Drug Knowledge Extraction Framework with Entity Pair Calibration for Chinese Drug Instructions

Xiaoliang Zhang
Nanjing Medical University
Lunsheng Zhou
Wuhan University of Science and Technology
Feng Gao
Wuhan University of Science and Technology
Zhongmin Wang
Nanjing Medical University
Yongqing Wang
Nanjing Medical University
Jianjun Guo
Nanjing Medical University
Shenqi Jing
Nanjing Medical University
Shumei Miao
Nanjing Medical University
Xin Zhang
Nanjing Medical University
Tao Shan
Nanjing Medical University
Yun Liu (liuyun@njmu.edu)
Nanjing Medical University

Research Article

Keywords: Information Extraction, Entity Pair Calibration, Pre-Training Model, Chinese Drug Instruction

Posted Date: December 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1143836/v1

RESEARCH

Drug Knowledge Extraction Framework with Entity Pair Calibration for Chinese Drug Instructions

Xiaoliang Zhang[1,2,4], Lunsheng Zhou[6], Feng Gao[6], Zhongmin Wang[1,2,3], Yongqing Wang[5], Jianjun Guo [1,2,3], Shenqi Jing[1,2,3], Shumei Miao[1,2,3], Xin Zhang[1,2,3], Tao Shan[1,2,3], Yun Liu[1,2,*]

[*]Correspondence: liuyun@njmu.edu.cn

[1] Department of Medical Informatics, School of Biomedical Engineering and Informatics, Nanjing Medical University, 101Longmian Avenue, Jiangning District, Nanjing 211166, P.R. China

[2] Institute of Medical Informatics and Management, Nanjing Medical University, 101Longmian Avenue, Jiangning District, Nanjing 211166, P.R. China

[3] Department of Information, the First Affiliated Hospital, Nanjing Medical University, No.300 Guang Zhou Road, Nanjing 210029, P.R. China

[4] Department of Data Application Management, the First Affiliated Hospital, Nanjing Medical University, No.300 Guang Zhou Road, Nanjing 210029, P.R. China

[5] Department of Pharmaceutical, the First Affiliated Hospital, Nanjing Medical University, No.300 Guang Zhou Road, Nanjing 210029, P.R. China

[6] Department of Computer Science, Wuhan University of Science and Technology, Wuhan, China.

Abstract

Existing pharmaceutical information extraction research often focus on standalone entity or relationship identification tasks over drug instructions. There is a lack of a holistic solution for drug knowledge extraction. Moreover, current methods perform poorly in extracting fine-grained interaction relations from drug instructions. To solve these problems, this paper proposes an information extraction framework for drug instructions. The framework proposes deep learning models with fine-tuned pre-training models for entity recognition and relation extraction, in addition, it incorporates an novel entity pair calibration process to promote the performance for fine-grained relation extraction. The framework experiments on more than 60k Chinese drug description sentences from 4000 drug instructions. Empirical results show that the framework can successfully identify drug related entities (F1 >= 0.95) and their relations (F1 >= 0.83) from the realistic dataset, and the entity pair calibration plays an important role (~5% F1 score improvement) in extracting fine-grained relations.

Keywords:

Information Extraction; Entity Pair Calibration; Pre-Training Model; Chinese Drug Instruction

1. Introduction

Drug instructions contain a wealth of drug knowledge, which can provide decision support for clinical diagnosis, prescription and healthcare management. However, textual descriptions provided in drug instructions usually take the form of long, complicated sentences, which are difficult to be processed by man or machine. The existing information extraction methods in the field of medicine include drug and disease entity recognition [1-3], drug-disease relationship extraction [4], etc. These methods mainly focus on specific tasks and do not present a complete framework for drug knowledge extraction. We argue that an complete drug knowledge extraction framework is crucial for drug knowledge construction and knowledge-based medical and pharmaceutical applications. The framework should take as input original drug instruction documents and generate relevant knowledge graphs based on drug instructions. Moreover, the relations extracted from drug instructions by current approaches are typically coarse-grained, e.g., most existing work provide solutions for identifying only the existence or polarity of drug-drug interactions, which is insufficient in clinical practice. For example, the drug knowledge provided by

Drugbank only provides the relevant text description of drug interactions as data properties (strings). It does not further provide word/phrase level semantics for detailed drug interactions knowledge, such as interaction mechanism, interaction result or clinical suggestion, thus prohibits fine-grained reasoning over the drug knowledge.

To overcome limitations in current approaches, this paper proposes a drug information extraction framework based on natural language processing techniques and deep learning models. It starts from automatically gathering the textual description of drug instructions from public sources. After data gathering and cleaning, it applies named entity recognition technology with natural language processing tools as well as deep learning models to identify 4 categories of entities: drugs, diseases, body parts and symptoms. After the entity name recognition, it derives a set of sentences with a list of relevant entities, then, an entity pair calibration (EPC) step is proposed. The goal of EPC is to distinguish between the main entity (the primary drug which the instruction is intended for) and the secondary entity (the drug mentioned in the sentence of the instruction, potentially associated with the primary drug entity), this step can reduce the noise for further drug relation extraction and facilitate accurate extraction of fine-grained relations. Finally, we use a combination of BERT (wwm) and BiGRU-ATT to extract the relationship between entity pairs.

The main contribution of this paper is two-fold:

1) a pretraining-based drug information extraction framework with a novel entity pair calibration process is proposed to extract drug-related entities and fine-grained relations from drug instructions;

2) to evaluate 1), a empirical study is carried out over a realistic dataset that contains over 60k sentences from 4000 drug instructions.

The remainder of the paper is organized as follows: Section 2 discusses the related work and the state-of-the-art with regard to drug information extraction; Section 3 presents the drug information extraction framework proposed in this paper, including the architectural design of the framework as well as specific neural models used in each module in the framework; Section 4 elaborates on the details of the entity pair calibration mechanism; Section 5 demonstrates and analyzes the empirical results before Section 6 concludes.

2. Related work

Rau et al. [5] first proposed the task of named entity recognition in 1991, the task of named entity recognition and relation extraction has developed rapidly. After the early methods based on traditional machine learning, such as Fang Xiaoshan et al. [6] proposed a method of named entity recognition based on rules, but this method relies heavily on the predefined high-frequency rules, The whole process of information extraction is not automatic enough. There are also neural network-based methods, such as the method of named entity recognition of electronic medical records based on BiLSTM-CRF model and attention mechanism, which is used by Chen Chen et al. [7] Has improved the performance of BiLSTM-CRF model to a certain extent, but it is only one of the tasks of the whole information extraction. Gong lejun [8] and others used a drug entity relationship extraction method based on GRU and CNN, which mainly extracted adverse drug reactions (ADRs).

In DDIE extraction 2013, the comprehensive evaluation rate was 75%. However, there was a lack of a complete set of information extraction processes, and the classification of the main entity and sub-entity was not well explained. Wang Yongchao [9] proposed an entity-relationship extraction method based on pointer network, but this method of pointer annotation does not pay enough attention to the continuity of entities, which will increase the probability of generating illegal sequence annotation in the entity recognition stage, resulting in low accuracy. There are a large number of entity names nested between diseases and drugs in the text of the medical field, Therefore, this method used in the field of medicine will produce a lot of wrong entity names. Only one of the tasks above presents a complete information extraction process, and the granularity level for the entity-relationship is not enough, thus prevents it to be used in practice.

In recent years, attention mechanism in natural language processing has gained interests. Ashish Vaswani [10] proposed the transformer model structure in 2017, which can decouple from the traditional neural network structures such as CNN or RNN, and it focused only on the attention mechanism. The bidirectional encoder representation from transformers (BERT) [11] pre-training model is developed based on the transformers, which pre-trains sentences by mask language model (MLM) and next sense prediction (NSP). But the original BERT model is used on English corpus. For Chinese corpus, the word-based "Mask" will lose some semantics. Coping to the special features of Chinese text expression, the whole word mask method proposed in BERT-wwm [12] is more effective in Chinese. Therefore, the BERT-wwm Chinese pre-training model is proposed. Instead of using the conventional character-based masking model, the whole word (consists of multiple characters) mask is used in BERT-wwm. To identify the whole word, a word segmentation tool is used to segment the text, then the whole word segment is marked and masked, with the help of several other optimizations, the experimental results on the public data set CMRC 2018 showed that BERT-wwm outperforms the original BERT. In this paper, we use BERT-wwm as a basic model for the drug information extraction framework.

3. Drug information extraction framework with pre-training

The drug information extraction framework proposed in this paper uses pre-training models as a key method to label sentences as well as entities and relations in sentences. In the following, we first present the architectural design of the framework and introduce the functions of its modules, then, we show the pretraining based models used in the relevant modules in the framework.

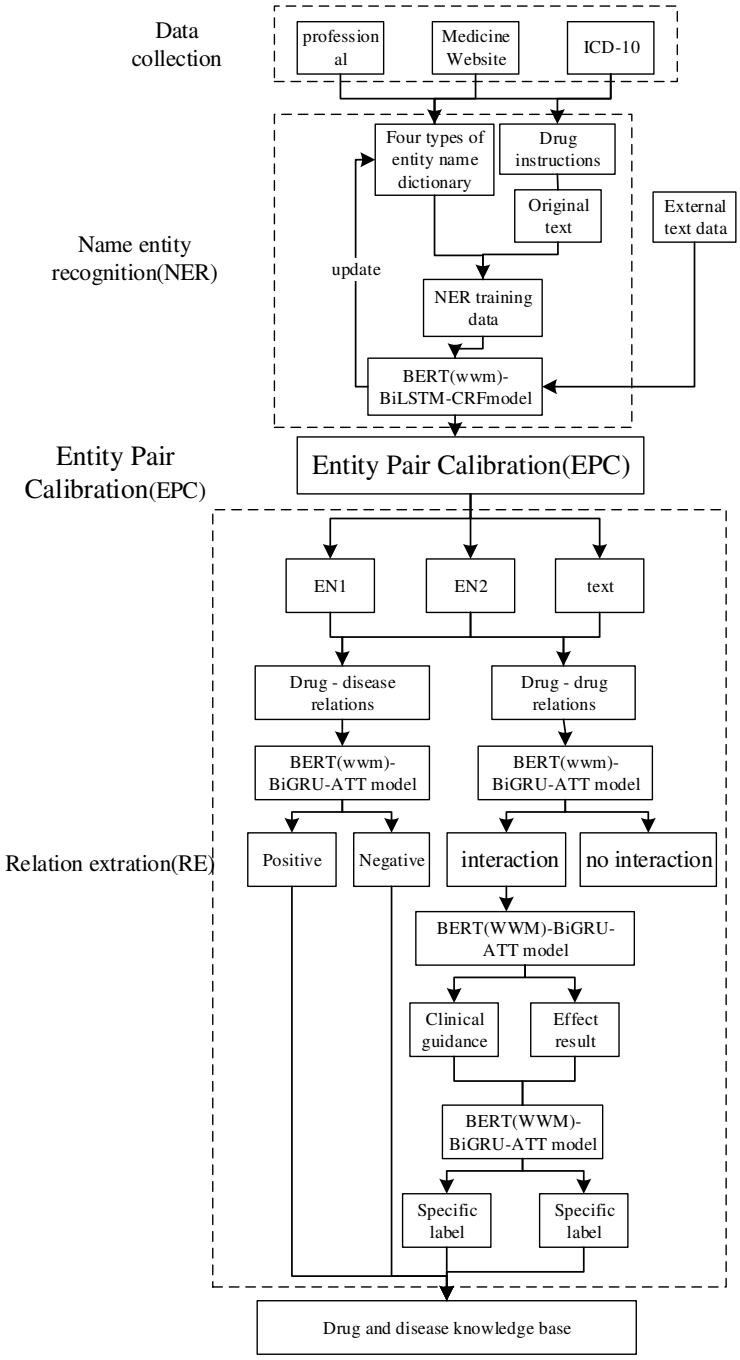

Figure 1. Overall framework of information extraction for drug instructions

3.1 Architectural design

The framework has four main modules: data collection and cleaning, named entity recognition, entity pair calibration and entity relation extraction, as depicted in Figure 1. The data collection is a automatic process that gathers drug instructions and organize them into semi-structured text inputs. The entity recognition and entity pair calibration are semi-supervised learning modules that identifies valid entity pairs for the relation extraction module, which is a supervised learning module.

3.1.1 Data collection and cleaning

The data collection module automatically collects drug instructions, disease diagnosis, treatment guidelines and other medical text data from public sources on the Internet. We

carry out this operation regularly with a careful selection of authoritative data sources, in order to provide rich and reliable data sources for drug information extraction. The structure of the collected and cleaned data is stored as spreadsheets, all textual description of a drug instruction is retained and categorized into columns using the section titles appears in the instructions, e.g., product name, primary chemical component, producer, drug reaction, drug interaction, etc.

3.1.2 Named entity recognition

The Named Entity Recognition (NER) module is mainly responsible for identifying four kinds of named entities from the text, which are "drugs", "diseases", "symptoms" and "body parts", thus preparing datasets for further relationship extraction. This module uses an entity name dictionary as a starting point and applies a BERT(wwm)-BiLSTM-CRF deep learning model (explained in Section 3.2) to learn new names. The entire procedure consists of 3 steps: Firstly, the original sentences in the instruction are converted to the format required by the training dataset for the named entity recognition algorithm model. Secondly, the sentences are processed with the longest string matching method with the limited dictionary names collected manually in the early stage. The longest string matching method can effectively avoid the wrong annotation of nested entity names in the dictionary; Thirdly, the deep learning model is trained by the dataset to generalize entity names, generalized entity names are added to the dictionary, so that the training data set of the model can be updated regularly, thus achieving a self-improvement solution.

3.1.3 Entity pair calibration

When drug and disease entities have been extracted from sentences, we can select sentences with n (n>=2) entities to further investigate entity relations. Since we are trying to learn binary relations, those n entities may produce C(n,2) entity pairs. However, not all those entity pairs are legitimate for relation extraction, e.g., the same drug/disease may have different names, occurring at different locations in the same sentence, or there may be a type-of/sub-class-of relation between them, in these cases, the entities pairs cannot enter drug/disease interaction extraction module without causing noise. To reduce noisy data caused by wrong entity pairs, the Entity Pair Calibration (EPC) module is used to verify the primary entity (the main drug described by the drug instruction, i.e., the intended object for the instruction) and secondary entity (the other drugs/diseases described in the sentence). In this paper, we use a modified word embedding method to splice the entity names with the original sentences and "Mask" entity names in the sentences. Finally, we feed the masked text into the BERT(wwm)-BiGRU-ATT model (explained in Section 3.2.3) to train the entity pair calibration module, so that it can identify the categories of entities in the text and remove entity pairs with unintended type-of/sub-class-of relations. We provide the details of the EPC process in Section 4.

3.1.4 Relation extraction

The Relation Extraction (RE) module extracts the relationship between the entity pair and its text information from the entity pair calibration module, which is mainly realized by the BERT(wwm)-BiGRU-ATT model (explained in Section 3.2.3). We distinguish between 2

main relations: drug-disease or drug-drug relation, based on the types of the secondary entity in entity pairs.

The relationship between "drugs" and "diseases" can be further divided into positive correlation and negative correlation; "Positive correlation" means that the drug can treat or alleviate the disease, "negative correlation" means that the drug can cause or aggravate the disease. The relationships between drugs are firstly divided into exist-interaction and non-interaction (sometimes a sentence states that there is no known interaction between two drugs); then for the sentences describing drug pairs that do have interactions, we further extract 3 types of relationships including interaction mechanism, clinical guidance and interaction results. Each drug-drug interaction relationship type can be categorized into more fine-grained labels. For example, "clinical guidance" can be further divided into use with caution (UC), prohibition of joint use (PU), joint use as per prescription (PP), joint use over different time spans (TS), different bottle injections (BI), etc. The "interaction results" are further divided into decreased efficacy (ED), improved efficacy (EI), increased adverse reactions (RI), improved efficacy and increased adverse reactions (EIRI), decreased efficacy and increased adverse reactions (EDRI), etc. Ultimately, fine-grained knowledge of drug interaction and drug usage can be established through automatic mining of text data in drug instructions. The extraction process is exemplified in Figure 2.

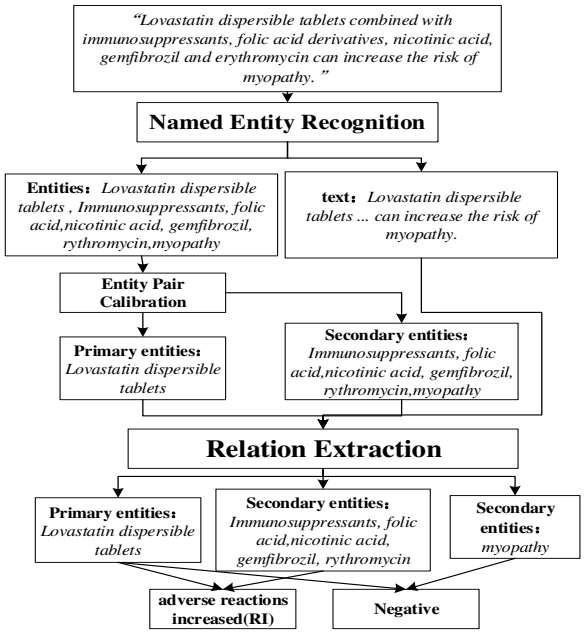

Figure 2. Examples of information extraction from drug instructions (translated from Chinese drug instruction)

3.2 Pre-Training based Model Design

The NER, EPC and RE modules are equipped with different deep learning models. All modules employ a variant of BERT model as the pretraining model, the NER module combines the pre-training model with BiLSTM and CRF, the EPC and RE modules combines the pre-training model with a BiGRU and a attention mechanism. In this section, we introduce the specific configurations of deep learning models used in relevant modules.

3.2.1 BERT(wwm) as the Pre-Training Model

We select the BERT(wwm) pre-training model as the first layer encoder. The BERT(wwm) is a variant of BERT with improvements on the masking method for Chinese text, that is, masking the entire word in Chinese instead of a single character, so as to keep the word and phrase level semantics in Chinese, as shown in Table 1.

Table1. The whole word mask example (translated from Chinese)

Categories	Examples
Original text	When sirolimus and voriconazole are used together, the blood drug level increase drastically, thus should not be used jointly.
MASK	When sirolimus and voriconazole are used together, the blood drug level [MASK]crease drastically, thus should [MASK] be used jointly.
WWM	When sirolimus and voriconazole are used together, the blood drug level [MASK] drastically, thus [MASK] be used jointly.

From the text semantics after "Mask", we can see that the text semantics after "WWM" is more complete, and the experimental effect of the final BERT(wwm) on the open dataset CMRC 2018 is also better than that of BERT, so we choose BERT(wwm) as the pre-training model.

The BERT(wwm) model is built based on BERT transformers, which is a two-way language model of transformers. Transformers is a method different from the traditional neural network, which completely relies on self -attention mechanism to calculate the vector representation of input and output. Firstly, three vectors are created for each word embedding vector of the encoder, which are query vector, key vector and value vector. These vectors are the weight matrix w obtained by combining the word vector with the three training vectors W_Q, W_K, W_V. The matrix is the parameter to be learned, and then the output is obtained by calculating the formula:

$$Z = Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

In formula (1): each line of Q, K and V represent the query vector, key vector and value vector; D is the length of the key vector, divided by dk1/2 is to get a more stable gradient in the process of back propagation. The structure of the BERT(wwm) model is shown in Figure 3.

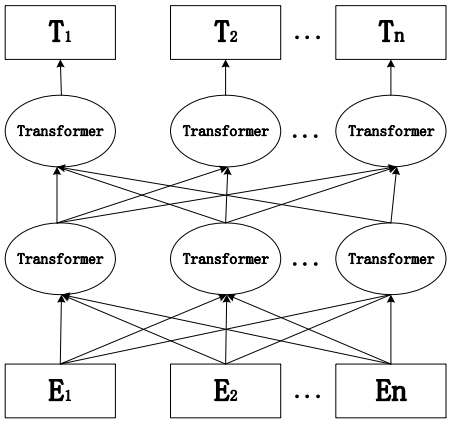

Figure 3. BERT(wwm) model architecture

3.2.2 BERT(wwm) -BiLSTM-CRF for Named Entity Recognition

The BERT(wwm)-BiLSTM-CRF model is the deep learning model of the NER module. The purpose of this model is to recognize the names of drugs, diseases, symptoms and

body parts from the medical text, so as to provide data for the subsequent relationship extraction task. The BERT(wwm) pre-training model trains an intermediate word vector for the encoder, which better integrates the syntax and semantics of the context of the word. Then the intermediate word vector is used as the input of the downstream BiLSTM model. BiLSTM is a bidirectional recurrent neural network, which can better capture the dependency of context semantics in text. The prediction score of each tag is labeled by the training output sequence of the BiLSTM network, which will be used as the input of the CRF layer. CRF layer can obtain the constraint rules of sequence labeling data from the training data, thus greatly reducing the probability of illegal sequence labels predicted.

The overall model architecture is shown in Figure 4. In this figure, the input text is first transformed into a dynamic semantic vector by BERT(wwm), The vector is trained by the BiLSTM network and the constraint rules of the CRF layer, and the label corresponding to each word is finally output. "B-DRUG" represents the beginning of an entity, "I-DRUG" represents the middle and end of an entity, "O" represents the part that is not an entity.

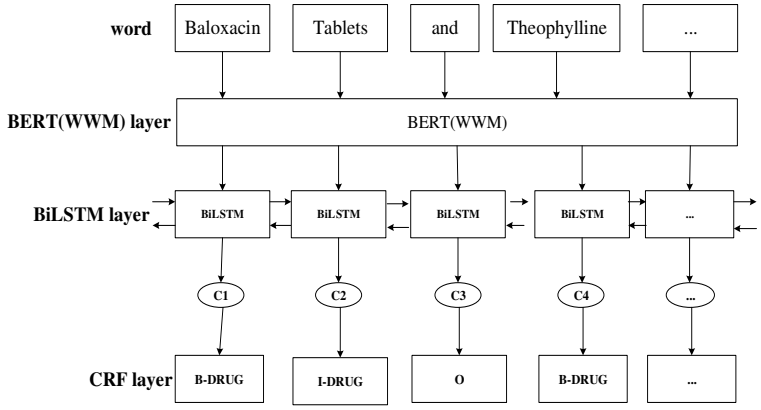

Figure 4. BERT(wwm)-BiLSTM-CRF model architecture

3.2.3 BERT(wwm)-BiGRU-ATT based Entity Pair Calibration and Relation Extraction

The BERT(wwm)-BiGRU-ATT model is used in the EPC and RE modules. Different word embedding can be used for different tasks. In this model, BERT (wwm) is still used as the upstream encoder to encode the text as the intermediate word vector, and the downstream uses a bidirectional GRU network and attention mechanism to train the upstream word vector. GRU network is a variant of the LSTM network, which can solve the problems of long-term dependence and gradient dispersion. Compared with the LSTM network, the GRU network has higher computational efficiency and occupies less memory. Therefore, considering the heavy workload of EPC and RE task of computation, the bidirectional GRU network is used. The entire model is accompanied by a final attention layer as shown in Figure 5.

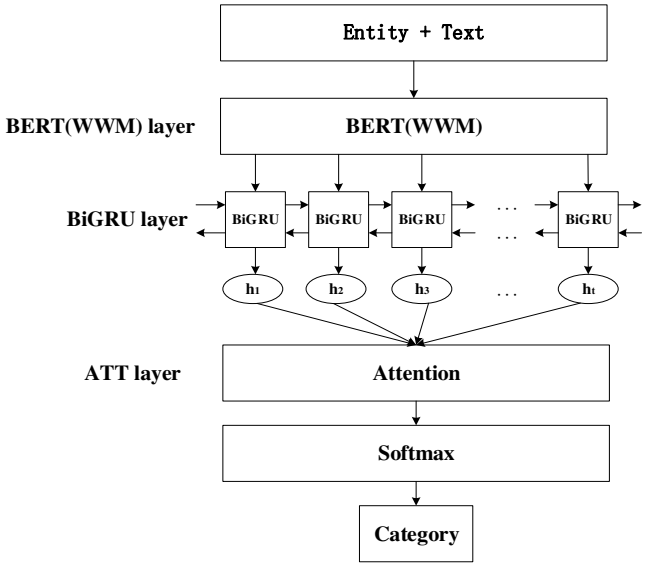

Figure 5. BERT(wwm)-BiGRU-ATT model architecture

We use a self-attention mechanism in Figure 5. For a text input, the model can generate different attention scores on different words, so that the model can pay more attention to certain words and improve the overall performance of the model. More specifically, given a sentence representation s=$\{X_1, X_2 \ldots X_n\}$, where X_n represents a word embedding, the model first randomly initializes the three vectors Q, K, V for each word, representing the query, key and value vectors. Then, it trains 3 weight matrices W_Q, W_K, W_V, and adjusts the word vectors using $q_n = X_n * W_Q$, $k_n = X_n * W_K$, $v_n = X_n * W_V$, Then, the preliminary word score is given by $f_1 = q_1 * \sum_{i=1}^{n} k_i$, the final score is given by $F = \sum_{j=1}^{n} v_j * \mathrm{Softmax}(f_j)$. This score determines the importance of other parts of the input sentence when encoding the word at the position. As shown in the example in Figure 6:

When combined with **voriconazole**, the blood concentration of **sirolimus** may **increase** significantly, so the two drugs **can not** be used at the same time

Figure 6. Text example of attention mechanism (translated from Chinese)

For a human brain, the two words "升高" (increase) and "不可"(cannot) in such a text determines the relation between the two entities. Similarly, through attention calculations, these two words are assigned with a greater attention score, thus indicate that the appearance of these two words w.r.t. their position is important to identify a relation between entities.

4. Entity pair calibration

In drug instructions, there are often multiple expressions of the same entity name, and there are inclusion relationships (for example, one entity is a component of another entity), sub-class and type-of relationships among different entities. There should be no interaction between such entity pairs. When this kind of entity pair is directly used as the input of the subsequent relation extraction task, it will cause a lot of noise.

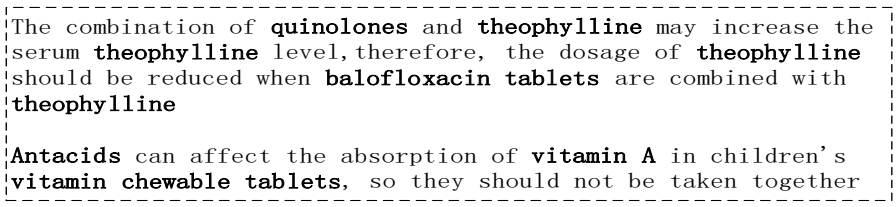

Figure 7. Examples of type-of/component-of relations in drug instructions (translated from Chinese)

For example, in the two texts in Figure 7, the first text involves three entity names:4-quinolones, theophylline and Balofloxacin Tablets. Among the three entity names, 4-quinolones is a drug type for balofloxacin tablets. Similarly in the second text, vitamin-A is a component of Multivitamin chewable tablets. In both cases, the entity pairs are noise. A naive approach will simply extract all entities from text and apply a Cartesian product to produce all entity pairs, which will result in very noisy input for the RE module. A natural improvement will be using a filter based on apriori knowledge after the Cartesian product to eliminate illegal entity pairs, such as Algorithm 1. However, this approach imposes a strong assumption that an accurate and complete knowledge base that describes the type-of, component-of and subclass-of relations between drugs exists. In practice, such apriori knowledge may be incomplete or even non-existence.

Algorithm 1. Traditional rule-based entity pair extraction

Input: set of entities: *ENs*
Output: set of entity pairs: *EN_tuple*

begin
ENre$_1$*, ENre*$_2$*, EN_tuple = [], [],[]*
for (*EN* **in** *ENs*):
if *EN* **in** *compositions*:
ENre$_1$.**append**(*EN*)
else:
ENre$_2$.**append**(*EN*)
EN_tuple = **Cartesian_product**(*ENre*$_1$*, ENre*$_2$)
EN_tuple = **FilterWithAprioriKB***(EN_tuple)*
return *EN_tuple*
end

In order to avoid using such wrong entity pairs, we propose to check the entity pair after the named entity recognition task and before the entity relationship extraction, so as to determine whether the entity pair is valid. In essence, we employ a deep learning model to label the primary and secondary entities in a entity pair to verify its validity. The primary entity set contains only the drug entity that the instruction is intended for, or its type, super-class or drug-composition, the secondary entity contains only the entity associated with the primary entity in the instruction. This way, we can produce correct entity pairs for relation extraction.

In order to successfully recognize the entity category from the text, we use the "WWM" mechanism similar to BERT (wwm) model. We mask the primary and secondary entities (with m‘#’symbols, where m is the length of the masked entity) when training the algorithm to identify them. Meanwhile, to pertain the information in the entity name, we use the same splicing method in BERT and concatenate the entity name (with a‘$ symbol) in front of the masked sentence. In this way, we can not only ensure a lossless text embedding but also obtain flexibility in the embedding method. Table2 depicts an example of EPC training data.

Table 2. Example of the training data for entity pair checking algorithm (translated from Chinese)

Label	Text (masked)
Primary Entity	quinolones$The combination of ########## and theophylline may increase the level of theophylline in serum, so the dose of theophylline should be reduced when balofloxacin tablets are combined with theophylline
Primary Entity	balofloxacin tablets$The combination of quinolones and theophylline may increase the level of theophylline in serum, so the dose of theophylline should be reduced when ########## are combined with theophylline
Secondary Entity	theophylline$The combination of quinolones and ######## may increase the level of ######## in serum, so the dose of theophylline should be reduced when balofloxacin tablets are combined with theophylline

Algorithm 2 details the procedure of the EPC module. It first transforms the named entity set ENs and the relevant sentence text with a word embedding function *trans_en_ text()*, the result of this function is the encoded text: *text_new*, The label of the entity can be generated when *text_ new* is fed into the neural network model, and the category of the entity can be determined according to the label. Finally, all the primary and secondary entities are processed by function: *cartesian_product()*, generating the output entity pairs: *EN_ tuple*.

Algorithm 2. Entity Pair Calibration extraction using pre-training

Input: set of entities: *ENs, text containing entities: text*
Output: set of entity pairs: *EN_tuple*

```
begin
ENnn1, ENnn2, EN_tuple = [], [],[]
for (EN in ENs):
        text_new = trans_en_text(EN, text)
        label = NN(text_new)
        if label == en_master:
                ENnn1.append(EN)
        else:
                ENnn2.append(EN)
EN_tuple = Cartesian_product(ENnn1, ENnn2)
return EN_tuple
end
```

5. Empirical results and analysis

In order to evaluate our approach, we experiment the modules in this framework over realistic datasets. In this section, we first present the experiment dataset and configuration, then, we demonstrate our experiment results and analysis.

5.1 Experiment dataset and configuration

The experimental data of this paper mainly comes from authoritative data sources on the Internet. Starting from a medical entity dictionary, the medical domain sequence annotation dataset with 3 million Chinese characters for named entity recognition is constructed, which is used to train and verify the NER module. There are four types of entity names: drug name, disease name, body part name and symptom name. As shown in Table 3.

Table 3. Entity name categories and examples

Entity	Examples

categories	
DRUG	oxazepam, atorvastatin, tadalafil
DISEASE	diabetes, AIDS, nasopharyngeal cancer
BODY	tongue, stomach, spleen
SYMPTOM	wheezing, fatigWue, ventosity

For the EPC and RE modules, we prepared 60k Chinese sentences from 4000 drug instructions. EPC labels entities with 2 categories: primary or secondary, RE labels relations among entities with 12 fine-grained drug-drug and drug-disease relations, as introduced in Section 3.1.4. The proportion for training and test dataset is 4:1. All training dataset is labelled manually by pharmaceutical experts.

In terms of parameter setting, this paper uses the base version of Chinese BERT(wwm) pre-training model, and the network structure consists of 12-layer, 768-hidden, 12-heads and 110M parameters. The maximum sentence length set by the named entity recognition part is 128; The number of sentences processed each time is 64; The learning rate is $5 * 10^{-5}$; The number of LSTM network nodes is set to 128; Dropout is set to 0.5 to prevent overfitting, and the training times are 3. The maximum sentence length of the entity pair verifier and relation extraction part is 80; The number of sentences processed each time is 16; The learning rate was 0.001; The number of GRU network nodes is set to 128; Dropout is set to 0.2 to prevent overfitting, and the training times are 10. The words embedded in BiGRU-ATT model adopt the character vector obtained by word2vec training. The number of characters in the table is 16116 and the vector dimension is set to 100. All the other parameters are initialized randomly.

5.2 Experiment result and analysis

In this section, we discuss the experiments results for the NER, EPC and RE modules.

5.2.1 Named entity recognition

The comparative experiment results of the NER module are shown in Figure 8. We evaluate the F1 scores of the 4 entity types using 3 different models, including 2 pretraining based models: BERTwwm-BiLSTM-CRF, BERT-BiLSTM-CRF and a conventional model BiLSTM-CRF.

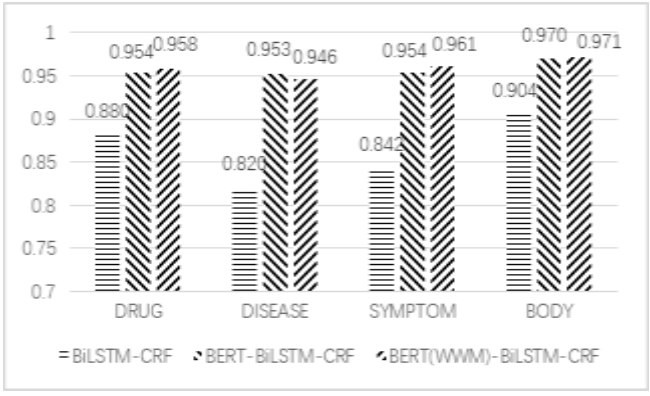

Figure 8. F1 score for named entity recognition model

As depicted in Figure 8, pre-training models (BERT/BERTwwm) significantly improve the performance of NER compared to conventional models. Also the BERT(wwm)-BiLSTM-CRF model generally performs the best in the recognition of drug, symptom and body names, except for disease names. In the prediction results of the entity recognition model, the BERT(wwm)-BiLSTM-CRF model can also successfully recognize the names that are not matched by dictionary matching, as shown in Figure 9.

Diclofenac	O	B-DRUG
sodium	O	I-DRUG
can	O	O
be	O	O
used	O	O
in	O	O
the	O	O
treatment	O	O
of	O	O
osteoarthritis	B-DISEASES	B-DISEASES

Figure 9. BERT(wwm)-BiLSTM-CRF model predicts result

In Figure 9, the first column is the text data labeled by sequence, the second column is the label labeled by the dictionary, and the third column is the label predicted by the model. It can be seen that the BERT(wwm)-BiLSTM-CRF model successfully identifies the drug name "双氯芬酸钠" (Diclofenac Sodium) which does not appear in the dictionary. This feature can help expand the original dictionary and facilitate a semi-supervised approach.

5.2.2 Entity Pair Calibration

To evaluate the performance of the model used in EPC, we compare the precision, recall and F1 score for identifying the primary (EN1) and secondary (EN2) entities using 4 different models: DPCNN, BiLSTM-ATT, BERT-BiGRU-ATT and BERTwwm-BiGRU-ATT. The experimental results of the EPC are shown in Table 4

Table 4. Experimental results of Entity Pair Calibration (EPC) model

model	Entity Role	Precision	Recall	F_1
DPCNN	EN 1	0.9256	0.9387	0.9321
	EN 2	0.9694	0.9626	0.9660
BiLSTM-ATT	EN 1	0.8935	0.9104	0.9019
	EN 2	0.9552	0.9463	0.9507
BERT-BiGRU-ATT	EN 1	0.9717	0.9836	0.9776
	EN 2	0.9920	0.9862	0.9891
BERT(wwm)-BiGRU-ATT	EN 1	**0.9719**	**0.9918**	**0.9817**
	EN 2	**0.9960**	**0.9862**	**0.9911**

The results in Table 4 suggests that the performance of DPCNN [14] without pre-training neural network is quite good. It was proposed by Tencent AI Lab in 2017 to extract long-distance text dependence by continuously deepening the network and using a network based on word level. It outperforms the BiLSTM-ATT model, but not better than the pre-training models. The results show that our approach performs the best among the 4 models, and it can achieve 98% and 99% F1 values for primary and secondary entity identifications, respectively.

5.2.3 Relation extraction

The experimental results of RE module w.r.t. 4 coarse-grained relations (polarity of drug-disease relations, existence of drug-drug interactions, clinical guidance and interaction results) are shown in Figure 10-13. We compare 4 different models: 1) conventional deep learning approach (BiGRU-ATT), 2) deep learning with pre-training model (BERT-BiGRU-ATT), 3) on top of 2), add an entity pair filter with aprioir knowledge (rule-based) matching (BERT-BiGRU-ATT(RE)) and 4) on top of 2, adding the EPC module as described in Section 4 (BERT-BiGRU-ATT(EPC)).

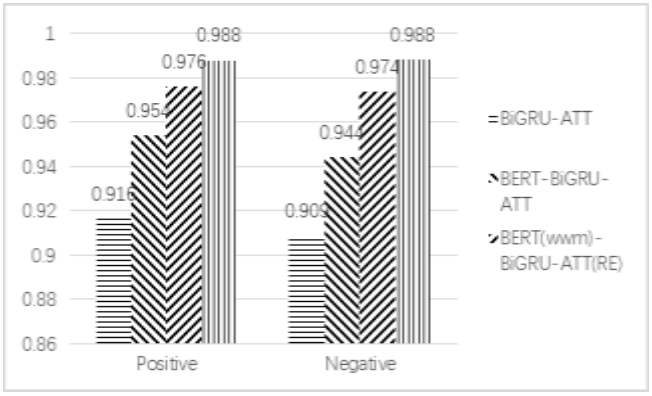

Figure 10. F1 score for positive and negative correlation extraction between drugs and diseases

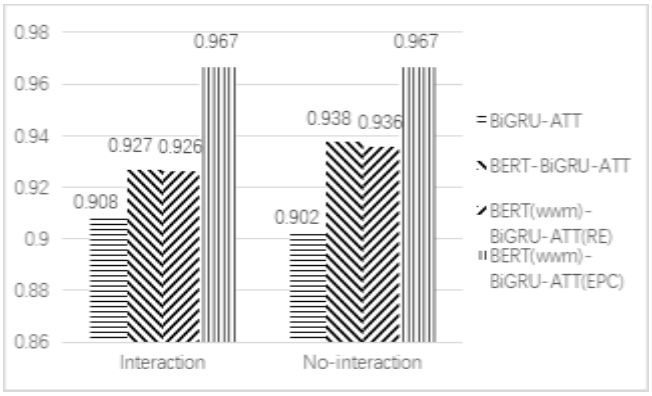

Figure 11. F1 score for extracting the existence of drug-drug interaction

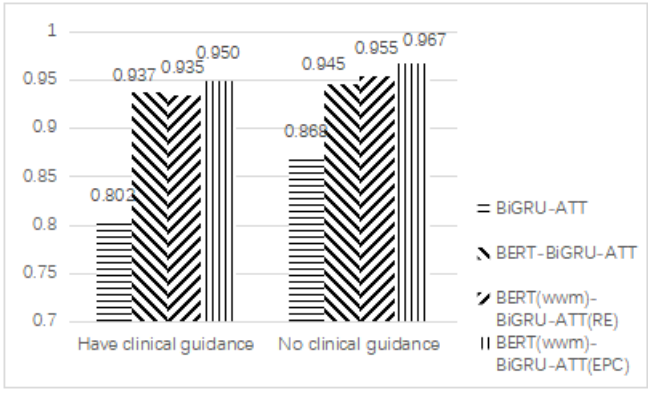

Figure 12. F1 score for extracting the existence of "clinical guidance" in drug-drug interaction

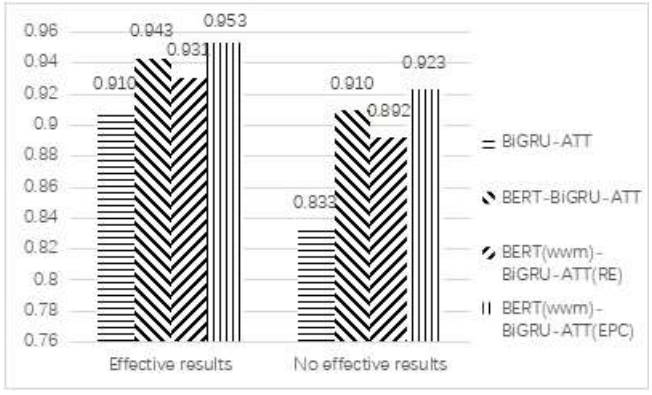

Figure 13. F1 score for extracting the existence of "interaction result" in drug-drug interaction

As can be seen from Figure 10-13, pre-training models outperform conventional deep learning models. Adding the EPC module performs the best as we expected. However, using apriori knowledge to eliminate wrong entity pairs performs similarly to not using filters. This demonstrates that the performance of the rule-based entity calibration heavily relies on the quality of the apriori knowledge. Figure 14 and 15 show the experiment results for 10 more fine-grained RE tasks as discussed in Section 3.1.4, using the same 4 models.

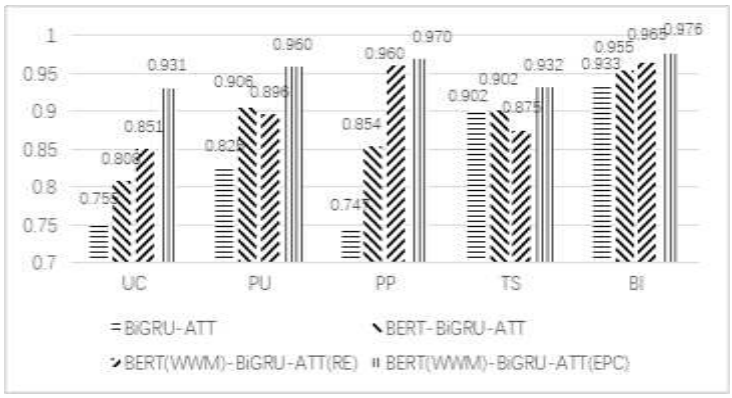

Figure 14. F1 score for fine-grained "clinical guidance" relation extraction

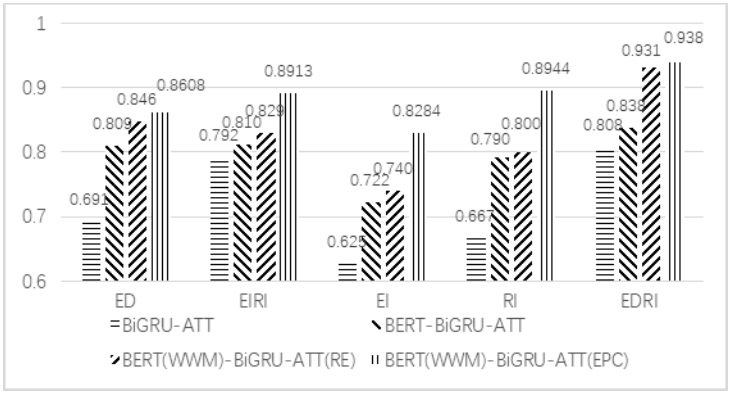

Figure 15. F1 score for fine-grained "interaction result" relation extraction

Results from Figure 14 and 15 suggests that the performance of fine-grained relation extraction is improved significantly by EPC. More specifically, for the relation of "use with caution (UC)", the F1 value is increased by 8.04%, and for the relation of "adverse reaction increase (RI)", the F1 value is increased by 9.44%. Compared to the coarse grain relation extraction result (~2% F1 improvement on average) in Figure 10-13, we can see that the EPC module has a much bigger contribution (5% F1 improvement on average) while extracting more fine-grained and complex relations.

6. Conclusion

Extracting fine-grained drug knowledge from professional drug instructions is a complicated task for both human and machines. In order to address this problem, we propose a holistic information extraction framework based on an integration of pre-training and deep learning models. The abstract structure of the framework follows a basic design of pipeline information extraction methodology that includes data gathering, Named Entity Recognition and Relation Extraction, on top of which, we design a novel procedure called Entity Pair Calibration and place it between the NER and RE tasks to reduce the input noise for drug relation extraction.

The framework is empirically evaluated over more than 60000 sentences from Chinese

drug instructions, covering 4000 different drugs. Experiment results demonstrate that the proposed framework can achieve >=0.95 F1 score in NER and >=0.83 in RE tasks. Moreover, the proposed EPC module can significantly improve the performance for fine-grained relation extraction (F1 score improvement up to 9.44%, 5% on average). Using this framework, we have successfully built a drug interaction knowledge base with over 1,300,000 RDF triples, describing more than 180,000 drug-drug and drug-disease relations. It is worth mentioning that although we design and evaluate the framework for Chinese drug instructions, it can be easily adapted for other languages by changing the pre-training model into BERT or other language-specific models. Also, the fundamental idea of EPC can be reused for other languages.

Our future work may carry out in 3 aspects. Firstly, the pre-training model BERT(wwm) used in this paper can be optimized with grammatical structure and semantic information[15,16]. Secondly, the architecture of the framework is configurable but not dynamically adaptable, we could improve on this aspect by monitoring the result and dynamically switch to better models. Also, adding confrontation training [17] in the word embedding stage of the model can improve the robustness of the model. Finally, there is an inherent problem in this domain, which is the imbalance of samples in drug-drug interactions. This problem is not easily solvable using pure learning-based approaches, we could explore using a hybrid approach of knowledge and learning-based approach to address this issue [18].

Appendix

Abbreviations

ADR: adverse drug reactions
ATT: attention mechanism
Bert: bidirectional encoder representation from transformers
BI: different bottle injections
BiLSTM: bidirectional long short-term memory
CRF: conditional random field
CNN: convolutional neural network
ED: decreased efficacy
EDRI: decreased efficacy and increased adverse reactions
EI: improved efficacy
EIRI: improved efficacy and increased adverse reactions
EPC: entity pair calibration
GRU: gated recurrent unit
LSTM: long short-term memory
MLM: mask language model
NER: Named entity recognition
NSP: next sense prediction
PP: joint use as per prescription
PU: prohibition of joint use
RE: relation extraction
RI: increased adverse reactions
RNN: recurrent neural network
TS: joint use over different time spans
UC: use with caution
WWM: whole word masking

Ethics and Consent to Participate

Not applicable.

Consent to Publish

Not applicable.

Availability of Data and Materials

The datasets generated and/or analyzed during the current study are available in the [GitHub] repository, [https://github.com/zhou-ls/Drug_Knowledge_Extraction_Framework]

Competing Interests

The authors declare that they have no competing interests.

Funding

Paper supported by National Natural Science Foundation of China [U1836118], grants from the National key Research & Development plan of Ministry of Science and Technology of China[Grant no. 2018YFC1314900, 2018YFC1314901], the industry prospecting and common key technology key projects of Jiangsu Province Science and Technology Department [Grant no. BE2020721], the Special guidance funds for service industry of Jiangsu Province Development and Reform Commission [Grant no. (2019)1089], the big data industry development pilot demonstration project of Ministry of Industry and Information Technology of China [Grant no. (2019)243), (2020)84].

Authors' Contributions

Zhang, Xiaoliang proposes the main idea of this paper and help organize the data annotations required for drug instructions.
Zhou, Lunsheng drafts experiment sections of the manuscript and conducted relevant experiments.
Gao, Feng provides the architectural design, model selection and configuration of the framework.
Wang, Zhongmin provides overall guidelines for pharmaceutical knowledge graph construction.
Wang, Yongqing provides specific instructions on building pharmaceutical knowledge graph and drug annotations.
Guo, Jiangjun provides drug knowledge based application design.
Jing, Shenqi develops drug knowledge graph construction program.
Miao, Shumei develops drug knowledge graph construction program.
Zhang, Xin provides the literature reviews.
Shan, Tao provides the literature reviews.
Liu, Yun is the guarantor of this paper.
All authors have reviewed and consented the publications.

Acknowledgements

Not applicable.

Authors' information

Zhang, Xiaoliang（1988-）, male, master, engineer, mainly engaged in medical informatics, Pharmaceutical knowledge map and medical artificial intelligence research. Email: 798494709@qq.com
Zhou, LunSheng (1998 -), male, postgraduate student, mainly engaged in knowledge graphs and deep learning.

Email:459548202@qq.com
Gao, Feng（1986-）, male, doctor, lecturer, mainly engaged in knowledge mapping research. Email: feng.gao86@wust.edu.cn
Wang, Zhongmin (1974 -), male, doctor, professor, mainly engaged in medical informatics, Pharmaceutical knowledge map and medical artificial intelligence research. Email: wzm.cn@139.com
Wang, Yongqing (1972 -), male, professor, chief pharmacist, doctoral supervisor, mainly engaged in Pharmaceutical management and intelligent pharmacy research.. Email: wyqjsph@163.com
Guo, Jianjun (1974 -), male, master, senior engineer, mainly engaged in medical informatics, Pharmaceutical knowledge map and medical artificial intelligence research. Email: 15996295776@139.com
Jing, Shenqi (1983 -), male, master, senior engineer, mainly engaged in medical informatics, Pharmaceutical knowledge map and medical artificial intelligence research. Email: jingshenqi@jsph.org.cn
Miao, Shumei (1988 -), female, master, engineer, mainly engaged in medical informatics, Pharmaceutical knowledge map and medical artificial intelligence research. Email: shm_miao@163.com
Zhang, Xin (1983 -), female, master, associate professor, mainly engaged in medical informatics, Pharmaceutical knowledge map and medical artificial intelligence research. Email: 79953912@qq.com
Shan, Tao (1986 -), male, master, engineer, mainly engaged in medical informatics, Pharmaceutical knowledge map and medical artificial intelligence research. Email: nanjingtaotao@126.com
Liu, Yun (1967 -), female, professor, chief physician, doctoral supervisor, mainly engaged in medical informatics, hospital information management and medical artificial intelligence research. Email: liuyun@njmu.edu.cn. Liu Yun is the corresponding author as well as the guarantor of this paper.

References

1. YANG Z H,LIN H F,LI Y P.Exploiting the performance of dictionary-based bio-entity name recognition in biomedical literature[J].Computational biology and chemistry,2008,32(4):287-291.
2. Kang N , Singh B , Bui C , et al. Knowledge-based extraction of adverse drug events from biomedical text[J]. Bmc Bioinformatics, 2014, 15(1):1-8.
3. D Mahendran, Mcinnes B T . Extracting Adverse Drug Events from Clinical Notes[J]. AMIA. Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, 2021, 2021:420-429.
4. WEI C H,PENG Y F,LEAMAN R,et al.Overview of the Bio Creative V chemical disease relation (CDR) task[C] // Proceedings of the 5th Bio Creative Challenge Evaluation Workshop.Oxford:Oxford University Press,2015:154-166.
5. Rau L F . Extracting company names from text[C]// Artificial Intelligence Applications, 1991. Proceedings. Seventh IEEE Conference on. IEEE, 1991.
6. Fang Xiaoshan, Sheng Huanye. An example-based method for rule extraction in Chinese name recognition (English) [A]. Chinese Information Society of China. Proceedings of the First Student Computational Linguistics Symposium [C]. Chinese Information in China Society: Chinese Information Society of China, 2002: 8.
7. Chen Chen,Liu Xiaoyun,Fang Yuhua.Electronic Medical Record Named Entity Recognition Integrated with Attention Mechanism[J].Computer Technology and Development,2020,30(10):216-220.
8. Gong Lejun, Liu Xiaolin, Gao Zhihong, Li Huakang.Drug interaction relationship extraction based on two-way GRU and CNN[J/OL].Journal of Shaanxi Normal University (Natural Science Edition):1-6[2020-11-25].http://kns.cnki.net/kcms/detail/61.1071.N.20201013.1330.002.html.
9. Wang Yongchao,Mu Hualing,Zhou Lingzhi,Xing Wei.A Joint Extraction Method of Entity and Relationship Based on Pointer Network[J/OL].Computer Application Research:1-5[2020-11-28].https://doi.org/10.19734/j.issn.1001-3695.2020.04.0113.
10. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.
11. Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.
12. Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for Chinese BERT [EB/OL]. [2019-10-29]. https://arxiv.org/pdf/1906.08101.pdf.
13. Lample G , Ballesteros M , Subramanian S , et al. Neural Architectures for Named Entity Recognition[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016.
14. Johnson R , Zhang T . Deep Pyramid Convolutional Neural Networks for Text Categorization[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017.
15. Liu Y , Ott M , Goyal N , et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[J]. 2019.
16. Sun Y , Wang S , Li Y , et al. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5):8968-8975.
17. Goodfellow I J , Shlens J , Szegedy C . Explaining and Harnessing Adversarial Examples[J]. Computer ence, 2014.
18. Li S , Xu H , Lu Z . Generalize Symbolic Knowledge With Neural Rule Engine[J]. 2018.