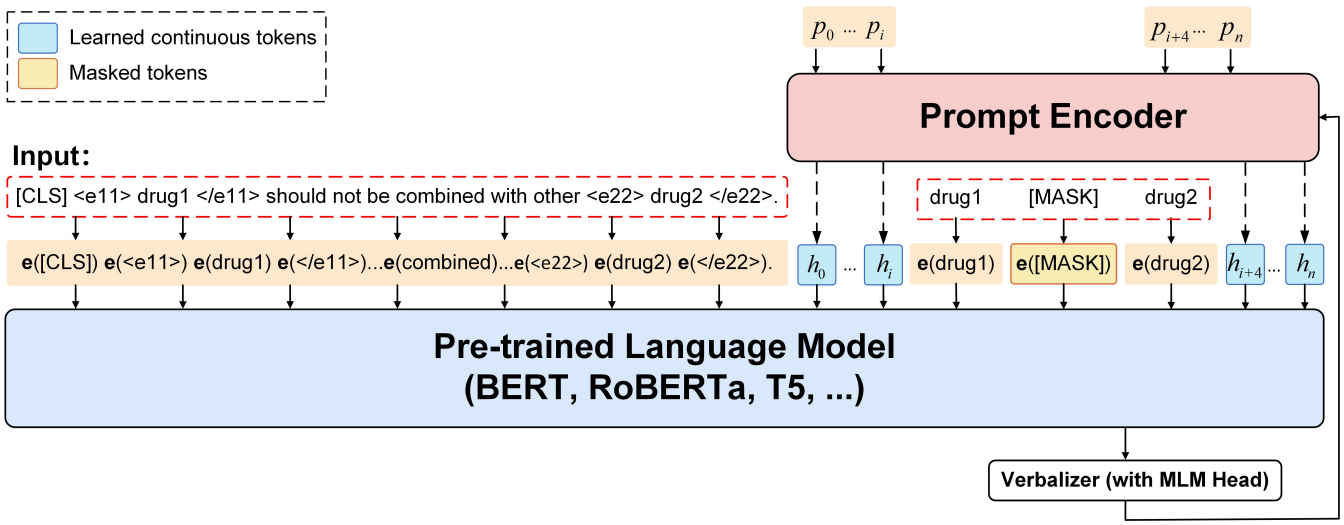


Fig. 4. The specific processes of predicting the interaction between "DRUG1" and "DRUG2" by prompt tuning.

$$\hat{h}_{0:n} = argmin\ Loss(M(x, y)) \tag{6}$$

Additionally, to associate the prompts embeddings $h_i (0 < i < n)$ with each other rather than an independent vector, we choose a bidirectional long-short-term memory network (Bi-LSTM) with ReLU-activated two-layer multilayer perceptron (MLP) to minimize their discreteness. Formally, the input embeddings $h_i^{'} (0 < i < n)$ to the PLM $M$ are derived from:

$$h_i^{'} = MLP(Bi - LSTM(P_{0:n})) \tag{7}$$

### E. Optimization Strategy

The Masked Language Model (MLM) is frequently used during the pre-training phase. This involves randomly masking some tokens in the input sequence and then allowing the model to predict the corresponding "[MASK]" section. The goal of the training is to enable the model to gradually comprehend natural language by repeating the above process. Please refer to Figure 1(a) for a detailed description of the entire process.

Prompt tuning shares similarities with MLM, we usually employ a mapping function $\psi : Y \to V$ to bridge the set of classes and set of label words during the training. In some studies, researchers have also called the function $\psi$ as a "verbalizer". With the mapping function $\psi$, we can formalize the probability distribution over $Y$ with the probability distribution over $V$ at the masked position:

$$p(y \mid x) = p(\psi(y) \mid T(x)) \tag{8}$$

The final class label can be determined by predicting the masked word and mapping it to the class set using a verbalizer $\psi$. Subsequently, we choose the Cross-Entropy as a loss function to optimize the objective, which is depicted in Equation (9):

$$J_{[MASK]} = -\frac{1}{|X|} \sum_{x \in X} y \log p(y \mid x) \tag{9}$$

## V. Experiments

In this chapter, we conducted a series of experiments on three standard relation extraction datasets.

### A. Metrics

Same as the previous research, we evaluate the model performance with *Precision*, *Recall* and *F1-micro*. The calculations can be obtained from Equations (10) - (12):

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 - micro = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{12}$$

where *TP* denotes the number of samples correctly predicted in positive cases, *FP* denotes the number of samples incorrectly predicted in negative cases, *FN* denotes the number of samples incorrectly predicted in positive cases and *F1-micro* is used for a comprehensive evaluation of *Precision* and *Recall* in the classification task.

### B. Experimental Settings

We benchmarked PTDA on three datasets, employing SciFive-Large [33] as the PLM for prompt tuning. In the experiments, the majority of the hyperparameters are kept consistent with the previous work. The significant parameter values can be obtained in Table VI.

TABLE VI
Hyperparameter Settings.

| Parameters | DDI 2013 | ChemProt | DTIs |
|---|---|---|---|
| Epochs | 12 | 18 | 25 |
| Batch size | 16 | 24 | 32 |
| Learning rate | 3e-5 | 2e-5 | 2e-4 |
| Optimizer | Adam | Adam | Adam |
| Dropout | 0.3 | 0.3 | 0.4 |
| Generate data ratio | 2 | 3 | 5 |
| Max length | 256 | 256 | 256 |

TABLE VII
COMPARISON WITH LITERATURE RESULTS ON DDI 2013 DATASET. "-" INDICATES EXPERIMENTAL RESULTS ARE NOT PUBLISHED WITH SOURCE PAPER AND THE REST ARE FROM THE ORIGINAL PAPER. THE BEST RESULT ARE BOLD.

| | Methods | *Pre (%)* | *Rec (%)* | *F1 (%)* |
|---|---|---|---|---|
| Based on CNN/RNN | SCNN [34] | 69.1 | 65.1 | 67.0 |
| | Joint AB-LSTM [30] | 73.4 | 69.7 | 71.5 |
| | Position-aware LSTM [35] | 75.8 | 70.4 | 73.0 |
| | Atten Bi-LSTM [36] | 78.4 | 76.2 | 77.3 |
| Based on Fine-tuning | DARE [7] | 82.0 | 74.0 | 78.0 |
| | ELECTRA-Med [37] | 80.1 | 78.2 | 79.1 |
| | Character-BERT [38] | - | - | 80.4 |
| | Multiple-entity-Att-BioBERT [39] | 81.0 | 80.9 | 80.9 |
| Based on Prompt tuning | **PTDA** | **85.7** | **84.1** | **84.9** |

TABLE VIII
COMPARISON WITH LITERATURE RESULTS ON CHEMPROT DATASET. "-" INDICATES EXPERIMENTAL RESULTS ARE NOT PUBLISHED WITH SOURCE PAPER AND THE REST ARE FROM THE ORIGINAL PAPER. THE BEST RESULT ARE BOLD.

| | Methods | *Pre (%)* | *Rec (%)* | *F1 (%)* |
|---|---|---|---|---|
| Based on CNN/RNN | SVM + Deep-Learning [40] | 72.7 | 57.4 | 64.1 |
| | Atten-RNN [41] | 65.4 | 64.8 | 65.2 |
| | Deep-Word-Representation [42] | 67.0 | 72.0 | 69.4 |
| | ELECTRA-Med [38] | 75.5 | 70.7 | 72.9 |
| Based on Fine-tuning | DARE [7] | 79.0 | 68.0 | 73.0 |
| | NCBI_BERT [43] | 73.4 | 75.5 | 74.4 |
| | BioBERT [6] | 77.0 | 75.9 | 76.5 |
| | KeBioLM [44] | - | - | 77.5 |
| Based on Prompt tuning | **PTDA** | **82.3** | **77.4** | **79.8** |

TABLE IX
THE RESULTS ON DTIS DATASET. EXPERIMENTAL RESULTS ARE FROM THE ORIGINAL PAPER. THE BEST RESULT ARE BOLD.

| Methods | *F1 (%)* | *AUPRC (%)* |
|---|---|---|
| BERE-AVE [29] | 46.0 | 38.4 |
| BERE-POOL [29] | 57.9 | 51.7 |
| BERE [29] | 62.5 | 52.4 |
| EGFI [1] | 71.2 | 58.1 |
| **PTDA** | **72.1** | **60.3** |

### *C. Comparison with Baselines*

So far, many DDI extraction methods have been proposed. We selected some classic approaches as the baseline to compare with PTDA on both DDI 2013, ChemProt and DTIs datasets.

Table VII-IX presents the performances of PTDA and baselines for biomedical relation extraction on various datasets. We can observe that on the DDI 2013 dataset, the *Precision*, *Recall* and *F1-micro* of PTDA reach 85.7%, 84.1%, and 84.9% respectively, outperforming all baseline methods. Meanwhile, on the ChemProt dataset, the PTDA also obtains the best results compared to other baseline methods, with *Precision*, *Recall* and *F1-micro* of 82.3%, 77.4% and 79.8%, respectively. In addition to the DDIs 2013 and ChemProt dataset, we compared the PTDA with other baseline models on the DTIs dataset. Table IX shows the performances of PTDA and other baseline models. Compared with other exited methods, PTDA achieves the highest *F1-score* and *AUPRC*.

### *D. Effect of Generated Data Size*

This section aims to examine how varying amounts of generated data affect PTDA performance. From our detailed description of contextual word embedding substitution in Section IV.C, it is clear that all the contextual information used to synthesize the examples is derived from a priori knowledge that PubMedBERT gained during pre-training. In an ideal scenario, we can utilize contextual word embedding substitution to continuously generate data and expand the training dataset size.

However, this procedure is not flawless. The prior knowledge embedded in PubMedBERT is finite, and as the amount of generated data increases, the valuable contextual information gradually diminishes. This could result in the homogenization of the generated samples or even the introduction of noisy data. Consequently, the model may no longer be able to learn significant information from the training data.

Therefore, our objective is to determine the number of generated samples, in other words, to search for the optimal ratio $r$. In this process, if the value of $r$ is too small, the improvement in model performance will be insignificant. But, if the value of $r$ is excessively large, the model risks being affected by noise. To gain empirical insight into the above question, we designed a comparative experiment, which

TABLE X
COMPARATIVE RESULTS OF PRECISION, RECALL AND F1-MICRO ON DDI 2013 AND CHEMPROT DATASETS UTILIZING DIFFERENT PROMPT TEMPLATES. "[X]" INDICATES THE ORIGINAL INPUT, SUCH AS "<E11> DRUG1 </E11> SHOULD NOT BE COMBINED WITH OTHER <E22> DRUG2 </E22>.", "[P]" IS A CONTINUOUS LEARNABLE PARAMETER AND "[MASK]" REFERS TO THE PREDICTED TARGET. THE BEST RESULTS ARE BOLD AND THE SECOND BEST OVERCOMES ARE ADDED UNDERLINES.

| Template Type | Input Example | DDI 2013 | | | ChemProt | | | DTIs | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | *Pre (%)* | *Rec (%)* | *F1 (%)* | *Pre (%)* | *Rec (%)* | *F1 (%)* | *Pre (%)* | *Rec (%)* | *F1 (%)* |
| Manual-Template | [CLS] [X], the relation of *DRUG1* and *DRUG2* is [MASK] [SEP]. | 80.2 | 77.5 | 78.8 | 78.4 | 73.4 | 75.8 | 68.7 | 65.0 | 66.8 |
| | [CLS] [X], the *DRUG1* [MASK] the *DRUG2* [SEP]. | 78.5 | 83.5 | 80.9 | 76.8 | 77.4 | 77.1 | 70.1 | 66.4 | 68.2 |
| Soft-Template | [CLS] [X], [P] [P] [P] [P] [P] [P] [MASK] [SEP]. | 80.4 | 82.2 | 81.3 | 76.1 | **79.4** | 77.7 | 70.2 | 67.5 | 68.8 |
| | [CLS] [X], [P] [P] [P] [MASK] [P] [P] [P] [SEP]. | 83.4 | 81.6 | 82.5 | 81.9 | 74.8 | 78.2 | 68.1 | 70.8 | 69.4 |
| Mixed-Template | [CLS] [X], [P] [P] *DRUG1* and *DRUG2* [P] [P] [MASK] [SEP]. | 84.1 | 83.6 | 83.8 | 79.2 | 77.8 | 78.5 | **72.9** | 70.3 | 71.6 |
| | [CLS] [X], [P] [P] *DRUG1* [MASK] *DRUG2* [P] [P] [SEP]. | **85.7** | **84.1** | **84.9** | **82.3** | 77.4 | **79.8** | 69.8 | **76.7** | **73.1** |

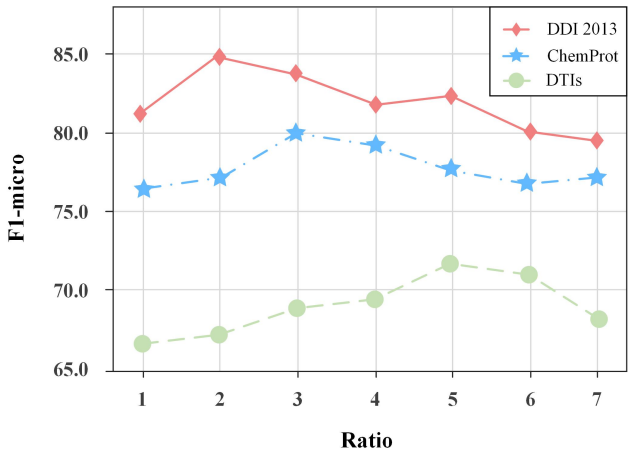


Fig. 5. F1-micro of PTDA on the DDI 2013, ChemProt and DTIs datasets with different ratios.

utilized the training data from the different datasets to construct a corresponding size of new samples according to different ratios $r$.

Figure 5 presents the test results for different ratios. It is obvious that the PTDA achieves the best performance on both the DDI 2013, ChemProt and DTIs datasets, when $r_{DDI\,2013} = 2$, $r_{ChemProt} = 3$ and $r_{DTIs} = 5$. The experimental result supports our assumption that additional data does not necessarily enhance the classifier performance, since the massive noisy data seems to cause adverse effects on model training.

### *E. Comparison of Prompt Template*

Template construction is the key to prompt tuning. An appropriate template can make better use of a priori knowledge embedded in the pre-trained language model and enhance the downstream task's performance.

At present, there are three mainstream template types: Manual-Template (the prompts are designed entirely by handcraft), Soft-Template (the prompts contain only continuous learnable parameters) and Mixed-Template (the prompts include both handcrafted tokens and continuous learnable parameters). To test the performance of different templates, we designed six templates based on the characteristics of each template type. The PTDA constructs input sequences for prompt tuning according to the following templates. The test results are listed in Table X.

We observe that the templates with continuous learnable parameters all outperform the handcraft templates. This phenomenon implies that the prompts with learnable parameters can automatically optimize the training process to better match the downstream tasks. Besides, we also find that adding a few anchor tokens to the prompt can help the model better understand the task. For instance, on the DDI extraction task, we add "DRUG1" and "DRUG2" as anchor tokens and insert the predicted "[MASK]" in between them. These operations make the model more biased to learn the association between the two drugs, thus reaching the purpose of classifying drug interactions.

### *F. Choice of Pre-trained Language Model*

Better performance in downstream tasks can be reached by fully exploiting the prior knowledge in the pre-trained model during prompt tuning.

There are some variations in the prior knowledge embedded in pre-trained language models, due to differences in the training method and training corpus. In order to seek the pre-training model that best meets the task requirements, we selected some mainstream pre-trained models in the biomedical domain for the experiments, such as SciBERT [45], BioBERT, PubMedRoBERTa [46], and SciFive. The comparison results are published in Table XI.

We observe that the "Large" version of each pre-trained model outperforms the corresponding "Base" version in performance. We believe that it is attributed to the greater number of trainable parameters in the "Large" version, which embeds more prior knowledge in the model. Secondly, it is confirmed that the SciFive exhibits the best overall performance among the four biomedical domain-specific models.

TABLE XI
COMPARISON RESULTS OF PRECISION, RECALL AND F1-MICRO FOR VARIOUS PRE-TRAINED MODELS IN THE BIOMEDICAL DOMAIN ON THE DDI 2013 DATASET AND CHEMPROT DATASET. THE BEST RESULTS ARE BOLD AND THE SECOND BEST ARE ADDED UNDERLINE.

| Dataset | Metrics | Base | | | Large | | | |
|---|---|---|---|---|---|---|---|---|
| | | BioBERT | PubMedRoBERTa | SciFive | SciBERT | BioBERT | PubMedRoBERTa | SciFive |
| DDI 2013 | *Pre (%)* | 80.9 | 83.9 | 83.8 | 81.7 | 84.3 | <u>85.6</u> | **85.7** |
| | *Rec (%)* | 83.5 | 81.7 | **84.6** | <u>84.5</u> | 80.2 | 81.9 | 84.1 |
| | *F1 (%)* | 82.2 | 82.8 | <u>84.2</u> | 83.1 | 82.2 | 83.7 | **84.9** |
| ChemProt | *Pre (%)* | 73.3 | 74.3 | 80.9 | **82.5** | 79.8 | 76.4 | <u>82.3</u> |
| | *Rec (%)* | 77.2 | <u>79.0</u> | 76.6 | 74.5 | 74.8 | **79.5** | 77.4 |
| | *F1 (%)* | 75.2 | 76.6 | <u>78.7</u> | 78.3 | 77.2 | 77.9 | **79.8** |
| DTIs | *Pre (%)* | 69.8 | 69.2 | 70.1 | 68.3 | <u>72.6</u> | **73.6** | 69.8 |
| | *Rec (%)* | 70.4 | 72.7 | 69.9 | <u>75.1</u> | 69.7 | 70.9 | **76.7** |
| | *F1 (%)* | 69.2 | 70.9 | <u>72.3</u> | 71.5 | 71.1 | 72.2 | **73.1** |

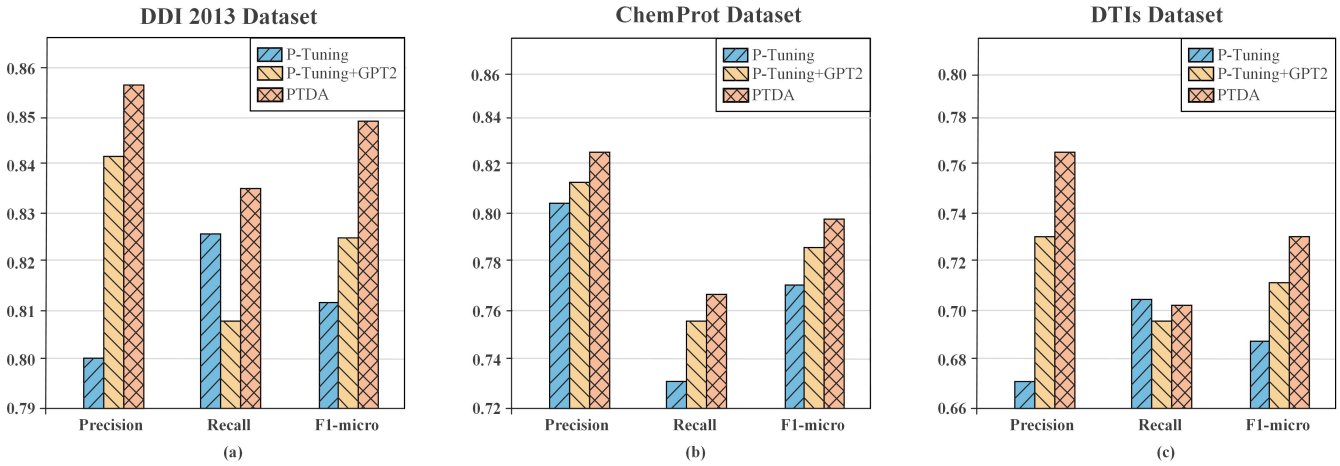


Fig. 6. The performances of different text augmentation methods on three datasets.

## G. Ablation Study on Text Data Augmentation

To better understand the specific contribution of contextual word embedding substitution to generate samples, we designed two experiments, single P-Tuning and P-Tuning + GPT-2, to compare with our proposed PTDA. Here, single P-Tuning means relying only on prompt tuning for the pre-trained language model to classify biomedical relations without any text data augmentation operations. The P-Tuning + GPT-2 suggests that we first generate massive training data using the GPT-2 and then classify DDI via prompt tuning. For generating data with GPT-2, we follow the work presented by Papanikolaou et al. [7]. We aim to address two issues through the above experiment. The first issue concerns the effect of text data augmentation on the model. The second is to detect the distinction between various approaches to generate data.

The experimental results are displayed in Figure 6. We observe that single P-Tuning performs the worst on both datasets, indicating that generated samples can alleviate the negative effects of data imbalance and boost model performance. Additionally, we notice that PTDA performs better than the P-Tuning + GPT-2, which points to the fact that generated data using contextual word embedding substitution is higher quality and more beneficial for model training than data generated using GPT-2.

In addition to evaluating the overall dataset, we also analyzed the alterations from each relation type before and after applying text data augmentation, to further explore the role of the contextual word embedding substitution. Taking the DDI 2013 dataset as an example, we recorded the evaluation results of the four positive types under different methods, the details are presented in Table XII.

By comparing these metrics, we observe that each type's score increased to some degree with data augmented technique. The "Int" type exhibited the most significant boost with a surprising 8.8% growth in F1-micro. The phenomenon can be attributed to the fact that the high-quality data constructed by employing contextual word embedding substitution greatly alleviates the severe lack of the "Int" type in the original dataset, thus boosting the model's ability to recognize the "Int" type. Simultaneously, it is also worth mentioning that the metric of the other three types displays a lesser growth with data augmentation compared to the "Int" type. With careful analysis, we concluded that if the size of one type in the dataset is relatively adequate, the benefit of utilizing data augmentation to generate instances to increase their number consistently will gradually diminish.

TABLE XII
COMPARISON RESULTS OF PRECISION, RECALL AND F1-MICRO FOR THE FOUR POSITIVE TYPES ON THE DDI 2013 DATASET IN THREE MODES: SIGNAL P-TUNING, P-TUNING + GPT-2 AND PTDA. THE BEST RESULT ARE BOLD.

| Enrich Strategy | Metrics | DDI 2013 dataset | | | |
|---|---|---|---|---|---|
| | | Advise | Effect | Mechanism | Int |
| P-Tuning (No Text Data Augmentation) | *Pre (%)* | 88.1 | 80.5 | 87.5 | 60.1 |
| | *Rec (%)* | 86.8 | 85.7 | 81.4 | 48.9 |
| | *F1 (%)* | 87.4 | 83.0 | 84.3 | 53.9 |
| P-Tuning + GPT-2 (No Contextual Word Embedding Substitution) | *Pre (%)* | **91.2** | 78.4 | 86.6 | **84.0** |
| | *Rec (%)* | 89.6 | 86.6 | **86.0** | 45.7 |
| | *F1 (%)* | 90.4 | 82.3 | 86.3 | 59.1 |
| PTDA | *Pre (%)* | 89.1 | **81.1** | **88.9** | 79.6 |
| | *Rec (%)* | **92.7** | **87.6** | 85.1 | **51.7** |
| | *F1(%)* | **90.9** | **84.2** | **86.9** | **62.7** |

### *H. Error Analysis*

In this section, we have made a similar effort to Zhu et al. [39], presenting a normalized confusion matrix to analyze classification results. The color density indicates the percentage of instances, so that we can see the percentage of misclassified instances. Our goal is to summarize the reasons for data misclassification by observing the final results. This will help us improve the model's performance in future studies. As an example, the DDI 2013 dataset's test results are still referenced, and Figure 7 shows the corresponding normalized confusion matrix.

According to the displayed results, we can roughly group the misclassified samples into three categories:

(1) Some samples in each positive type are incorrectly predicted as negative type. For this error, we believe that the main reason is too many negative samples in the training set. In our earlier work, we have utilized data augmentation techniques to mitigate the adverse effects of data imbalance to some extent. However, there are still substantially more negative samples than positive ones, which leads to the classifier being less sensitive to the positive samples after training, thus causing misclassification.

(2) A few samples in the negative type are also misclassified into the positive types. Through examining these negative samples, we discovered that they typically include one or more relation trigger words, similar to those appearing in positive classes. This similarity results in the model making incorrect predictions.

(3) So many "Int" type instances are wrongly classified as "Effect" type. The normalized confusion matrix shows that the PTDA model misclassifies approximately 40% of "Int" samples as "Effect" type. Upon careful comparison, we find that these examples contain some keywords that are similar to the "Effect" type, such as the sentence "conversely, <e10> diethylpropion </e10> may interfere with <e22> antihypertensive </e22> (i.e., guanethidine, a-methyldopa).", where the word "interfere" serves as a significant relation trigger in the "Effect" type. Consequently, the model may easily misclassify a similar sentence as the "Effect" type.

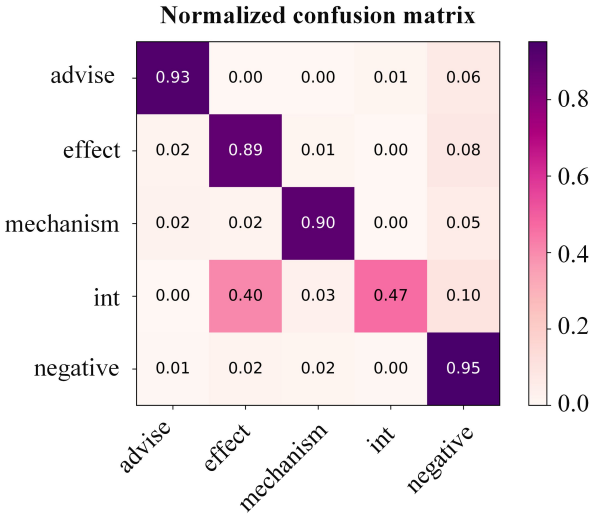


Fig. 7. The normalized confusion matrix for classification results.

## VI. CONCLUSIONS

In this paper, we present a novel approach (PTDA) for DDI extraction based on prompt tuning and text data augmentation techniques. The method can not only utilize contextual word embedding substitution to generate high-quality samples to mitigate the adverse effects of insufficient data or class imbalance, but also better exploit the role of prior knowledge in biomedical pre-trained language models. Test results on three relation classification datasets demonstrate that PTDA significantly outperforms the baselines without manual annotation and the introduction of extra knowledge.

In the future, we plan to introduce some external knowledge as a supplement to our current work, such as the molecular structure of drugs, artificial annotation of drug action, and so on. We hope to further boost the performance of the model by incorporating additional knowledge.

## REFERENCES

[1] L. Huang, J. Lin, X. T. Li, L. Q. Song and Z. T. Zheng, "EGFI: drug-drug interaction extraction and generation with fusion of enriched entity and sentence information," *Briefings in Bioinformatics*, vol. 23, no.1, pp. 153-167, 2022.

[2] N. WariKoo, Y. C. Chang and W. L. Hsu, "LBERT: Lexically aware Transformer-based Bidirectional Encoder Representation model for learning universal bio-entity relations," *Briefings in Bioinformatics*, vol. 37, no. 3, pp. 404-412, 2021.

[3] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language

Understanding," in *Proc. NAACL*, Minnesota, USA, pp. 4171-4186, 2019.
[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, pp. 9-17, 2019.
[5] Y. Gu, R. Tinn, H. Chen, M. Lucas and N. Usuyama, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1-23, 2021.
[6] J. Lee, W. Yoon, S. Kim, D. Kim and S. Kim, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234-1240, 2020.
[7] Y. Papanikolaou and A. Pierleoni, "DARE: Data Augmented Relation Extraction with GPT-2," *arXiv 2020*, arXiv:2004.13845, 2020.
[8] Z. Lin, D. Yang, H. Jiang, and H. Yin, "Learning Patient Similarity via Heterogeneous Medical Knowledge Graph Embedding", *IAENG International Journal of Computer Science*, vol. 48, no. 4, pp.868-877, 2021.
[9] T. G. Soares, A. Azhari, N. Rokhman and E. Winarko, "Education Question Answering Systems: A Survey", *Lecture Notes in Engineering and Computer Science: Proceedings of The International Multi-Conference of Engineers and Computer Scientists 2021*, Hong Kong, pp.24-34, 2021.
[10] G. L. He, C. Y. Chi and Y. Y. Zhan, "Combining N-gram Statistical Model with Pre-trained Model to Correct Chinese Sentence Error", *Engineering Letters*, vol. 30, no. 2, pp.476-484, 2022.
[11] X. Zhang, G. Yu, J. Shang, and B. Zhang, "Short-term Traffic Flow Prediction with Residual Graph Attention Network", *Engineering Letters*, vol. 30, no. 4, pp.1230-1236, 2022.
[12] Y. Wang, X. Cheng, and X. Meng, "Sentiment Analysis with An Integrated Model of BERT and Bi-LSTM Based on Mult-Head Attention Mechanism", *IAENG International Journal of Computer Science*, vol. 50, no. 1, pp.255-262, 2023.
[13] X. Yu, Z. Li, J Wu, and M. Liu, "Multi-module Fusion Relevance Attention Network for Multi-label Text Classification", *Engineering Letters*, vol. 30, no. 4, pp.1237-1245, 2022.
[14] M. Ueda, Y. Matsunami, P. Siriaraya, and S. Nakajima, "Developing Evaluation Expression Dictionaries for the Cosmetic Review Recommendation", *Lecture Notes in Engineering and Computer Science: Proceedings of The International Multi-Conference of Engineers and Computer Scientists 2019*, Hong Kong, pp.236-241, 2019.
[15] A. A. Syed, Y. H. Lukas, and A. Wibowo, "A Comparision of Machine Learning Classifiers on Laptop Products Classification Task", *Lecture Notes in Engineering and Computer Science: Proceedings of The International Multi-Conference of Engineers and Computer Scientists 2021*, Hong Kong, pp.104-110, 2021.
[16] D. Klein and C. D. Manning, "Accurate Unlexicalized Parsing," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 7-12 July, 2003, Sapporo, Japan, pp. 423-430.
[17] D. G. Huang, Z. C. Jiang, L. Zou and L. S. Li, "Drug–drug interaction extraction from biomedical literature using support vector machine and long short-term memory networks," *Information Sciences*, vol. 415-416, pp. 100-109, 2017.
[18] S. Y. Liu, B. Z Tang, Q. C. Chen and X. L. Wang, "Drug-Drug Interaction Extraction via Convolutional Neural Networks," *Computational & Mathematical Methods in Medicine*, vol. 2016: 6918381, pp. 1-8, 2016.
[19] S. Y. Liu, K. Chen, Q. C. Chen and B. Z. Tang, "Dependency-based convolutional neural network for drug-drug interaction extraction," *In Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 15-18 December, 2016, Shenzhen, China, pp. 1074-1080.
[20] K. Wang, X. Fu, Y. Liu, W. Chen, and J. Chen, "Att-FMI: A Fusing Multi-Information Model with Self-Attentive Strategy for Relation Extraction," *IAENG International Journal of Applied Mathematics*, vol. 53, no.3, pp961-971, 2023.
[21] P. Zhou, W. Shi, J. Tian, Z. Y. Qi and B. C. Li, "Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification," *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 7–12 August, 2016, Berlin, Germany, pp. 207-212.
[22] J. Wei and K. Zhou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," *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)*, 3-7 November, 2019, Hong Kong, China, pp. 6382-6388.
[23] S. Kobayashi, "Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations," *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1-6 June 2018, Louisiana, USA, pp. 452-457.
[24] R. Luo, L. Sun, Y. Xia, T, Qin and S. Zheng, "BioGPT: generative pre-trained transformer for biomedical text generation and mining." *Briefings in Bioinformatics*, vol. 23, no. 6, pp. 409-423, 2022.
[25] T. Schick and H. Schütze, "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference," *In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 21-23 April 2021, Online, pp. 255-269.
[26] J. Han, S. Zhao, B. Cheng, S, Ma and W. Lu, "Generative Prompt Tuning for Relation Classification," *In Findings of the Association for Computational Linguistics: EMNLP 2022*, 5-9 December 2022, Abu Dhabi, United Arab Emirates, pp. 3170-3185.
[27] X. Liu, Y. Zheng, Z. Du, M. Ding and Y. Qian, "GPT Understands, Too," *arXiv 2021*, arXiv:2103.10385.
[28] I. S. Bedmar, P. Martínez and M. H. Zazo, "SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDI Extraction 2013)," I*n Second Joint Conference on Lexical and Computational Semantics (*SEM)*, 13-14 June 2013, Atlanta, Georgia, USA, pp. 341-350.
[29] L. Hong, J. Lin, S. Li, F. Wan and H. Yang, "A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories," *Nature Machine Intelligence*, vol 2, pp. 347-355, 2020.
[30] S. K. Sahu and A. Anand, "Drug-drug interaction extraction from biomedical texts using long short-term memory network," *Journal of Biomedical Informatics*, vol. 86, no. 1, pp. 15-24, 2018.
[31] M. Krallinger, O. Rabal and S. A. Akhondi, "Overview of the Bio-Creative VI chemical-protein interaction Track," *In Proceedings of the Bio-Creative VI Workshop*, 2-4 July 2017, Bethesda Maryland, USA pp. 141-146.
[32] nlpaug. [Online]. Available: https://pypi.org/project/nlpaug
[33] L. N. Phan, J. T. Anibal, H. Tran, S. Chanana and E. Bahadroglu, "SciFive: a text-to-text transformer model for biomedical literature," *arXiv 2021*, arXiv:2106.03598.
[34] Z. Zhao, Z. Yang, L. Luo, H. Lin and J. Wang, "Drug-drug interaction extraction from biomedical literature using syntax convolutional neural network," *Bioinformatics*, vol. 32, no. 22, pp.3444-3453, 2016.
[35] D. Zhou, L. Miao and Y. He, "Position-aware deep multi-task learning for drug–drug interaction extraction," *Artificial Intelligence in Medicine*, vol. 87, no. 1, pp.1-8, 2018.
[36] W. Zheng, H. Lin, L. Luo, Z. Zhao and Z. Li, "An attention-based effective neural model for drug-drug interactions extraction," *BMC Bioinformatics*, vol. 18, no. 1, pp.445-467, 2017.
[37] G. Miolo, G. Mantoan and C. Orsenigo, "ELECTRAMed: a new pre-trained language representation model for biomedical NLP," *arXiv 2021*, arXiv:2104.09585.
[38] H. E. Boukkouri, O. Ferret, T. Lavergne, H. Noji and P. Zweigenbaum, "CharacterBERT: Reconciling ELMo and BERT for Word-Level Open-Vocabulary Representations from Characters," *In Proceedings of the 28th International Conference on Computational Linguistics*, 8-13 December 2020, Barcelona, Spain (Online), pp. 6903–6915.
[39] Y. Zhu, L. Li, H. Lu, A. Zhou and X. Qin, "Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions," *Journal of Biomedical Informatics*, vol. 106, no. 1, 103451.
[40] Y. Peng, A. Rios, R. Kavuluru and Z. Lu, "Extracting chemical-protein relations with ensembles of SVM and deep learning models," *Database*, vol. 2018, bay073, 2018.
[41] H. Lu, L. Li, X. He, Y. Liu and A. Zhou, "Extracting chemical-protein interactions from biomedical literature via granular attention-based recurrent neural networks," *Computer Methods and Programs in Biomedicine*, vol. 176, no. 2, pp. 61-68, 2019.
[42] C. Sun, Z. Yang, L. Luo, L. Wang. "A Deep Learning Approach with Deep Contextualized Word Representations for Chemical–Protein Interaction Extraction from Biomedical Literature," *IEEE. Access*, vol. 7, no. 15, pp. 1034-1046, 2019.
[43] Y. Peng, S. Yan and Z. Lu, "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets," *In Proceedings of the 18th BioNLP Workshop and Shared Task*, 1-5 August 2019, Florence, Italy, pp. 58–65.
[44] Z. Yuan, Y. Liu, C. Tan, S. Huang and F. Huang, "Improving Biomedical Pretrained Language Models with Knowledge," *In Proceedings of the 20th Workshop on Biomedical Language Processing*, 11-15 June 2021, Online, pp. 180-190.

[45] I. Beltagy, K. Lo and A. 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)*, 3-7 November 2019, Hong Kong, China, pp. 3615–3620.
[46] Y. Liu, M. Ott, N. Goyal, J. Du and M. Joshi, "Roberta: A Robustly Optimized BERT Pretraining Approach," *arXiv 2019*, arXiv:1907.11692.

**Xuefeng Fu** received a Ph.D. degree from Southeast University, China, in 2016. His major research areas are deep learning-based knowledge graph construction and inference. He specializes in symbolic logic for representing and reasoning about imprecise knowledge. He has worked as an associate professor at the Nanchang Institute of Technology (NIT) since 2005. He is in charge of the National Natural Science Foundation of China project "Research on efficient non-standard reasoning techniques for graph-based description logic" and presided over the project of Jiangxi Provincial Natural Science Foundation "Research on OWL ontology debugging method based on the graph". He is also a senior member of the CCF (China Computer Federation).

**Kailiang Wang** is a graduate student at the Nanchang Institute of Technology (NIT). His main research interests include machine learning and information extraction.

**Yanping Liu** is a graduate student at the Nanchang Institute of Technology (NIT). Her main research interest is multiple sentiment analysis.

**Weikun Chen** is a graduate student at Nanchang Institute of Technology (NIT). His main research interests include knowledge graphs and recommend systems.

**Jun Chen** is a graduate student of Nanchang Institute of Technology (NIT). His main research interests include deep learning and recommend systems.