



# Joint Model of Entity Recognition and Relation Extraction with Self-attention Mechanism

MAOFU LIU and YUKUN ZHANG, Wuhan University of Science and Technology  
WENJIE LI, The Hong Kong Polytechnic University  
DONGHONG JI, Wuhan University

In recent years, the joint model of entity recognition (ER) and relation extraction (RE) has attracted more and more attention in the healthcare and medical domains. However, there are some problems with the prior work. The joint model cannot extract all the relations for a specific entity, and the majority of joint models heavily rely on complex artificial features or professional natural language processing (NLP) tools. In this article, we construct a novel joint model that can simultaneously extract all medical entities and relations from medicine Chinese instructions. Moreover, the self-attention mechanism is introduced to the joint model to learn word intra-sentence dependencies. The proposed model is evaluated using a medicine Chinese instruction dataset that we collect and an open dataset provided in CoNLL-2004. Experimental results show that the model with self-attention achieves the state-of-the-art performance.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Natural language processing**;

Additional Key Words and Phrases: Joint model, entity recognition, relation extraction, self-attention, medicine chinese instruction

## ACM Reference format:

Maofu Liu, Yukun Zhang, Wenjie Li, and Donghong Ji. 2020. Joint Model of Entity Recognition and Relation Extraction with Self-attention Mechanism. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 19, 4, Article 59 (May 2020), 19 pages.

<https://dx.doi.org/10.1145/3387634>

## 1 INTRODUCTION

The past decade has witnessed that more and more attention has been paid to the processing of healthcare and medical texts, such as electronic medical records, electronic health records, biomedical literatures from PubMed,<sup>1</sup> medicine instructions, and question answering pairs from the medical community. Unlike the texts from social media online platforms, the medicine

<sup>1</sup><https://pubmed.ncbi.nlm.nih.gov/>.

The work presented in this article is supported by the Major Projects of National Social Science Foundation of China under Grant No. 11&ZD189 and the National Natural Science Foundation of China under Grant No. 61672445.

Author's addresses: M. Liu and Y. Zhang, School of Computer Science and Technology, Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan 430065, China; emails: liumaofu@wust.edu.cn, 978523050@qq.com; W. Li, PQ707, Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong; email: cswjli@comp.polyu.edu.hk; D. Ji, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China; email: dhji@whu.edu.cn.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2020 Association for Computing Machinery.

2375-4699/2020/05-ART59 \$15.00

<https://dx.doi.org/10.1145/3387634>

instructions are usually well written, authoritative, comprehensive, and formally organized. Moreover, the medicine instructions contain the rich semantic data expressed with plenty of medical entities and relations among them. Therefore, entity classification/recognition and relation extraction contribute to the semantic understanding of the medicine instructions.

Entity classification/recognition (EC/NER) and relation extraction (RE) from medical texts refer to identifying biomedical entities, such as medicine, disease, gene, protein, and bacteria, from the unstructured texts in the healthcare and medical domain, and extracting semantic relations among these entities, e.g., gene and disease, protein and protein interaction, and so on. They are the basic and critical tasks in medical event extraction, medical information retrieval, and clinical medication guideline and can be beneficial for medical knowledge graph construction, medical question answering, medicine recommendation, and many other applications related to healthcare and medicine.

In the previous work, most methods have adopted pipeline models to tackle these tasks. Given the medical texts, the EC/NER task is usually carried out by conditional random fields (CRF) and then paired up to commit the RE task by support vector machines (SVM). There are two drawbacks with these pipeline models: (1) error propagation, that is, errors in the EC/NER task are transmitted to the RE task, and (2) the ignorance of the inherent association or interaction between EC/NER and RE.

Recently, due to the popularity of deep learning, more and more researchers have turned to explore joint models based on neural networks, performing both EC/NER and RE tasks simultaneously, to solve the aforementioned problems. They have achieved state-of-the-art performance on many datasets, such as adverse drug effect (ADE) [1] and Dutch real estate classifieds (DREC) [2]. However, there still exist some problems: (1) Most models heavily rely on hand-crafted or artificial features and NLP tools, and (2) they cannot identify all semantic relations of a specific entity in a sentence. For instance, in the following Example 1, “尿路感染 (urinary tract infection)” could be involved in not only the cure relationship with “谷氨酸诺氟沙星注射液 (norfloxacin glutamate injection)” but also the causality relationship with “敏感菌 (sensitive bacteria)” simultaneously.

Example 1: 谷氨酸诺氟沙星注射液适应症: 本品适用于敏感菌所致的呼吸道、尿路感染、淋病、前列腺炎、肠道感染和伤寒及其他沙门菌感染。

*Norfloxacin glutamate injection: This product is suitable for respiratory tract, urinary tract infection, gonorrhea, prostatitis, intestinal infection, typhoid caused by sensitive bacteria, and other infections by salmonella.*

Meanwhile, most studies have focused on English datasets, and so far, there is no public dataset available in Chinese. Therefore, we construct our own dataset of medicine Chinese instructions and propose a novel joint model to solve the two problems mentioned above, with the deep features automatically learned by BiLSTM and extracting all the relations among entities by self-attention mechanism and multi-head selection. In our constructed dataset, there are also some discontinuous entities. As shown in Example 1, “呼吸道 (respiratory tract)” actually represents “呼吸道感染 (respiratory tract infection)” according to its successive “尿路感染 (urinary tract infection),” and the “呼吸道感染 (respiratory tract infection)” consists of two parts, i.e., “呼吸道 (respiratory tract)” and “感染 (infection)”.

The self-attention mechanism is usually used to learn word dependency in a sentence and capture the internal structure of the sentence. Compared with the convolutional neural network (CNN) and the recurrent neural network (RNN), the self-attention ignores the distance between words and directly calculates word dependency, which is very effective for both long-distance and local dependencies. More important, it has much faster computation speed and fewer parameters than RNN. At present, the self-attention mechanism has been successfully applied in many

NLP tasks, including reading comprehension [3], text summarization [4], machine translation [4], language inference [7, 8], relation extraction [9], and semantic role labeling (SRL) [10]. Different from the previous studies, to the best of our knowledge, our work is the first to apply the self-attention mechanism to the joint model for entity recognition and relation extraction.

In this article, we propose a joint neural network model with the self-attention mechanism for medical entity classification/recognition and relation extraction from medicine Chinese instructions. The contributions of our work are summarized in three parts.

- (1) We manually annotate 400 pieces of medicine Chinese instructions according to carefully formulated guideline and scheme and construct a corpus containing 5,505 medical entities and 5,512 semantic relations.
- (2) We label the discontinuous entity taking the Chinese characteristic of medical instruction into consideration and propose a joint model to process the discontinuous entity automatically.
- (3) The self-attention mechanism is integrated into our joint model to learn word dependency, especially between two medical entities.

The remainder of this article is organized as follows. Section 2 reviews related work. Section 3 describes the dataset constructed. Section 4 discusses the proposed joint model with the self-attention mechanism. Section 5 then presents experiments and discussions. Finally, Section 6 concludes the article.

## 2 RELATE WORK

The work related to EC/NER and RE, joint learning models, and self-attention in healthcare and medical domain will be described in detail in this section.

In the early stage, medical EC/NER and RE were mainly based on rules or dictionaries [11, 12]. At present, the mainstream methods are based on machine learning with artificial features. Frunza et al. [13] investigated on three types of disease-treatment relations, i.e., cure, prevention, and side effects, in the Medline abstract dataset and support vector machines– (SVM) based classification model with the artificial semantic features via the external medical knowledge resource. Liu et al. [14] adopted a  $k$ -nearest neighbor– (KNN) based model to extract relations from medicine Chinese instructions, and the clue words about the corresponding relations had been introduced to the model. These models heavily rely on complex artificial features and NLP tools, which have poor portability and increase the complexity and training difficulty.

With the rise of deep learning, some researchers have turned to deep neural networks for EC/NER and RE. Jagannatha and Yu [15] put forward a BiLSTM with a CRF model to identify clinical events. Lin et al. [16] completed the relation extraction task based on CNN using the attention mechanism at the sentence level. Quan et al. [17] constructed a multi-channel CNN model for RE by introducing into the convolution channel the word vectors pre-trained from PubMed, PMC, Medline, and Wikipedia.

All the above studies have regarded EC/NER and RE as two independent tasks in a pipeline model. The joint models have recently attracted more and more attention, because they can solve the problems mentioned in the Introduction well. Yang and Claire [18] presented a joint model that sought a globally optimal solution from the optimal results of subtasks. Singh et al. [19] proposed a single joint graphical model to represent the various dependencies between subtasks. Zheng et al. [20] used the shared BiLSTM to encode the sentence and adopted an LSTM for NER and a CNN for RE. Gupta et al. [21] proposed a method that relied on RNNs but used a lot of hand-crafted features and additional NLP tools to extract features like POS tags. Adel and Schütze [22] solved the EC and RE tasks using an approximation of a global normalization objective, i.e., CRF,

and they replicated the context of the sentence including the left and right parts of the entities to feed one entity pair once a time to a CNN for relation extraction. Pershina et al. [23] proved that the relation extractors trained with distant supervision can benefit significantly from a small number of manually labeled examples. Fu et al. [24] learned unified representation of relations via multi-task learning between multiple relation datasets and obtained significant improvement [25]. They also adopted a domain adversarial neural network to learn cross-domain features and obtained improvement on all three test domains at ACE-2005. Chan et al. [26] presented a system for rapidly customizing event extraction capability to find new event types and their arguments. The aforementioned works have not simultaneously inferred other potential entities and relations within the same sentence and only completed the EC and RE tasks.

Miwa and Bansal [27] introduced the dependency information into the RE task through parameter sharing. They classified the relations according to their shortest paths in the dependency trees. Li et al. [28] also proposed a model with tree-LSTMs to learn the dependency information for entity and relation extraction from the biomedical texts. Katiyar and Cardie [29] and Bekoulis et al. [30] investigated attention-based RNNs for extracting relations between entity mentions without using any dependency parse tree feature. Zhang et al. [31] established an end-to-end relationship extraction model based on global optimization, and used new LSTM features to better represent context. Zheng et al. [32] employed a new annotation strategy to transform the joint task into a sequence labeling problem. However, all of them assumed that the relations are mutually exclusive. There are limitations with these approaches. First, an entity can only participate in one relation. Second, the time complexity of entity recognition is increased compared to the standard approaches that are of linear complexity. In our constructed dataset, we find that the entity often corresponds to multiple relationships. We solve this problem by casting relation extraction as multi-head prediction.

Vaswani et al. [4] introduced self-attention to neural machine translation (NMT) and achieved state-of-the-art performance. They also proposed multi-head attention based on self-attention. Paulus et al. [5] exploited the distance between relative positions or sequence elements to expand self-attention, which effectively improved the translation quality and efficiency. Tan et al. [10] used self-attention in the SRL task and achieved state-of-the-art performance on the CoNLL-2005 and CoNLL-2012 datasets. Verga et al. [9] combined self-attention with the biomedical relation extraction. They proposed a document-level model and obtained the significant experimental results in the dataset of chemical disease relations (CDR). In this article, we put forward the joint model with the self-attention mechanism to commit medical EC/NER and RE simultaneously from the medicine Chinese instructions. Lin et al. [7] used self-attention to learn the embedding of a sentence, expressed as a two-dimensional matrix instead of a vector where each line in the matrix represented different parts of the sentence. Cheng et al. [3] applied self-attention to reading comprehension, solving the limitations of traditional RNN for processing the inherently structured input. Shaw et al. [6] introduced self-attention mechanism to the relative position representation between elements, and the relative position representation of self-attention achieved a significant improvement over absolute position encoding of self-attention in two machine translation tasks. Shen et al. [8] adopted self-attention to natural language reasoning and made the state-of-the-art performance on multiple datasets.

### 3 DATASET CONSTRUCTION

#### 3.1 Medicine Instruction

In the healthcare and medical domain, there are already some datasets in English, e.g., ADE and DREC, and many studies have achieved good performances on these datasets. However, they are

not publicly available, and there are no available domain-specific Chinese datasets. In this work, we construct a new linguistic corpus with medicine Chinese instructions.

The medicine instruction is a document that contains the important information about the medicine. It is an important guiding document for clinicians and patients to use medicine safely and effectively, which is of great significance in diagnosis and treatment of medicine. The description of the antibacterial medicine usually contains dense diseases, bacteria, and medicines with the clear and identifiable boundaries, and the relationship between entities is much easier to judge, so we select the antibacterial medicine instructions as the language material in our dataset. A total of 1,053 antibacterial medicine Chinese instructions have been crawled from Chinese websites,<sup>2</sup> and 400 medicine instructions with rich semantic information and clear structure are chosen to construct the medicine Chinese instruction corpus.

### 3.2 Annotation Guideline

The annotation of medical entity and relationship between entities usually involves semantic understanding expertise. Under the guideline put forward by medical experts and iterations via annotation practices, we have formulated the complete medical entity and relationship annotation guideline and system.

Referring to the medical glossary [33] and the suggestions from medical experts, we develop the following annotation guideline.

- (1) Our corpus contains seven types of medical entities, i.e., disease, symptom, medicine, body-region, patient, bacteria, and treatment. There are nine kinds of medical entity relations in our corpus, i.e., cure, hyponymy, causality, anaphora, meronymy, prevention, inhibition, sensitivity, and drug-resistance.
- (2) The longest principle is adopted to ensure the semantic integrity of a medical entity. If there exists more than one type of a medical entity in a text span, then the longest one is retained according to the longest principle and the other shorter are discarded.
- (3) A medical entity may consist of two discontinuous parts. We refer to this kind of the medical entity as the discontinuous entity and annotate two parts as a single medical entity with the same entity type.
- (4) After entity annotation, the semantic relationship between specific medical entities is represented by a structured triple (former-entity, relation-type, later-entity). For the entity combination rule, the structured triple complies with the order of linguistic reading of semantic relationship.

Example 2: 注射用美洛西林钠适应症: 用于大肠埃希菌、肠杆菌属、变形杆菌等革兰阴性杆菌中敏感菌株所致的呼吸系统、泌尿系统、消化系统、妇科和生殖器官等感染, 如败血症、化脓性脑膜炎、腹膜炎、骨髓炎、皮肤及软组织感染及眼、耳、鼻、喉科感染。

*Mezlocillin sodium injection: This product is suitable for respiratory, urinary, digestive, gynecological, reproductive, skin, soft tissue, eye, ear, nose or throat infections caused by the sensitive strains of gram-negative bacilli, such as Escherichia coli, enterobacter, and proteus, and these infections include sepsis, purulent meningitis, peritonitis, and osteomyelitis.*

In Example 1, the string “谷氨酸诺氟沙星注射液 (norfloxacin glutamate injection)” is labeled as a medicine entity altogether according to our annotation guideline (2). Only when there exists the string “谷氨酸诺氟沙星 (norfloxacin glutamate)” in the medicine instruction, we will mark “谷氨酸诺氟沙星” as a medicine entity. In Example 2, there exists an abstract or concrete disease

<sup>2</sup>[http://top.chinaz.com/site\\_www.yiwan.cn.html](http://top.chinaz.com/site_www.yiwan.cn.html).

Table 1. Medical Entities in Example 2

Entity type	Entity
medicine	mezlocillin sodium injection
specific disease	sepsis; purulent meningitis; peritonitis; osteomyelitis
abstract disease (infection)	respiratory; urinary; digestive; gynecological; reproductive; skin; soft tissue; eye; ear; nose; throat
bacteria	sensitive strains; gram-negative bacilli; <i>E. coli</i> ; enterobacter; proteus

about body-region, e.g., “喉科感染 (throat infection),” we only annotate the disease “喉科感染 (throat infection)” according to our annotation guideline (2) and the medical entity of body-region “喉科 (throat)” is discarded. In Example 2, “皮肤及软组织感染” in Chinese actually means “skin and soft tissue infection” in English. So, the two diseases “皮肤感染 (skin infection)” and “软组织感染 (soft tissue infection)” are labeled according to our annotation guideline (3), and they share the common linguistic component, i.e., “感染 (infection).” Afterward, the semantic relationship is represented by the structured triple, e.g., (*Escherichia coli*, causality, digestive infection). The former entity indicates the reason, and the latter one indicates the result in this causality relation, according to our annotation guideline (4).

In Example 2, as shown in Table 1, there are 20 medical entities, including one medicine, four specific diseases, 11 abstract body-region infections, and five bacteria in a single Chinese sentence. Among these entities, we can find a number of relationships, e.g., the cure relationships among medicine mezlocillin sodium injection and all diseases, the causality relationships among all bacteria and diseases, the hyponymy relationships among gram-negative bacilli and *E. coli*, enterobacter, proteus, the hyponymy relationships among sensitive strains and *E. coli*, enterobacter, and proteus.

### 3.3 Dataset

For each medicine Chinese instruction, five annotators manually annotate the entity, the entity type, the semantic relation, and the relation type independently according to the annotation guidelines. We adopt the majority principle, i.e., the minority subordinating to the majority, to settle the different opinions among annotators at the entity or the relation level. When the number of annotators holding two different opinions is close on one instance, i.e., 3 to 2, we invite another medical expert as the arbitrator to confirm reliability and consistency of our annotation.

The annotation results for 400 medicine instructions are shown in Table 2. In total, we have annotated 5,505 entities and 5,512 relationships. The dataset is available via the github<sup>3</sup> link.

## 4 MODEL

In this section, we explain our model, as shown in Figure 1, in detail. We develop a joint model based on BiLSTM to (1) identify the boundary and the type of the medical entity and (2) extract semantic relations among entities. The model is divided into five layers from bottom to top: (1) an embedding layer, (2) a self-attention layer, (3) a BiLSTM layer, (4) a NER layer, and (5) a sigmoid layer.

The input to the model is a sequence of tokens, which are then converted into the corresponding embedding vectors. The self-attention and BiLSTM layers are used to extract deep contextual information. Afterward, the NER and sigmoid layers perform the NER and RE tasks, respectively.

<sup>3</sup><https://github.com/zhangyukun230/medical-dataset>.



Table 2. Statistics of Medical Entities and Relations in Medicine Chinese Instruction Corpus

Entity	Amount	Relation	Amount
medicine	1,046	cure	1,734
disease	2,914	hyponymy	1,273
bacteria	1,280	causality	1,742
symptom	32	anaphora	254
patient	141	meronymy	7
body-region	37	prevention	29
treatment	55	inhibition	148
<b>all</b>	<b>5,505</b>	sensitivity	293
		drug-resistance	28
		<b>all</b>	<b>5,512</b>

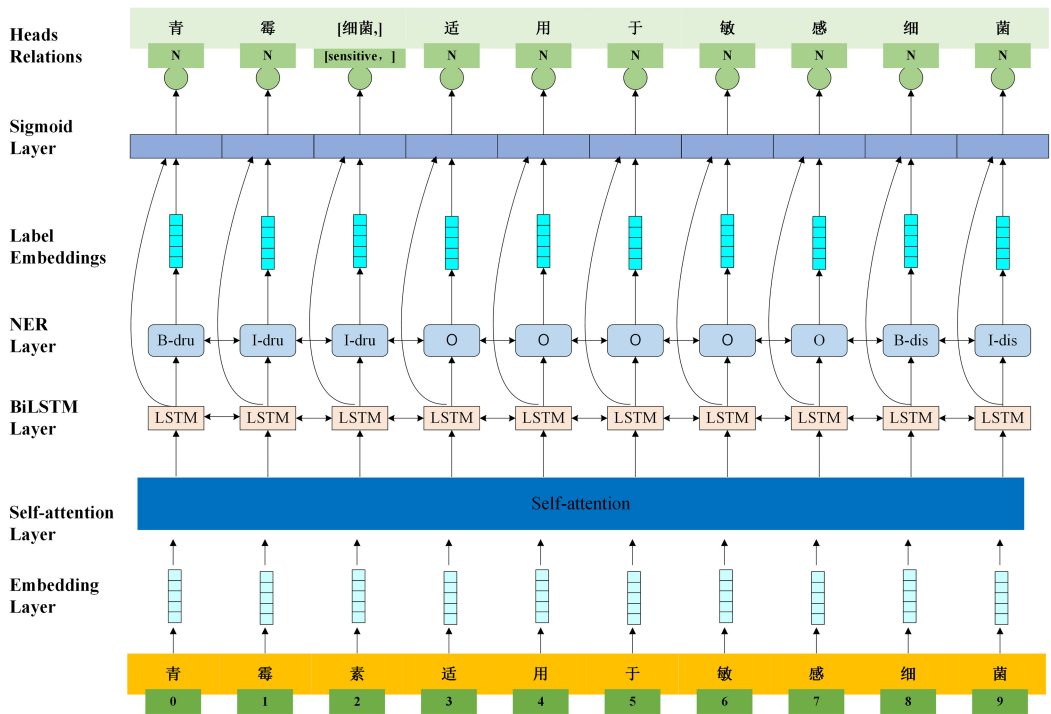


Fig. 1. Overview of our joint model. The model consists of five layers, i.e., embedding, multi-head attention, BiLSTM, CRF for NER, and sigmoid.

The output for each input token has two parts: (1) a recognized entity label, e.g., I-drug, which denotes the token inside a named entity of drug, and (2) a set of tuples comprising the head tokens of the entity and the types of relations between them, e.g., {(细菌 (*bacteria*), sensitive)}. An entity often contains more than one token, and we only consider the last token of the entity for multi-head selection. For example, there is a sensitive relation between the entities “青霉素 (*penicillin*)” and “细菌 (*bacteria*).” Instead of connecting all tokens of the entities, we only link “素” with “菌”.

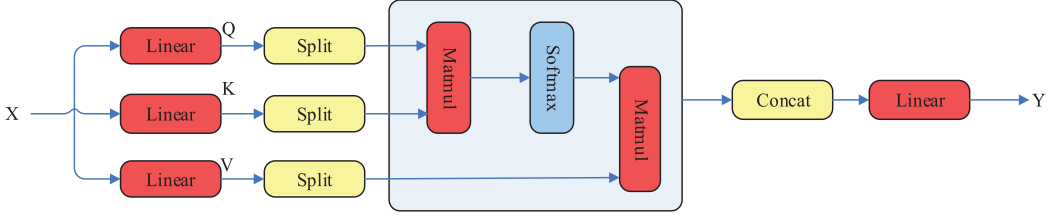


Fig. 2. The computation graph of multi-head attention mechanism.

Moreover, if there is no semantic relation between the two entities, then we assign the label “N” and regard the token itself as the head.

#### 4.1 Embedding Layer

Given a sentence  $X = (x_1, x_2, \dots, x_n)$  as a sequence of tokens, the embedding layer transforms these tokens into the vector matrix by pre-trained embedding so that they can be calculated in a neural network model.

#### 4.2 Self-attention Layer

Self-attention is a specific case of the attention mechanism that only requires a single sequence to compute its representation. Following Vaswani et al. [4], we use multi-head attention based on self-attention in this article. Figure 2 depicts the computation graph of the multi-head attention mechanism.

Given a matrix of  $n$  query vectors  $Q \in \mathbb{R}^{n \times d}$ , keys  $K \in \mathbb{R}^{n \times d}$ , and values  $V \in \mathbb{R}^{n \times d}$ , the scaled dot-product attention computes the attention scores with the following formula (1):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad (1)$$

where  $Q = K = V$  in the self-attention layer.

For multi-head attention,  $X$  is converted to  $(Q, K, V)$  through different linear elements. Then,  $h$  parallel heads are employed to focus on a different part of channels of value vectors. We denote the learned linear maps as  $W_i^Q \in \mathbb{R}^{n \times d/h}$ ,  $W_i^K \in \mathbb{R}^{n \times d/h}$ ,  $W_i^V \in \mathbb{R}^{n \times d/h}$ , which correspond to queries, keys, and values, respectively. For the  $i$ th head, the scaled dot-product attention is calculated, and the mathematical formulation is shown as follows:

$$M_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V). \quad (2)$$

Finally, all vectors produced by parallel heads are concatenated together to form a single vector. A linear map is used to mix different channels from different heads:

$$M = \text{Concat}(M_1, M_2, \dots, M_h), \quad (3)$$

$$Y = MW, \quad (4)$$

where  $M \in \mathbb{R}^{n \times d}$ ,  $W \in \mathbb{R}^{d \times d}$ , and  $Y$  is the output of self-attention layer and also the input of BiLSTM layer.

#### 4.3 BiLSTM Layer

LSTM has the ability to well capture long-term dependencies and its extension BiLSTM can make use of the contextual information in both past and future. Therefore, BiLSTM models have achieved excellent results in many NLP tasks. In this work, we adopt a BiLSTM to encode the text of medicine



Chinese instruction. We empirically set the number of LSTM layers to 3. For each vector  $y_i$  of  $Y$ , we concatenate the forward output ( $\vec{h_i}$ ) and the backward output ( $\overleftarrow{h_i}$ ) at the timestep  $i$ . The entire output of BiLSTM at the timestep  $i$  is then as follows:

$$h_i = \left[ \left( \vec{h_i} \right), \left( \overleftarrow{h_i} \right) \right]; i = 0, 1, \dots, n.. \quad (5)$$

#### 4.4 NER Layer

Similarly to Miwa and Bansal [27], Zheng et al. [32], and Katiyar and Cardie [29], we formulate the NER task as a sequence labeling problem. In sequence labeling, an entity usually contains more than one token and one tag is assigned to each token in the entity. We adopt the BIO tagging scheme, i.e., using B, I, and O to denote the beginning, inside, and outside of the entity, respectively. We annotate the first character of an entity as B-type, the other tokens of the entity as I-type, and those non-entity tokens as O. With the BIO tagging scheme, we can identify the boundary and the type of entity simultaneously. In Figure 1, the Chinese token “青” has been tagged as B-drug, “霉” and “素” are tagged as I-drug, and “适” is tagged as O. The scores of the tags are calculated by the following formula:

$$s^{(e)}(h_i) = V^{(e)} f(W^{(e)} h_i + b^{(e)}) V \in \mathbb{R}^{l \times n(entity)}, W \in \mathbb{R}^{l \times 2d}, b \in \mathbb{R}^l, \quad (6)$$

where  $e$  is the symbol denoting NER,  $f(\cdot)$  is the activation function such as relu or tanh,  $d$  is the number of LSTM hidden units,  $n(entity)$  is the number of NER tags, and  $l$  is the length of input  $w$ .

In the NER task, the BIO labeling scheme faces several restrictions. For example, B-drug cannot be followed by I-disease. In the softmax, the tags assigned to the tokens according to their probabilities are independent of each other, although BiLSTM is able to learn the contextual information. CRF has a transition characteristic that takes the order among tags into consideration. Therefore, we use a linear-chain CRF (Lample et al. [34]) for the NER task. For input token  $w_i$ , the score sequence can be expressed as follows:

$$S(y_1^{(e)}, \dots, y_n^{(e)}) = \sum_{i=0}^n s_{i, y_i^{(e)}}^{(e)} + \sum_{i=1}^{n-1} T_{i, y_i^{(e)}}, \quad (7)$$

where  $s_{i, y_i^{(e)}}^{(e)}$  represents the tag score when the tag of  $w_i$  is  $y_i$ ,  $T$  denotes the transition score from tag  $y_i$  to  $y_{i+1}$ , and  $T \in \mathbb{R}^{(p+2, p+2)}$ . Then, the probability of a given sequence of tags over all possible tag sequences for the input sentence  $w$  can be defined by the following formula (7):

$$\Pr(y_1^{(e)}, \dots, y_n^{(e)} | w) = \frac{e^{s(y_1^{(e)}, \dots, y_n^{(e)})}}{\sum_{\tilde{y}_1^{(e)} \dots \tilde{y}_n^{(e)}} e^{s(\tilde{y}_1^{(e)}, \dots, \tilde{y}_n^{(e)})}}, \quad (8)$$

where we adopt Viterbi to find the tag sequence  $\tilde{y}_i^{(e)}$ , which has the highest score.

Entity tags are then fed into the RE task as the corresponding label embedding, considering that the types of entities involved benefit relation prediction. Our model takes the label embedding as a parameter layer and learns it in the training phase. The input of the sigmoid layer is the concatenation of the BiLSTM state  $h_i$  and the label embedding  $g_i$  of  $x_i$ ,

$$z_i = (h_i, g_i). \quad (9)$$

#### 4.5 Sigmoid Layer

Inspired by the work of Bekoulis et al. [2], we regard the RE task as a multi-head selection problem, which can successfully solve the problem with most existing joint models that cannot recognize all semantic relations for a specific entity in a sentence. For token  $w_i$ , it may hold any type of

semantic relation  $r_k$  with other token  $w_j$ . In this article, we make the assumption that the semantic relations between one entity and others are independent.

Given a sequence of tokens  $w$  and a set of relation labels  $R$ , the goal of our model is to identify the vector of the most probable heads  $w_j$  and the vector of the most probable corresponding relation label  $r_k$  for each token  $w_i$ . We calculate the score between tokens  $w_i$  and  $w_j$  given a label  $r_k$  as follows:

$$s^{(r)}(w_i, w_j, r_k) = V^r f \left( W_1^{(r)} z_i + W_2^{(r)} z_j + b^{(r)} \right), V \in \mathbb{R}^{l \times n(relation)}, W \in \mathbb{R}^{l \times 2d+e}, b \in \mathbb{R}^l, \quad (10)$$

where  $r$ ,  $V^r$ ,  $d$ ,  $e$ , and  $l$  represent the symbol of RE, the weight matrix, the number of LSTM hidden units, the dimension of label embedding, and the length of input  $w$ , respectively.  $f(\cdot)$  and  $n(relation)$  denote the activation function, relu or tanh, and the number of RE tags. The probability of the token  $w_j$  to be selected as the head of the token  $w_i$  with the relation label  $r_k$  between them is defined as

$$\Pr(head = w_j, label = r_k | w_i) \sigma \left( s^{(r)}(w_i, w_j, r_k) \right), \quad (11)$$

where  $\sigma$  is the sigmoid function. Unlike softmax, sigmoid assumes that all relations are independent of each other, and it does not add up all probabilities of relations to 1. If the value of  $\Pr$  is more than 0.5, then sigmoid will determine whether there is a semantic relation between the two entities; otherwise, no relation exists, and it marks “N.” The goal of sigmoid is to minimize the cross-entropy loss  $\mathcal{L}_{re}$ . during the training phase,

$$\mathcal{L}_{re} = \sum_i^n \sum_j^m -\log \Pr(head = y_{i,j}, re = r_{i,j} | w_i). \quad (12)$$

For the joint entity and relation extraction task, we calculate the final objective  $\mathcal{L} = \mathcal{L}_{re} + \mathcal{L}_{ner}$ . We set the weights of  $\mathcal{L}_{re}$  and  $\mathcal{L}_{ner}$  as the hyper-parameters in our experiments and find that the performance is optimal when both the weights of them are 1.

## 5 EXPERIMENTS

### 5.1 Datasets and Settings

From Table 2, we can find that the medical entities and relationships in the medicine instructions are distributed unevenly. In this work, we focus on the entities of medicine, disease, bacteria, patient, and treatment, which hold the almost even distribution in the dataset. Meanwhile, this article only investigates the semantic relations of cure, hyponymy, causality, inhibition, and sensitivity. Thirty-two pieces of medical instructions, which do not contain any of the five types of semantic relations mentioned above, are removed from the data collection.

There are 839 discontinuous entities in our constructed dataset, accounting for approximately 17.90% of all annotated medical entities. In Example 3, we use “尿路 (urinary tract),” “呼吸道 (respiratory tract),” and “皮肤 (skin)” to represent the discontinuous entities “尿路感染 (urinary tract infection),” “呼吸道感染 (respiratory tract infection),” and “皮肤 (skin infection).” “尿路 (urinary tract),” “呼吸道 (respiratory tract),” and “皮肤 (skin)” are labelled as diseases rather than body-regions. In our model, “阿莫西林克拉维酸钾颗粒 (amoxicillin and clavulanate potassium granules)” and “尿路 (urinary tract)” form a cure relationship, with the underlying contextual information guidance. Our joint model can recognize “尿路 (urinary tract)” as a disease. In addition, the entity type information will be converted to label embedding and conveyed to the RE task.

Table 3. Statistics of Medical Entities

Entity	Disease	Bacteria	Medicine	Patient	Treatment	All
<b>training</b>	2,118	784	723	63	33	3,721
<b>test</b>	515	239	176	26	11	967
<b>all</b>	2,633	1,023	899	89	44	4,688

Table 4. Statistics of Relations among Medical Entities

Relation	Hyponymy	Cure	Causality	Inhibition	Sensitivity	All
<b>training</b>	908	1,301	1,262	108	216	3,795
<b>test</b>	285	258	321	34	61	959
<b>all</b>	1,193	1,559	1,583	142	277	4,754

Example 3: 阿莫西林克拉维酸钾颗粒适应症:本品适用于克雷伯菌属所致的呼吸道、尿路和皮肤及软组织感染等;

*Amoxicillin and clavulanate potassium granules: This product is suitable for respiratory tract urinary tract and skin soft tissue infection caused by Klebsiella.*

In the end, a total of 368 medicine Chinese instructions have been selected. Eighty percent of them are used for training and the remaining 20% for testing. The statistics of the entities and relations in the training set and the test set are presented in Tables 3 and 4, respectively. Besides our constructed dataset, we also use the CoNLL-2004 dataset [35] to evaluate the generalization of our model.

For comparison, we build a baseline model, which consists of four layers, i.e., embedding, BiLSTM, CRF, and sigmoid. Compared to our model, the baseline model removes the self-attention layer. In this way, we can not only explore the advantages of the joint model in relation to the other works but also understand how self-attention improves the experimental results.

In this article, for Precision, Recall, and F1-score, we also use two different evaluation standards, i.e., strict and relaxed, to be able to compare our experimental results with previous studies.

- (1) Strict: An entity is considered correct if both the boundaries and the type of the entity are correct.
- (2) Relaxed: We regard a multi-token entity correct if at least one of its composite token types is correct.

For CoNLL-2004, we use the embedding vectors in the previous work [21, 22]. We adopt the same splits as in Gupta et al. [21] to do the EC/RE task and evaluate our model on the NER/RE task similar to Miwa and Sasaki [36], Bekoulis et al. [2] on the same dataset using 10-fold cross validation. We also perform the same two tasks mentioned above on our own constructed dataset. We use pre-trained embedding vectors by word2vec based on Wikipedia. The dimension of the vectors is 50.

The hyper-parameters of the model are shown in Table 5. The parameters of baseline and our model are the same except for h-head. We use Adam algorithm to optimize the parameters, and the number of parallel heads  $h$  is 2. We commit dropout after every layer and fix the hyper-parameters, i.e. dropout values, best epoch, and learning rate, on the validation and test datasets.

Table 5. The Hyper-parameter Settings

Hyper-parameter	Optimizer	Activation	Learning rate	Dropout	LSTM unit	LSTM layers	Label embedding	h-head
our dataset	Adam	tanh	0.003	0.9	64	3	20	2
CoNLL-2004	Adam	tanh	0.003	0.8	64	3	25	2

Table 6. The Experimental Results

Dataset	Model	Method	F	Standard	Entity	Relation	Overall
EC/RE							
CoNLL-2004	Kate and Mooney (2010)	P	✓	Relaxed	91.36	66.28	78.82
	Kate and Mooney (2010)	J	✓	Relaxed	91.70	66.36	79.03
	Gupta et al. (2016)	P	✓	Relaxed	91.00	97.1	79.05
	Gupta et al. (2016)	J	✓	Relaxed	92.40	69.90	81.15
	Gupta et al. (2016)	J	×	Relaxed	88.80	58.30	73.60
	Adel and Schütze (2017)	J	×	Relaxed	82.10	62.50	72.30
	Bekoulis et al. (2018)	J	×	Relaxed	93.26	67.01	80.14
	Our model	J	×	Relaxed	<b>94.65</b>	<b>69.39</b>	<b>82.03</b>
our dataset	Baseline	J	×	Relaxed	97.39	73.93	85.65
	Our model	J	×	Relaxed	<b>97.87</b>	<b>75.88</b>	<b>86.87</b>
	Our model -LE	J	×	Relaxed	96.13	74.98	85.56
NER/RE							
CoNLL-2004	Miwa and Sasaki (2014)	J	✓	Strict	80.70	61.00	70.85
	Bekoulis et al. (2018)	J	×	Strict	83.04	61.04	72.04
	Our model	J	×	Strict	<b>85.26</b>	<b>62.59</b>	<b>73.93</b>
our dataset	Baseline	J	×	Strict	91.80	72.43	82.12
	Our model	J	×	Strict	<b>93.25</b>	<b>73.57</b>	<b>83.41</b>
	Our model -LE	J	×	Strict	92.45	71.79	82.12

The models are as follows: (i) baseline for NER/RE, (ii) baseline for EC (predicting only entity classes)/RE, (iii) our model for NER/RE, and (iv) our model for EC/RE. The ✓ and × symbols indicate whether the models rely on external NLP tools. We include different evaluation standards (Strict and Relaxed). P means pipeline method and J means joint model. “-LE” indicates that label embedding is removed from our model.

## 5.2 Experimental Results

The experimental results are shown in Table 6. On the two datasets we use both strict and relaxed standards. In the relaxed setting, we perform an EC task instead of NER, assuming that the boundaries of entities are given.

For the EC task on the CoNLL-2004 dataset, due to a large number of experimental results, we present the overall F1-score of different experiments in the form of histogram. We can see the details in Figure 3, where the symbols “✓” and “×” mean whether the models combine the artificial features.

Compared with the experiments of Kate and Mooney [37] and Gupta et al. [21], we can find that the joint model performs better than the pipeline model under the same conditions. This verifies the advantages of the joint model. The 93.25% F1-score of NER task on our dataset verifies that our annotation strategy and joint model are very effective.

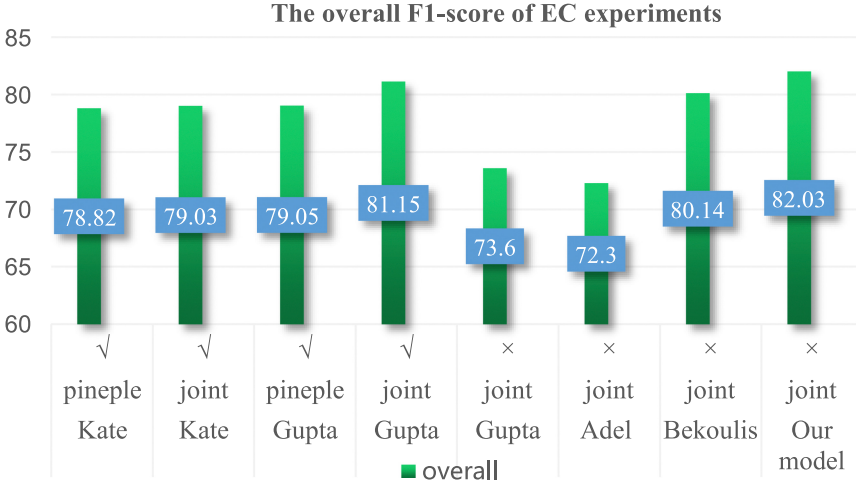


Fig. 3. The overall F1-score of EC experiments.

Since the pipeline model cannot learn the semantic relevance between the entity and relation extraction tasks, it brings in the entity extraction errors and transfers this type of errors to the relation extraction. For Example 4, the pipeline model of the NER task detects medical entities with the bacteria entity type like “肺炎衣原体 (mycoplasma pneumoniae),” “沙眼衣原体 (chlamydia trachomatis),” and “人型支原体 (mycoplasma humanity).” In subsequent relation extraction, the causality relations between them are naturally misidentified. With our joint model, all these entities and relations are extracted correctly.

**Example 4:** 注射用阿奇霉素适应症: 本品适用于敏感致病菌株所引起的下列感染。1. 由肺炎衣原体、流感嗜血杆菌、嗜肺军团菌、卡他摩拉菌、肺炎支原体、金黄色葡萄球菌或肺炎链球菌引起的需要首先采取静脉滴注治疗的社区获得性肺炎。2. 由沙眼衣原体、淋病奈瑟菌、人型支原体引起的需要首先采取静脉滴注治疗的盆腔炎。

*Azithromycin injection: This product is suitable for the following infections caused by sensitive pathogenic strains. 1. Community acquired pneumonia caused by Chlamydia pneumoniae, Haemophilus influenzae, Legionella pneumophila, Moraxella catarrhalis, Mycoplasma pneumoniae, Staphylococcus aureus, or Streptococcus pneumoniae, which needs to be treated by intravenous drip first. 2. Pelvic inflammation caused by Chlamydia trachomatis, Neisseria gonorrhoeae, and Mycoplasma hominis, which needs to be treated by intravenous drip first.*

#### Entity:

medicine: 注射用阿奇霉素 (azithromycin for injection); 本品 (this product).

treatment: 静脉滴注治疗 (intravenous drip); 静脉滴注治疗 (intravenous drip).

disease: 感染 (infections); 社区获得性肺炎 (community acquired pneumonia); 盆腔炎 (pelvic inflammation).

#### Pipeline model:

bacteria: 流感嗜血杆菌 (*Haemophilus influenzae*); 嗜肺军团菌 (*Legionella pneumophila*); 卡他摩拉菌 (*Moraxella catarrhalis*); 金黄色葡萄球菌 (*Staphylococcus aureus*); 肺炎链球菌 (*Streptococcus pneumoniae*); 淋病奈瑟菌 (*Neisseria gonorrhoeae*).

#### Our joint model:

bacteria: **肺炎衣原体** (*Chlamydia pneumoniae*); 流感嗜血杆菌 (*Haemophilus influenzae*); 嗜肺军团菌 (*Legionella pneumophila*); 卡他摩拉菌 (*Moraxella catarrhalis*); **肺炎支原体**

Table 7. The Comparative Experimental Results with Our Model and Bekoulis et al. (2018)

Dataset	Model	Entity	Relation
CoNLL-2004	Bekoulis et al. (2018)	896	257
CoNLL-2004	our model	920	264
our dataset	Bekoulis et al. (2018)	888	695
our dataset	our model	902	706

(*Mycoplasma pneumoniae*); 金黄色葡萄球菌 (*Staphylococcus aureus*); 肺炎链球菌 (*Streptococcus pneumoniae*); 沙眼衣原体 (*Chlamydia trachomatis*); 淋病奈瑟菌 (*neisseria gonorrhoeae*); 人型支原体 (*mycoplasma hominis*).

#### Relation:

cure: (注射用阿奇霉素, 感染); (注射用阿奇霉素, 社区获得性肺炎); (注射用阿奇霉素, 盆腔炎).

#### Pipeline model:

causality: (流感嗜血杆菌, 社区获得性肺炎); (嗜肺军团菌, 社区获得性肺炎); (卡他摩拉菌, 社区获得性肺炎); (肺炎支原体, 社区获得性肺炎); (金黄色葡萄球菌, 社区获得性肺炎); (肺炎链球菌, 社区获得性肺炎); (淋病奈瑟菌, 盆腔炎); (致病菌株, 感染).

#### Our joint model:

causality: (肺炎衣原体, 社区获得性肺炎); (流感嗜血杆菌, 社区获得性肺炎); (嗜肺军团菌, 社区获得性肺炎); (卡他摩拉菌, 社区获得性肺炎); (肺炎支原体, 社区获得性肺炎); (金黄色葡萄球菌, 社区获得性肺炎); (肺炎链球菌, 社区获得性肺炎); (淋病奈瑟菌, 盆腔炎); (致病菌株, 感染); (人型支原体, 盆腔炎); (沙眼衣原体, 盆腔炎).

Compared with the works of Gupta et al. [21] and Adel and Schütze [22], our model achieves about 6% improvement on both EC and RE tasks, even without complex hand-crafted features. Moreover, compared with the model of Gupta et al. [21] that relies on complex hand-crafted features, our model also improves about 1%. For the NER task, we improve overall F1-score with about 3% and 2% compared to the works of Miwa and Sasaki [36] and Bekoulis et al. [2]. This indicates that our model has better performance on the CoNLL-2004 dataset. It suggests that our neural network model can obtain more semantic information than the feature-based model or the existing neural network model. Unlike the prior studies that obtain the entity types and the corresponding relations from the pairs of entities, our model takes the entire sentence as input, rather than just the entity pairs.

In NER/RE task, Comparing with the work of Bekoulis et al. (2018) [2] on the CoNLL-2004 dataset, our model makes around 2% overall improvement. On our constructed dataset, our model achieves more than 1% overall improvement in experiments. The comparative experimental results with our model and Bekoulis et al. (2018) [2] are shown in Table 7.

From Table 7, we can find that more 24 entities and 7 relations are identified from CoNLL-2004 dataset via our model comparing with the model of Bekoulis et al. (2018) [2], and our model also recognizes additional 14 entities and 11 relations from our constructed dataset. Our model has designed one more self-attention layer comparing with Bekoulis et al. (2018) [2], which can learn the relationship between entities in the sentence or instruction globally. In fact, our model has learned word dependencies in the sentence or instruction via the self-attention mechanism, especially between the entities. Taking “青霉素适用于敏感细菌 (penicillin is suitable for sensitive bacteria)” as an example, and its self-attention weight map is shown in Figure 4. In our experiment, the “素” and “菌” represent the last tokens of the entities “青霉素” and “细菌”, respectively. The



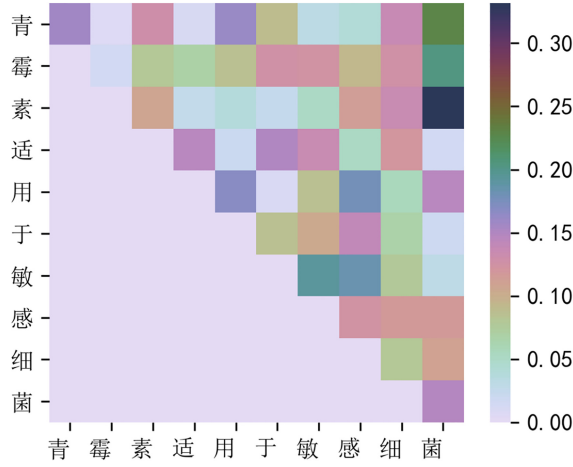


Fig. 4. The weight map of “青霉素适用于敏感细菌 (penicillin is suitable for sensitive bacteria)” by our model with the self-attention mechanism.

weight by our model between the “素” and “菌” is the maximum in Figure 4, denoting a sensitive relationship between the two entities “青霉素” and “细菌”. This verifies that the self-attention mechanism can strengthen the connection between the related entities “青霉素” and “细菌” via learning entity dependencies.

In Example 5, the model proposed by Bekoulis et al. (2018) [2] fails to recognize the disease entity “上” and its relationships. But our joint model identifies the entity “上” and its related relations, including (肺炎球菌, causality, 上), (金黄色葡萄球菌, causality, 上), (表皮葡萄球菌, causality, 上), (肺炎球菌, causality, 上), (金黄色葡萄球菌, causality, 上), (表皮葡萄球菌, causality, 上), and (乙酰吉他霉素片, cure, 上).

**Example 5:** 乙酰吉他霉素片适应症: 本品主要适应于革兰阳性菌所致的各种感染, 特别适应于金黄色葡萄球菌、肺炎球菌及表皮葡萄球菌引起的上、下呼吸道感染及表皮软组织感染。据文献报道, 本品对百日咳、猩红热、中耳炎等也有良好的疗效。

*Acetylkitasamycin tablet: This product is mainly suitable for various infections caused by gram-positive bacteria, especially for upper and lower respiratory tract infections caused by Staphylococcus aureus, pneumococcus, and S. epidermidis, as well as skin soft tissue infections. According to the literature, this product has good curative effect on pertussis, scarlet fever, otitis media and so on.*

#### Entity:

medicine: 乙酰吉他霉素片 (acetylkitasamycin tablets); 本品 (this product).

bacteria: 革兰阳性菌 (Gram-positive bacteria); 金黄色葡萄球菌 (*Staphylococcus aureus*); 肺炎球菌 (*pneumococcus*); 表皮葡萄球菌 (*S. epidermidis*).

#### Model of Bekoulis et al. (2018) [2]:

disease: 百日咳 (pertussis); 猩红热 (scarlet fever); 中耳炎 (otitis media); 感染 (infections); 表皮软组织感染 (soft tissue infections); 下呼吸道感染 (lower respiratory infections).

#### Our model:

disease: 百日咳 (pertussis); 猩红热 (scarlet fever); 中耳炎 (otitis media); 感染 (infections); 表皮软组织感染 (soft tissue infections); **上(upper)**; 下呼吸道感染 (lower respiratory infections).

**Relation:****Model of Bekoulis et al. (2018) [2]:**

causality: (表皮葡萄球菌, 表皮软组织感染); (肺炎球菌, 表皮软组织感染); (金黄色葡萄球菌, 下呼吸道感染); (肺炎球菌, 下呼吸道感染); (表皮葡萄球菌, 下呼吸道感染); (金黄色葡萄球菌, 表皮软组织感染).

cure: (乙酰吉他霉素片, 感染); (乙酰吉他霉素片, 表皮软组织感染); (乙酰吉他霉素片, 猩红热); (乙酰吉他霉素片, 百日咳); (乙酰吉他霉素片, 中耳炎); (乙酰吉他霉素片, 下呼吸道感染).

**Our model:**

causality: (肺炎球菌, 上); (金黄色葡萄球菌, 上); (表皮葡萄球菌, 表皮软组织感染); (肺炎球菌, 表皮软组织感染); (金黄色葡萄球菌, 下呼吸道感染); (肺炎球菌, 下呼吸道感染); (表皮葡萄球菌, 上); (表皮葡萄球菌, 下呼吸道感染); (金黄色葡萄球菌, 表皮软组织感染).

cure: (乙酰吉他霉素片, 感染); (乙酰吉他霉素片, 表皮软组织感染); (乙酰吉他霉素片, 猩红热); (乙酰吉他霉素片, 百日咳); (乙酰吉他霉素片, 中耳炎); (乙酰吉他霉素片, 上); (乙酰吉他霉素片, 下呼吸道感染).

Meanwhile, we find that the same model achieves better performance on our dataset than the CoNLL-2004 dataset. The reason is that medical entities have obvious character characteristics in the Chinese antibacterial medical instructions. We can see from Example 2 that the diseases usually end with the Chinese characters “炎” and “症”, e.g., “脑膜炎 (meningitis),” “骨髓炎 (osteomyelitis),” and “败血症 (Septicemia),” and the bacteria generally end with the Chinese character “菌”, e.g., “大肠埃希菌 (*Escherichia coli*)” and “肠杆菌 (*Enterobacter*).” Another interesting linguistic phenomenon is that the name of medicine usually appears at the beginning of a medicine instruction.

Finally, we verify the impact of label embedding (LE) on the experiment results. After implementing the model with the LE feature, we conduct a comparative experiment to uncover its significance. The experimental results are shown in Table 6. We can find that the performances of both EC and NER decrease by around 1% when the LE feature is removed. Therefore, we can make the conclusion that the category information of the medical entities is beneficial to relation extraction.

**5.3 Error Analysis**

The distributions of entity type and relation type are uneven, so the small number of entity or relation instances in the training dataset makes our model under-fitting for a type of entity or relation, which results in misidentification. In Example 6, “蛔虫肌肉 (ascaris muscle)” rarely appears in our dataset and our model misidentifies it as the bacteria type. The relation between “哌嗪 (piperazine)” and “蛔虫 (ascaris)” is also missed by our model due to its few occurrences.

Example 6: 枸橼酸哌嗪糖浆适应症: 用于蛔虫和蛲虫感染。哌嗪具有麻痹蛔虫肌肉的作用, 其机制可能为哌嗪在虫体神经肌肉接头处发挥抗胆碱作用, 阻断乙酰胆碱对蛔虫肌肉的兴奋作用。

*Piperazine citrate syrup is used for ascaris and enterobiasis infection. Piperazine can paralyze ascaris muscle, and its mechanism may make piperazine play an anticholinergic role at the neuromuscular junction of ascaris and block the excitatory effect of acetylcholine on ascaris muscle.*

**Entity:**

medicine: 枸橼酸哌嗪糖浆 (piperazine citrate syrup); 哌嗪 (piperazine); 乙酰胆碱 (acetylcholine).

disease: 蛲虫感染 (enterobiasis infection); 蛔虫 (Ascaris).

bacteria: 蛔虫肌肉 (Ascaris muscle).

**Relation:**

cure: (枸橼酸哌嗪糖浆, 蛔虫); (枸橼酸哌嗪糖浆, 蛲虫感染).

Besides our constructed dataset, we also made the experiments to the remaining unlabeled 653 pieces of medicine Chinese instructions, the following Example 7 has shown the automatic labelling results with our model.

Example 7: 利福平注射液适应症: 结核病。本品与其他抗结核药联合使用, 用于治疗各种类型结核, 包括初治, 进展期的, 慢性的及耐药病例。本品对大多数非典型的分枝杆菌菌株也有效。其他感染: 本品可以治疗难治性军团菌属及重症耐甲氧西林葡萄球菌感染。

*Rifampin injection: tuberculosis. This product is used in combination with other antituberculosis medicines to treat various types of tuberculosis, including initial treatment, progressive, chronic and drug-resistant cases. This product is also effective for most atypical mycobacterium strains. The other infection: This product can treat refractory legionella and severe methicillin resistant staphylococcus infection.*

**Entity:**

medicine: 利福平注射液 (rifampin injection); 本品 (this product); 本品 (this product); 抗结核药 (antituberculosis medicines); 本品 (this product); 甲氧西林 (methicillin).

disease: 感染 (infection); 结核 (tuberculosis); 葡萄球菌感染 (staphylococcus infection); 结核病 (tuberculosis); 难治性军团菌属 (refractory legionella); 重症耐 (severe resistant).

bacteria: 非典型的分枝杆菌菌株 (atypical mycobacterium strains).

**Relation:**

inhibition: (利福平注射液, 非典型的分枝杆菌菌株).

cure: (利福平注射液, 难治性军团菌属); (利福平注射液, 感染); (利福平注射液, 结核病); (利福平注射液, 结核).

We can find that entity extraction achieves good performance on the unlabeled medicine Chinese instruction. Moreover, our model makes satisfactory result in relation extraction with the unlabeled dataset to some extent. Our model recalls all the medical entities but the extracted 重症耐 (severe resistant) is not correct. All the extracted five relations are correct, but our model fails to identify the other three relations, i.e., (利福平注射液, cure, 葡萄球菌感染), (抗结核药, cure, 结核), and (利福平注射液, anaphora, 本品).

## 6 CONCLUSIONS

In this work, we construct a language corpus with Chinese medical instructions according to our specified annotation guidelines. A novel joint model is then developed to extract medical entities and relations simultaneously. Moreover, the self-attention mechanism is introduced into the joint model to learn the intra-sentence word dependencies. Although the proposed model does not use any artificial features and NLP tools, its performance is the state of the art.

In the future, we would like to focus on data generalization and try to train the learning ability of our model on the small dataset. We will conduct more experiments using emergent and popular pre-trained language models, such as BERT and GPT, and evaluate their contributions to the tasks. Our constructed dataset contains a large number of discontinuous entities, currently, only the main parts of the discontinuous entities can be labelled. In the future, we will introduce more language rules to solve the problem of discontinuous entity.

## REFERENCES

- [1] Harsha Gurulingappa, Abdul M. Rajput, Augus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J. Biomed. Inf.* 45, 5 (2012), 885–892.
- [2] Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Syst. Appl.* 114 (2018), 34–45. <https://bmcbioinformatics.biomedcentral.com/track/pdf/10.1186/1471-2105-6-S1-S14>.

- [3] Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*, 551–561.
- [4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of 31st Conference on Neural Information Processing Systems (NIPS'17)*, 5999–6009.
- [5] Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *Proceedings of the 6th International Conference on Learning Representations (ICLR'17)*.
- [6] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'18)*, 464–468.
- [7] Zhouhan Lin, Minwei Feng, Cicero N. Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *Proceedings of the 6th International Conference on Learning Representations (ICLR'17)*.
- [8] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2018. DiSAN: Directional self-attention network for RNN/CNN-free language understanding. In *Proceeding of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*, 5446–5455.
- [9] Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'18)*, 872–884.
- [10] Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018. Deep semantic role labeling with self-attention. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*, 4929–4936.
- [11] Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. 2005. ProMiner: Rule-based protein and gene entity recognition. *BMC Bioinf.* 6, Suppl 1 (2005), S14. <https://bmcbioinformatics.biomedcentral.com/track/pdf/10.1186/1471-2105-6-S1-S14>.
- [12] Christopher S. G. Khoo, Syin Chan, and Yun Niu. 2000. Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL'00)*, 336–343.
- [13] Oana Frunza and Diana Inkpen. 2010. Extraction of disease-treatment semantic relations from biomedical sentences. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, 91–98.
- [14] Maofu Liu, Li Jiang, and Huijun Hu. 2017. Automatic extraction and visualization of semantic relations between medical entities from medicine instructions. *Multimedia Tools Appl.* 76, 8 (2017) 10555–10573.
- [15] Abhyuday N. Jagannatha and Hong Yu. 2016. Bidirectional RNN for medical event detection in electronic health records. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'16)*, 473–482.
- [16] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*, 2124–2133.
- [17] Chanqin Quan, Lei Hua, Xiao Sun, and Wenjun Bai. 2016. Multichannel convolutional neural network for biological relation extraction. *BioMed Research International* 2016 (2016), 1–10. <http://downloads.hindawi.com/journals/bmri/2016/1850404.pdf>.
- [18] Bishan Yang and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, 1640–1649.
- [19] Sameer Singh, Sebastian Riedel, Brian Martin, Jiaping Zheng, and Andrew McCallum. 2013. Joint inference of entities, relations, and coreference. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, 1–6.
- [20] Suncong Zheng, Yuexing Hao, Dongyuan Lu, Hongyun Bao, Jiaming Xu, Hongwei Hao and Bo Xu. 2017. Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing* 257 (2017), 59–66. <https://www.sciencedirect.com/science/article/abs/pii/S0925232117301613>.
- [21] Pankaj Gupta, Hinrich Schütze, and Bernt Andrassy. 2016. Table filling multi-task recurrent neural network for joint entity and relation extraction. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING'16)*, 2537–2547.
- [22] Heike Adel and Hinrich Schütze. 2017. Global normalization of convolutional neural networks for joint entity and relation classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP'17)*, 1723–1729.
- [23] Maria Pershina, Bonan Min, Wei Xu, and Ralph Grishman. 2014. Infusion of labeled data into distant supervision for relation extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*, 732–738.

- [24] Lisheng Fu, Bonan Min, Thien H. Nguyen, and Ralph Grishman. 2018. A case study on learning a unified encoder of relations. In *Proceedings of the 4th Workshop on Noisy User-generated Text (W-NUT) at EMNLP*, 202–207.
- [25] Lisheng Fu, Thien H. Nguyen, Bonan Min, and Ralph Grishman. 2017. Domain adaptation for relation extraction with domain adversarial neural network. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP'17)*, 425–429.
- [26] Yee S. Chan, Joshua Fasching, Haoling Qiu, and Bonan Min. 2019. Rapid customization for event extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL'19)*, 31–36.
- [27] Makoto Miwa, and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*, 1105–1116.
- [28] Fei Li, Meishan Zhang, Guohong Fu, and Donghong Ji. 2017. A neural joint model for entity and relation extraction from biomedical text, *BMC Bioinf.* 18, 1 (2017) 198.
- [29] Arzoo Katiyar and Claire Cardie. 2017. Going out on a limb: Joint extraction of entity mentions and relations without dependency trees. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)*, 917–928.
- [30] Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. An attentive neural architecture for joint segmentation and parsing and its application to real estate ads. *Expert Syst. Appl.* 102 (2018), 100–112. <https://www.sciencedirect.com/science/article/abs/pii/S0957417418301192>.
- [31] Meishan Zhang, Yue Zhang, and Guohong Fu. 2017. End-to-End Neural Relation Extraction with Global Optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP'17)*, 1730–1740.
- [32] Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)*, 1227–1236.
- [33] Supot Nitsuwat and Wansa Paoon. 2004. Development of ICD-10-TM ontology for a semi-automated morbidity coding system in Thailand. *Methods Inf. Med.* 51, 6 (2004) 519–528.
- [34] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'16)*, 260–270.
- [35] Dan Roth, and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the 8th Conference on Computational Natural Language Learning (CoNLL'04) at NAACL-HLT*, 1–8.
- [36] Makoto Miwa and Yutaka Sasaki. 2014. Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, 1858–1869.
- [37] Rohit J. Kate and Raymond Mooney. 2010. Joint entity and relation extraction using card-pyramid parsing. In *Proceedings of the 14th Conference on Computational Natural Language Learning (CoNLL'10)*, 203–212.

Received August 2019; revised December 2019; accepted March 2020