



Article

A Joint Learning Model to Extract Entities and Relations for Chinese Literature Based on Self-Attention

Li-Xin Liang ¹, Lin Lin ^{1,*}, E Lin ^{2,*}, Wu-Shao Wen ² and Guo-Yan Huang ²¹ College of Big Data and Internet, Shenzhen Technology University, Shenzhen 518118, China; lianglixin@sztu.edu.cn² School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou 510006, China; wenwsh@mail.sysu.edu.cn (W.-S.W.); huanggy7@mail2.sysu.edu.cn (G.-Y.H.)

* Correspondence: linlin@sztu.edu.cn (L.L.); sysuline@gmail.com (E.L.)

Abstract: Extracting structured information from massive and heterogeneous text is a hot research topic in the field of natural language processing. It includes two key technologies: named entity recognition (NER) and relation extraction (RE). However, previous NER models consider less about the influence of mutual attention between words in the text on the prediction of entity labels, and there is less research on how to more fully extract sentence information for relational classification. In addition, previous research treats NER and RE as a pipeline of two separated tasks, which neglects the connection between them, and is mainly focused on the English corpus. In this paper, based on the self-attention mechanism, bidirectional long short-term memory (BiLSTM) neural network and conditional random field (CRF) model, we put forth a Chinese NER method based on BiLSTM-Self-Attention-CRF and a RE method based on BiLSTM-Multilevel-Attention in the field of Chinese literature. In particular, considering the relationship between these two tasks in terms of word vector and context feature representation in the neural network model, we put forth a joint learning method for NER and RE tasks based on the same underlying module, which jointly updates the parameters of the shared module during the training of these two tasks. For performance evaluation, we make use of the largest Chinese data set containing these two tasks. Experimental results show that the proposed independently trained NER and RE models achieve better performance than all previous methods, and our joint NER-RE training model outperforms the independently-trained NER and RE model.

Keywords: Chinese named entity recognition; relation extraction; joint learning; long short-term memory neural network; self-attention mechanism

MSC: 68T50

Citation: Liang, L.-X.; Lin, L.; Lin, E.; Wen, W.-S.; Huang, G.-Y. A Joint Learning Model to Extract Entities and Relations for Chinese Literature Based on Self-Attention. *Mathematics* **2022**, *10*, 2216. <https://doi.org/10.3390/math10132216>

Academic Editor: Radu Tudor Ionescu

Received: 13 April 2022

Accepted: 20 June 2022

Published: 24 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the era of information explosion, how to extract effective information from massive and heterogeneous text and organize it into a structured representation is a hot research topic. In order to solve the problem of automatic extraction of massive information in the text field, NER (named entity recognition) and RE (relation extraction) technology came into being. Named entity recognition is to identify the entity that represents the information unit in natural language, for example, to identify the entity, such as person's name, time, place, etc. Relation extraction is to identify the semantic relationship between pairs of entities in the text. As shown in Figure 1, this is an example in a Chinese corpus. In the sentence "In the autumn, mother goes up the mountain and chops, when she comes back, her pocket always carries a few grains of sour jujube, or hawthorn . . . that is the most extravagant snacks of childhood mother gives me", named entity recognition technology identifies the time entity, person entity, and thing entity, such as "autumn", "mother", "pocket", etc.;

relation extraction technology identifies relationships between entity pairs, for example, <sour jujube, pocket> and <hawthorn, pocket> are both the located relationship.

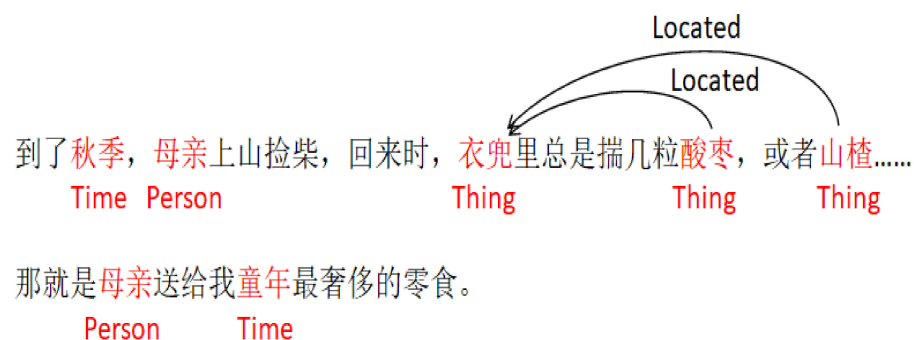


Figure 1. Example from the Chinese literature text corpus.

The research on named entity recognition and relation extraction uses rules-based and dictionary-based methods [1] at the beginning. These methods obtain entity label prediction results by comparing template library of entity identification rules with input text sequence, and on the other hand, obtain the semantic relationship between entity pairs according to the matching result between relationship examples and rule templates prepared by experts. Although these methods have certain effects, the performance depends largely on the rule template written by linguistic experts and the quality of the dictionary: if the rules are designed with too much detail, it is easy to cause frequent conflicts of entity recognition, while too few rules may cause insufficient coverage. In the meantime, the rule template designing process is time consuming, labor intensive, very error prone, and has language limitations.

The named entity recognition technology based on the hidden Markov model [2], maximum entropy model [3], conditional random field [4] and other statistical machine learning methods has a certain performance improvement compared with the rule-based method. These statistics-based methods do not require much linguistic knowledge and are more portable. However, these methods have a higher requirement for feature selection, and it is necessary to select features from the text that have a greater impact on entity recognition, such as word feature, contextual feature, lexicon and part-of-speech feature, stop word feature, core word feature and semantic feature. The selection of these features directly affects the performance of the named entity recognition model. In addition, the relation extraction method based on feature vectors [5], kernel functions [2] also depends on the selection of features. The choice of feature space, such as dependency syntax, syntax tree, tree kernel, and convolution tree kernel, has a great influence on the performance of the model.

Named entity recognition and relation extraction models based on neural network models such as convolutional neural network (CNN) and recurrent neural network (RNN) are the current mainstream methods, with the trend of “de-artificial features”. These methods outperform traditional ones without a large number of artificial features. In addition, if more semantic information or high-quality external dictionary information input is added to these neural network models, the performance can be further improved. In particular, in the current work, the research on named entity recognition and relation extraction mainly focuses on the English corpus but less on the Chinese corpus. The Chinese corpus contains more complex named entity recognition and relation extraction problems. Moreover, most researchers regard named entity recognition and relation extraction as two independent subtasks in the information extraction task pipeline. This method shares the underlying semantic connection between the two tasks. In addition, the research on these two tasks pays less attention to the influence of mutual attention between sentence characters on the annotation results of entity sequences and the influence of attention at different semantic levels of sentences on the classification results. In view of the shortcomings of

the current research, this paper proposes the Chinese named entity recognition model based on BiLSTM-Self-Attention-CRF and relation extraction model based on BiLSTM-Multilevel-Attention, on the basis of the self-attention mechanism, BiLSTM (bidirectional long short-term memory) and conditional random fields (CRF). This paper uses word vector and feature representation to connect these two tasks and share the same underlying module. It proposes a neural network model for the joint learning of these two models. The context is the vector feature extracted by BiLSTM, and the attention matrix is trained from the context. Experimental results show that the proposed model is better than other neural network models and has achieved good performance in the open Chinese corpus of Peking University.

In summary, the main contributions of this study are as follows:

1. We designed the BiLSTM-Self-Attention-CRF model to realize the Chinese named entity recognition. The BiLSTM model is used to obtain the text features. The feature weights are calculated by combining the self-attention mechanism. Finally, CRF is used to decode, and the entity recognition results are obtained.
2. We designed the BiLSTM-Multilevel-Attention model to realize relation extraction. The BiLSTM is used to consider the context and semantic connection of the text entirely, and the key feature words are obtained by combining a multilevel attention mechanism.
3. We put forward a joint model to make the data information of the two parts more closely linked and fully capture the inline information of the two tasks.

In Section 2 of this paper, we introduce the related work of neural network-based named entity recognition and relation extraction models. In Section 3, we elaborate on the neural network model and its submodules for the joint learning of named entity recognition and relation extraction models. In Section 4, we show the experimental data, training parameter settings and experimental design, and analyze the experimental results in detail. Finally, in Section 5, we summarize the research work of this paper.

2. Related Work

The deep learning technology based on neural network model has achieved great success in many natural language processing tasks, such as machine translation, part-of-speech labeling, sentiment classification, etc. This section mainly introduces related research and its result of named entity recognition and relation extraction methods based on the neural network model.

2.1. Named Entity Recognition

Named entity recognition is to identify the entity representing the information unit in the natural language. For this task, Collobert et al. [6] first introduced the neural network model using two network structures: the window method and the sentence method. The experimental results show that the effect of the sentence method using the CNN + CRF structure is obviously better than traditional neural network in the window method. Chiu et al. [7] proposed the network structure of BiLSTM + CNN, where the input vector integrated the word vector and character vector and completed the entity recognition task by combining two dictionaries constructed from a public corpus. Lample et al. [8] implemented the network structure of LSTM + CRF, which could achieve the best performance at that time without adding any artificial features. Ma et al. [9] added the CRF layer on the bidirectional LSTM + CNN, constructed the network structure of LSTM-CNNs-CRF, and achieved the highest F_1 value on the CoNLL-2003 English data set without the dictionary. Rei et al. [10] introduced the attention model into the named entity recognition task, used the attention mechanism to weight the character vector and the word vector so that the model could generate better text features. The experimental effect indicated that it was better than the model of directly splicing the character vector and the word vector. Bharadwaj et al. [11] attempted to introduce the international phonetic symbol into the neural network model of entity recognition, providing a solution to the problem of

multilingual entity recognition tasks. Peng et al. [12] used the character vector, the word vector, and the character vector trained by the position of the character in the word as the input of the CRF model; they combined the objective function of the word vector with the objective function of the entity recognition task for joint training and achieved good results. Peng et al. [13] also introduced the joint training for the Chinese word segmentation model [14] and the named entity recognition model, improved the performance of named entity recognition by jointly training these two models to share the word vector representation layer and introduce word segmentation location information. Realizing that abbreviations for Chinese named entities may lead to problems with reduced recognition performance, Zhang et al. [15] proposed a network structure called the recurrent architecture dynamic dictionary (RADD) based on RNN and the dynamic dictionary discriminating module. RADD is a two-classifier that first determines whether the input character sequence can constitute a word, then the determination result is input to the entity recognition model. It obtained the current optimal performance in a corpus with Chinese abbreviations. Dong et al. [16] added the order of the radicals of Chinese characters into the training of word vector as one of the characteristics of a word and achieved the current optimal F_1 value on the SIGHAN-MSRA data set.

Some researchers have made some improvements to the BiRNN (bidirectional-RNN)-CRF structure. The bidirectional-RNN network can maximize the context characteristics of a specific time series, and the CRF layer can effectively learn and apply the rules between entity labels. The researchers are making the model based on the structure of BiRNN + CRF better to improve the performance of named entity recognition, for example, by introducing the attention mechanism to add the representation of the word vector plus the character vector [10], the phonological characteristics of the characters [11], and the position information of Chinese characters in Chinese word segmentation [12]; they are using pretrained Wikipedia entity classification results as dictionary features [17] and other methods to construct a word vector that incorporates more semantic features and high-quality dictionary features as one of the inputs of the BiRNN + CRF network structure.

2.2. Relation Extraction

Relation extraction is to identify the semantic relationship between pairs of entities in the text. Nguyen et al. [18] applied CNN to the relation extraction problem for the first time, firstly embedded the word vector and the relative position of two entities, then spliced the two vectors as the input of the CNN, and finally obtained the result of the relation extraction through the convolutional layer, the pooling layer, and the softmax layer. Santos et al. [19] designed an improved CNN structure called CR-CNN, added a “relationship type score layer” after the convolutional layer and calculated the score of the sentence in each relationship category, thereby improving the performance of the model. Chen Yu et al. [3] applied DBN (deep belief nets) to the Chinese named entity relation extraction and achieved a better experimental effect than such models as supporting vector machines. Xu et al. [20] proposed the SDP-LSTM (shortest dependency path-LSTM) network structure based on the shortest dependency path, used the word vector, part-of-speech labeling, grammatical relationship, etc., of each node on the shortest dependency path between two entities as input to the LSTM unit to improve the performance of the model. Cai et al. [21] proposed a bidirectional recurrent convolutional neural network (BRCNN), used the shortest dependency path (SDP) for the forward and backward layers as input to the bidirectional LSTM network, set up two separate streams in LSTM which are the words in SDP and the relationship between words in SDP, input the output of the bidirectional LSTM into two corresponding forward and backward CNNs, and finally obtained the relationship classification result through the pooled layer and the softmax layer. Zhou et al. [22] proposed the LSTM network structure based on the attention mechanism, Wang et al. [23] added two layers of attention mechanism in the CNN structure, gave attention to the position of each word in the input sequence relative to two entities and the relatively different relationship categories on the pooling layer. Zeng et al. [24]

proposed PCNN (piecewise convolutional neural networks) to extract text features, where the difference from CNN was that the text was divided into before entity 1, between entity, after entity 2 after the pooling layer according to the position of the two entities, and the maximum pooling was performed on the three texts.

2.3. Joint Learning

Most studies treat named entity recognition and relation extraction as two separate subtasks in the information extraction task pipeline, thus breaking the underlying semantics connection between the two tasks. Miwa et al. [25] introduced the end-to-end idea to relation extraction task, where the named entity recognition and relation extraction module share the underlying bidirectional LSTM module, input the sequence labeling result and the text dependency analysis information into the relation extraction module of tree bidirectional LSTM model, and finally obtained the relationship classification result. Zheng et al. [26] also performed joint learning of the entity recognition task and the relation extraction task by sharing the underlying LSTM representation of the neural network. He used the encoder–decoder model in the named entity recognition module and the CNN structure in the relation extraction module. Zheng et al. [27] also proposed a new labeling strategy that combines entity sequence labeling and relationship result labeling. This major innovation transformed the classification task into a sequence labeling task; however, the labeling strategy failed to solve the problem of multiple relationships between one entity and other entities in one processing unit.

From the current research, we can see that the relation extraction research pays more attention to the influence weight of each character on relationship classification; the sentence abstract information used for relationship classification is a one-dimensional vector; the representative sentence features are not rich enough; and there is little exploration of the influence of different semantic levels of sentences on the results of relationship classification. In addition, the effect of joint learning on named entity recognition and relation extraction on the performance of these two tasks has not been studied.

In view of these shortcomings, this paper introduces the self-attention mechanism into the Chinese named entity recognition and relation extraction task, and explores the impact of joint learning of named entity recognition and relation extraction on the performance of these two tasks.

3. Model Framework

3.1. Long-Term and Short-Term Memory Networks

The long short-term memory neural network (LSTM) is a variant of RNN that solves gradient disappearance and gradient explosion problems and can effectively capture long-range sequence features. When the backpropagation algorithm is running, LSTM can largely preserve the difference between the predicted result and the actual result during the training process and keep the difference at a more constant level so that the circular network can learn many sequence moments, thus establishing long-distance causal connection of sentence sequences.

LSTM removes or adds information to the cell state through the “gate”, which is composed of sigmoid function and point-by-point multiplication group. The sigmoid function outputs a value between 0 and 1, describing the degree to which the information can pass the threshold: 0 for not passing any information, and 1 for all information to pass. The “gate” determines how much information is passed by screening the information using the weight parameter and the offset parameter. These parameters are obtained during the network training process. Each LSTM unit controls the influence of information on the state of the unit through three gates, which are the forgetting gate, the input gate and the output gate. The forgetting gate is mainly responsible for controlling how much information in the memory unit at the last moment can be accumulated in the memory unit at the current moment; the input gate mainly controls how much information in the candidate memory unit can enter the current memory unit; and the output gate mainly controls how much

information in the memory unit can enter the calculation of the current hidden layer. The design of the forgetting gate, the input gate, the output gate and the memory unit enables the LSTM unit to store, read and update long-distance history information. Its structure is shown in Figure 2.

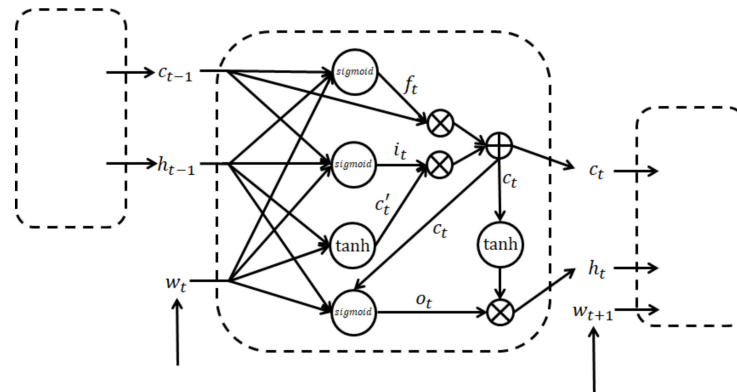


Figure 2. LSTM unit structure.

We need to obtain the output of the hidden layer at current moment, which depends on the weight parameter and the offset parameter, the output of the previous memory unit, the output of the previous hidden layer, and the output of each gate. Suppose the current time is t , the output of the previous memory unit is c_{t-1} , the output of the previous hidden layer is $h(t-1)$, the input is the output of the current candidate memory unit is c'_t , the output of the current memory unit is c_t , and the output of the current hidden layer is h_t ; the activation function is sigmoid or tanh function, then the calculation formula is as follows:

$$\begin{aligned}
 \text{sigmoid}(x) &= \frac{1}{1+e^{-x}} \\
 \text{tanh}(x) &= \frac{e^x - e^{-x}}{e^x + e^{-x}} \\
 f_t &= \text{sigmoid}(W_{fc}c_{t-1} + W_{fh}h_{t-1} + W_{fe}e_t + b_f) \\
 c'_t &= \text{tanh}(W_{ch}h_{t-1} + W_{ce}e_t + b_c) \\
 i_t &= \text{sigmoid}(W_{ic}c_{t-1} + W_{ih}h_{t-1} + W_{ie}e_t + b_i) \\
 c_t &= f_t c_{t-1} + i_t c'_t \\
 o_t &= \text{sigmoid}(W_{oc}c_t + W_{oh}h_{t-1} + W_{oe}e_t + b_o) \\
 h_t &= o_t \text{tanh}(c_t)
 \end{aligned} \tag{1}$$

3.2. Traditional Attention Mechanism and Self-Attention Mechanism

The encoder–decoder framework in natural language processing is a general processing framework for solving the task of how to generate the output text sequence Y from input text sequence X . In this framework, when the decoder generates the target sequence Y , the abstract semantics C used for each output y_i in Y is the same, and C is generated by encoder encoding for each x_i in the text sequence X . This means that the influence of each x_i is the same for generating y_i . However, that is not the actual case. Taking one named entity recognition task as an example, when the sentence “Zhuge Liang, Guan Yu, etc. guard Jingzhou” generates the corresponding entity label, it is obvious that “guard” has more influence on “Jingzhou” to be marked as a place name than other words, i.e., each word’s impact on the generation of entity labels is not the same, but the traditional encoder–decoder framework cannot handle this uneven influence problem.

In order to solve the problem of uneven influence in the decoding process, the attention mechanism is introduced into the decoding process, and each y_i in the generated target sequence is affected by the different attention distribution information of each input in the sequence. This framework is shown in Figure 3.

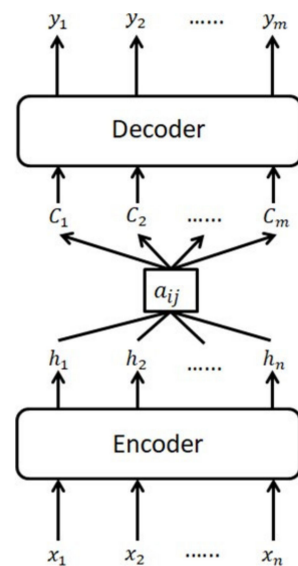


Figure 3. The framework of encoder–decoder based on attention mechanism.

It can be seen from the figure that attention information C_i , which is constantly changing according to current generation of y_i , is added to the model, and the calculation of y_i becomes

$$y_i = g(C_i, y_1, y_2, \dots, y_{i-1}), 1 \leq i \leq m \quad (2)$$

Suppose a_{ij} is the correlation between h_j which is obtained by the encoder from the input of the j -th time sequence and the decoder's i -th phase. Then the semantic representation of the attention information C_i for the i -th phase is derived from the weighted sum of h_j and a_{ij} :

$$C_i = \sum_{j=1}^n a_{ij} h_j, 1 \leq i \leq m \quad (3)$$

where n is the length of the input sequence, m is the length of the output sequence, and the correlation a_{ij} is obtained during the training of the model. This traditional attention mechanism is more like a mechanism of input and output alignment, with the calculation of the correlation a_{ij} introducing the external information of the stage i of the Decoder. The self-attention mechanism only updates the weight parameters through its own information. The calculation of the attention does not need to introduce external information and is more focused on the influence between input sequences. It is very suitable for the characteristics of the named entity recognition and relation extraction task. For example, for the sentence “Meng Yiping, President of Mengniu Group, came to our institute for visits and exchanges” and its given entity pair of <Sun Yiping, Mengniu Group>, the influence of the word “president” on the “ownership” relationship of the entity pair should be greater than other words. Effectively abstracting the attention influence matrix of different semantic layers on the classification result of the sentence is a crucial step to improve the performance of the model.

3.3. Joint Learning Model

The joint learning model consists of a shared underlying module, a self-attention-CRF module for the Chinese named entity recognition task, and a multilevel-attention module for the relation extraction task. The structure is shown in Figure 4. The joint learning model can train a richer vector representation by sharing the underlying parameters. By joint training these two tasks, it considers more the connection of the word vector and context feature representation in the neural network model between the two tasks. Through joint learning, the performance of the named entity recognition and relation extraction task is improved.

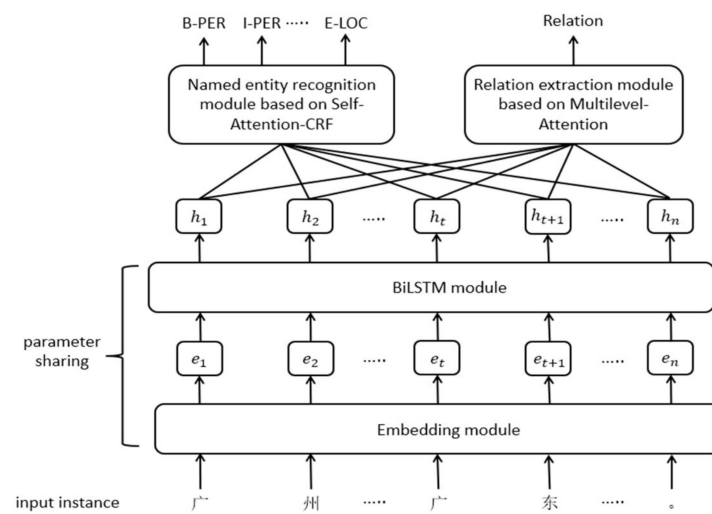


Figure 4. The framework of joint learning model based on self-attention.

3.3.1. Shared Underlying Module

The shared underlying module consists of the embedding module and the BiLSTM module. The training parameters are updated and shared during the training process. The embedding module converts the input characters into a word vector containing the semantic information of the character. The word vector e_t is spliced together by the pre-trained Chinese word vector e_{ce} , and the Chinese character stroke order feature vector e_{se} :

$$e_t = [e_{ce}, e_{se}] \quad (4)$$

Among them, the Chinese character stroke order feature vector e_{se} is the output derived from the stroke order of the word through the LSTM network. Figure 5 is an example of the stroke order feature vector generation of the “Guo” word. The “Guo” word stroke order is “vertical fold horizontal horizontal vertical horizontal down horizontal”, the corresponding stroke order number is 25112141, and the stroke order number sequence is input into the LSTM network to obtain the stroke order feature output.

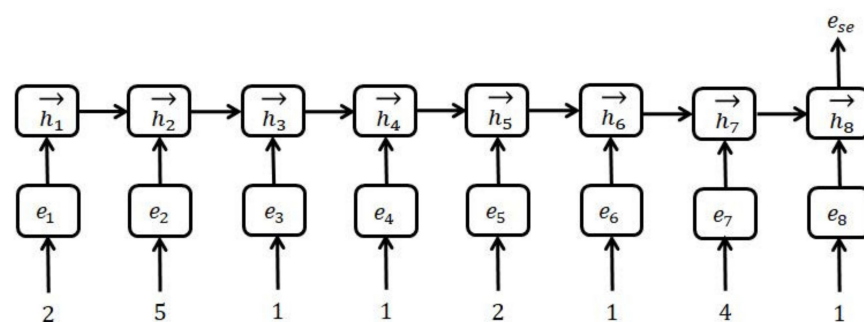


Figure 5. The stroke order feature generation of “Guo”.

The input of the BiLSTM module is the output of the embedding module. The bidirectional LSTM structure is used to extract the context characteristics of the input text. The module consists of the LSTM forward layer and backward layer. Each layer consists of several LSTM units. The LSTM forward layer output \vec{h}_t at time t can be calculated by the

formula, and the calculation of backward layer output \overleftarrow{h}_t is similar. Therefore, \overrightarrow{h}_t , \overleftarrow{h}_t and the output of the BiLSTM module at that time h_t can be expressed as

$$\begin{aligned}\overrightarrow{h}_t &= \text{lstm}(c_{t-1}, h_{t-1}, e_t) \\ \overleftarrow{h}_t &= \text{lstm}(c_{t+1}, h_{t+1}, e_t) \\ h_t &= [\overrightarrow{h}_t, \overleftarrow{h}_t]\end{aligned}\quad (5)$$

3.3.2. Named Entity Recognition Module Based on Self-Attention-CRF

Different from the current mainstream traditional attention mechanism, this paper introduces self-attention mechanism into the named entity recognition task, focusing more on extracting the influence of the attention between words and words on the sequence labeling results in the text sequence. The input of the module is the output of the BiLSTM module, and the entity label prediction results of the text sequence are obtained after the linear layer, the self-attention layer, the output layer and the CRF layer.

The linear layer is responsible for performing three linear transformations on the output H of the BiLSTM module to obtain three matrices of Q , K and V :

$$\begin{aligned}Q &= HW_q, W_q \in R^{\text{dim} \times \text{dim}} \\ K &= HW_k, W_k \in R^{\text{dim} \times \text{dim}} \\ V &= HW_v, W_v \in R^{\text{dim} \times \text{dim}}\end{aligned}\quad (6)$$

where dim is the representation vector dimension of each word in the output H of the BiLSTM module, and W_q , W_k , and W_v are the linear transformation parameters of H .

In the process of training, the self-attention layer obtains the attention matrix representing the influence between word and word in the input sequence. This layer first performs matrix multiplication of Q and K , and then obtains attention weight matrix α through the *softmax* function:

$$\alpha = \text{softmax}(QK^T), \alpha \in R^{n \times n}\quad (7)$$

where n is the length of the text sequence. Use matrix multiplication of α and V to give the influence weight of all words on each word in the sequence, and obtain the resulting sentence representation vector M :

$$M = \alpha V, \alpha \in R^{n \times \text{dim}}\quad (8)$$

The input of the output layer is M , and the output is the entity label score vector l corresponding to each word in the sentence:

$$l = \tanh(MW_l + b_l), W_l \in R^{\text{dim} \times \text{tag}_{num}}, b_l \in R^{\text{tag}_{num}}\quad (9)$$

where W_l is the weight parameter of M , b_l is the offset parameter, \tanh is the activation function, and tag_{num} is the total number of labels.

The input of the CRF layer is the entity label score vector l , which is used to add constraints to the entity label prediction to ensure that the predicted label conforms to the rules. For example, the entity label of the first character in the sentence sequence is definitely not the label "I-" and "E-"; in the label prediction sequence, "B-1, I-2, E-3", 1, 2, 3 should be the same entity type, for example, "B-PER, I-PER" is a legal sequence, "B-PER, I-LOC" is an illegal sequence. These constraint rules are learned by the CRF layer according to the entity label score vector l of the output layer and the corresponding correct label during the training process. The mathematical representation is an entity label probability transformation matrix. In the process of model training, the CRF layer is responsible for the learning of the probability transformation matrix. In the process of model testing and

use, the transformation matrix is used, combined with the entity label score vector l input by the layer, to obtain the final entity prediction label by the Viterbi algorithm.

3.3.3. Relation Extraction Module Based on Multilevel-Attention

At present, the one-dimensional self-attention sentence vector representation is commonly used in relation extraction research. In this paper, the self-attention sentence matrix is adopted to extract more abundant sentence information, which is used to extract the relationship between entity pairs. The input of this module is the output of the BiLSTM module. After the entity feature integration layer, the multilevel-attention layer and the output layer, the classification result of the relationship instance is obtained. The entity feature integration layer adds the following two features to the output vector of the BiLSTM module:

1. Relative entity location feature. There are two values, which represent the distance between the current word and entity 1 and entity 2, respectively. For example, in the statement “Guangzhou is the capital of Guangdong”, the values of relative entity location feature of each word for the entity pair <Guangzhou, Guangdong> are shown in Table 1.
2. Entity type feature. There are two values, entity type 1 of entity 1 and entity type 2 of entity 2. As shown in the above example, the entity pairs <Guangzhou, Guangdong> are both place name entities, and their corresponding entity type features are location.

Table 1. Instance of relative entity location.

Word	广	州	是	广	东	的	省	会	.
Feature pos1	0	1	2	3	4	5	6	7	8
Feature pos2	3	2	1	0	−1	−2	−3	−4	−5

The entity feature integration layer separately embeds the two features of relative entity location and entity type to obtain a vector representation, and splices the feature vector with the output h_t of the BiLSTM module to obtain a vector ω_t of the corresponding character at time t :

$$\omega_t = [h_t, pos1_emb, pos2_emb, entitytype1_emb, entitytype2_emb] \quad (10)$$

The multilevel-attention layer introduces a multilayer self-attention mechanism into the output of the feature integration layer, which is mathematically represented as a weight matrix α , calculated by the matrix representation W of the sentence, the weight parameters S_1 and S_2 :

$$W = [W_1, W_2, \dots, W_n], W \in R^{n \times \text{dim}} \quad (11)$$

$$\alpha = \text{softmax}(\tanh(WS_1)S_2), S_1 \in R^{\text{dim} \times d}, S_2 \in R^{\text{dim} \times L} \quad (12)$$

where n is the length of the sentence, and d and L are hyperparameters. From the above formula, the α dimension is $n \times L$, and L represents how many levels of each word in the sentence are extracted to get the weight information. Therefore, each row of α represents the weight information of the word at the L levels, Then the representation matrix W of the original sentence is multiplied by the weight matrix α to obtain the sentence feature vector H after the weight is given by the hierarchical attention matrix:

$$H = \tanh(W^T \alpha), H \in R^{\text{dim} \times L} \quad (13)$$

Finally, the predicted relationship classification label y' is obtained through the output layer:

$$\begin{aligned} h &= W_h H + b_h \\ p &= \text{softmax}(h) \\ y' &= \text{argmax}(p) \end{aligned} \quad (14)$$

where W_h is the weight parameter of H , and b_h is the offset parameter.

4. Experiment and Analysis

4.1. Corpus

Peking University's open Chinese corpus [28] is one of the corpora of Chinese entity corpus with many entity types and relationship instances. The corpus has 28,896 sentences in the named entity recognition data set, and a total of 159,862 entities are marked. Its entity categories and quantity ratio are shown in Table 2.

Table 2. The set of entity labels.

Category	Example	Quantity	Ratio
Thing	Apple	56,940	35.6%
Person	Li Qiu	52,265	32.7%
Location	Paris	27,442	17.2%
Time	One Day	11,757	7.4%
Metric	One Liter	5818	3.6%
Organization	Guerrillas	3275	2.0%

The corpus has a total of 14,501 relationship instances in the relation extraction data set. Its categories and quantity ratios are shown in Table 3.

Table 3. The set of relation labels.

Category	Example	Quantity	Ratio
Located	Youlan valley-Valley	5730	39.5%
Part-Whole	Hongkong-China	3254	22.4%
Family	Liu Pi-Liu Bei	1395	9.6%
General-Special	Fish-Sole	949	6.5%
Social	Mother-Neighborhood	884	6.1%
Ownership	Village-villager	838	5.8%
Use	Grandpa-quill	697	4.8%
Create	Ba Jin-Home	383	2.6%
Near	Fo Shan-Guang Zhou	371	2.6%

4.2. Metrics

This paper conducts a 5-fold cross-validation on the Chinese literary corpus published by Peking University. The experiment uses the weighted precision, recall and F_1 value of each entity category and relationship category to evaluate the named entity recognition and relation extraction task. F_1 is calculated as follows:

$$F_1 = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (15)$$

In order to facilitate the comparative analysis of the experiment, the parameters of the model training are set according to reference [28,29]. The main training parameter settings of the model are shown in Table 4.

Table 4. The set of model train parameters.

Parameter	Value
epoch	100
batch size	32
learning rate	0.001
L2 regularization	0.0001
dropout rate	0.5
optimizer	AdaDelta
Chinese character embedding size	200
stroke num embedding size	5
stroke sequence LSTM output size	10
LSTM embedding size	200
entity relative position embedding size	5
entity type embedding size	5
L	50
embedding initializer	Xavier

4.3. Experimental Results and Analysis

4.3.1. Named Entity Recognition Task

The experimental results of the named entity recognition task are shown in Table 5. Among the experimental results, the BiLSTM and CRF models are data set baseline standards given in reference [28], and BiLSTM-CRF and BiLSTM encoder–decoder models are the mainstream models for solving the named entity recognition task based on neural networks. This paper also implements the latter two models as performance comparison models.

Table 5. Results of named entity recognition.

Model	Indicator	Thing	Person	Location	Organization	Time	Metric	All
BiLSTM [28]	P	67.07	80.30	58.09	0	64.47	46.15	70.52
	R	62.37	78.50	46.79	0	45.51	22.18	62.36
	F	64.63	79.39	51.83	0	53.36	29.96	66.19
CRF [28]	P	75.72	87.92	68.41	46.69	76.20	70.50	77.72
	R	65.42	82.27	50.98	45.26	60.93	38.42	65.91
	F	70.19	85.00	58.42	45.96	67.72	49.74	71.33
BiLSTM-CRF [8]	P	71.96	87.04	60.4	55.20	67.52	52.77	74.79
	R	69.05	86.99	60.13	28.40	61.10	47.04	72.84
	F	70.48	87.01	60.26	37.50	64.15	49.74	73.80 (+2.47)
BiLSTM-Encoder-Decoder [26]	P	64.89	84.46	63.44	47.22	73.85	66.44	72.99
	R	72.49	88.06	56.35	27.28	56.80	32.57	73.37
	F	68.48	86.23	59.69	35.14	64.21	43.71	73.18 (+1.85)
BiLSTM-Self-Attention-CRF	P	70.62	86.15	62.61	51.13	72.66	59.91	75.30
	R	70.80	89.16	59.15	27.98	54.03	43.75	74.10
	F	70.71	87.63	60.83	36.17	61.98	50.57	74.51 (+3.18)
NER RE joint learning	P	70.51	85.85	62.82	51.01	72.81	59.75	75.12
	R	70.98	89.51	59.01	28.55	53.79	44.23	74.30
	F	70.75	87.63	60.82	36.57	61.86	50.80	74.53 (+3.20)

It can be seen from the experimental results that the BiLSTM-Self-Attention-CRF model proposed in this paper achieves the best effect in the single named entity recognition model, which is 3.18 higher than the best-performing CRF model in the data set baseline standard. It also outperforms the mainstream BiLSTM-CRF and BiLSTM-Encoder-Decoder models by 0.71 and 1.33 respectively, which verifies the validity of the single named entity recognition model. In the joint learning model, the named entity recognition module achieves the F_1 value of 74.53, having an improvement of 0.02 over the singlenamed entity recognition

model proposed in this paper. It can be seen that the joint learning model improves the performance of the named entity recognition task, but it is not significant.

It can be seen from Table 5 that the neural network model embedded in the CRF layer can enhance the learning of the dependencies between global entity labels, which is a more effective way to improve performance. Although the BiLSTM-Encoder-Decoder model adds the last time entity label input to the decoder, its acquisition of the previous time sequence information is still limited, and the performance is not as good as the BiLSTM-CRF model. However, the single CRF easily obtains high accuracy (77.72) and low recall rate (65.91), and the bidirectional LSTM network for efficient extraction of text features can improve performance.

Specific to each entity category, it can be seen from Table 5 that the model proposed in this paper has improved performance compared to other models in four entity categories: thing, person, location and metric. However, in two entity categories, organization and time, the performance is not as good as the CRF model. This is because the CRF model can artificially formulate multiple feature functions for specific entity types. These feature functions are very important for identifying complex Chinese institution name and time entities.

The model proposed in this paper uses the self-attention mechanism to carry out the weight learning of the influence between words and words in the text sequence. The mathematical representation is the attention matrix in the model. Visualizing the attention matrix helps to analyze the role it plays in the entity recognition. After comparative analysis, the effect of attention can be divided into the following four types:

1. The mutual influence between entities of the same type contributes to the labeling of the entity in the sentence. As shown in Figure 6, the words “Guizhou”, “Hunan” and “province” have far-reaching influence on “Jiangxi”, a lot more than the influence of other words; the influence of the words “Guizhou”, “Jiangxi” and “province” on “Hunan” is much greater than the influence of other words. These influences tend to increase the probability that “Jiangxi” and “Hunan” are labeled as the same type of location as “Guizhou”.
2. The interaction between entities with referential relationships contributes to the labeling of entities in sentences. As shown in Figure 7, the influence of “wetland” on “paradise” is much greater than the influence of other words in sentences. The influence of the referential relationship tends to increase the probability that the “paradise” is labeled as the same type of location as the “wetland”.
3. The attention between words and words in a longer entity contributes to the accurate labeling of the long entity. As shown in Figure 8, the influence of the eight words in the “1960s and 1970s” entity is much greater than the influence of other words.
4. The influence of verbs in sentences contributes to the recognition of noun entities. As shown in Figure 9, the influence of “crawl” on “body” is much greater than the influence of other words.



Figure 6. Instance of self-attention visualization 1.



Figure 7. Instance of self-attention visualization 2.

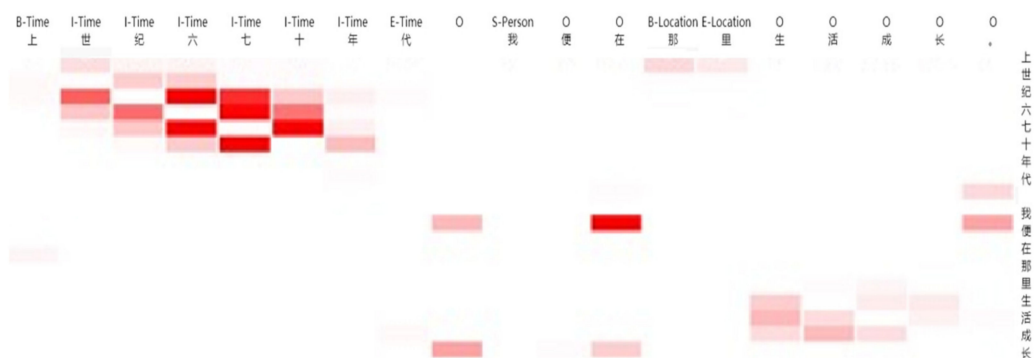


Figure 8. Instance of self-attention visualization 3.



Figure 9. Instance of self-attention visualization 4.

4.3.2. Relation Extraction Task

The results of the relation extraction task are shown in Table 6. The SVM, RNN, CNN, CR-CNN, SDP-LSTM, DepNN, and BRCNN models are the data set baseline standards given in reference [28]; the SR- BRCNN model [29] reached the current optimal level in the relation extraction task of the data set. That paper was published in NAACL 2018; this paper also implements the BiLSTM-Attention model [22] as one of the performance comparison models, which is a popular model of current attention mechanism in the relational extraction task.

Table 6. Results of relation extraction.

Model	Feature	F_1
SVM	Word embeddings, NER, WordNet, HowNet, POS, dependency parse, Google n-gram	48.9
RNN	Word embeddings + POS, NER, WordNet	48.3 49.1
CNN	Word embeddings + word position embeddings, NER, WordNet	47.6 52.4
CR-CNN	Word embeddings + word position embeddings	52.7 54.1
SDP-LSTM	Word embeddings + POS, NER, WordNet	54.9 55.3
DepNN	Word embeddings, WordNet	55.2
BRCNN	Word embeddings + POS, NER, WordNet	55.0 55.6
SR-BRCNN	Word embeddings + POS, NER, WordNet	65.2 65.9
BiLSTM-Attention	Word embeddings	64.6
BiLSTM-Multilevel-Attention	Word embeddings + stroke embeddings	73.3 (+7.4) 73.6 (+7.7)
NER RE joint learning	Word embeddings + stroke embeddings	78.8 (+12.9) 79.2 (+13.3)

It can be seen from the experimental results that, in the single relation extraction model, the BiLSTM-Multilevel-Attention model proposed in this paper achieves the best F_1 value of 73.6, 18.0 higher than the F_1 value of 55.6 of the best-performing BRCNN model in data set baseline standard, and 7.7 higher than the F_1 value of the current SR-BRCNN model which has the best relation extraction performance on the data set. This verifies the validity of the single relation extraction model proposed in this paper. In the joint learning model, the relation extraction module obtained a F_1 value of 79.2, which is further improved by 5.6 on the basis of 73.6. The F_1 value is accumulatively 13.3 higher than that of the current SR-BRCNN model, which has the best relation extraction performance on the data set. The joint learning model considers the correlation between the word vector and context feature by sharing parameters during the training process, and the experimental results fully verify that the joint learning model with shared underlying parameters can effectively improve the performance of the relation extraction task.

In terms of feature introduction, models such as SR-BRCNN and SDP-LSTM need to rely on the shortest dependency path and part-of-speech labeling, etc., generated by external NLP tools. These external features will improve the performance of the model to a certain degree and can integrate more natural languages grammatical information, but they also introduce errors generated by NLP tools into the model, affecting the performance. In the case of using only the objective features of the relative position of the entity, the entity type, etc., and not using any NLP tools, the BiLSTM-Multilevel-Attention model obtains a large increase in F_1 value, which fully proves the validity of the model.

In order to facilitate the analysis of the role played by the multilevel-attention mechanism in the relation extraction process, the weight matrix α representing the sentence-level information in the model is visualized. The following two examples are taken to intercept the representative sentence-level information:

1. As shown in Figure 10, in the relationship visualization example of “In the autumn, mother goes up the mountain and chops, when she comes back, her pocket always carries a few grains of sour jujube, or hawthorn that is the most extravagant snacks of childhood mother gives me.”, it can be seen that the matrix pays more attention to the words such as “pocket”, “inside” and “carries”, which greatly affects the entity pair {several jujube, and the pocket} to be judged as the social relationship.
2. As shown in Figure 11, in the relationship visualization example of “Sun Jinyou’s home is holding a blue tile in the south of Wuye’s house; and only a short wall away from his east neighbor Wang Endian”, it can be seen that the words “east”, “and” and “neighbor” have greatly influenced the entity pair “Sun Jinyou, Wang Endian” to be judged as the social relationship.

到了秋季, 母亲上山捡柴, 回来时, 衣兜里总是揣几粒酸枣, 或者山楂.....那就是母亲送给我童年最奢侈的零食。
 到了秋季, 母亲上山捡柴, 回来时, 衣兜里总是揣几粒酸枣, 或者山楂.....那就是母亲送给我童年最奢侈的零食。
 到了秋季, 母亲上山捡柴, 回来时, 衣兜里总是揣几粒酸枣, 或者山楂.....那就是母亲送给我童年最奢侈的零食。
 到了秋季, 母亲上山捡柴, 回来时, 衣兜里总是揣几粒酸枣, 或者山楂.....那就是母亲送给我童年最奢侈的零食。
 到了秋季, 母亲上山捡柴, 回来时, 衣兜里总是揣几粒酸枣, 或者山楂.....那就是母亲送给我童年最奢侈的零食。
 到了秋季, 母亲上山捡柴, 回来时, 衣兜里总是揣几粒酸枣, 或者山楂.....那就是母亲送给我童年最奢侈的零食。
 到了秋季, 母亲上山捡柴, 回来时, 衣兜里总是揣几粒酸枣, 或者山楂.....那就是母亲送给我童年最奢侈的零食。
 到了秋季, 母亲上山捡柴, 回来时, 衣兜里总是揣几粒酸枣, 或者山楂.....那就是母亲送给我童年最奢侈的零食。
 到了秋季, 母亲上山捡柴, 回来时, 衣兜里总是揣几粒酸枣, 或者山楂.....那就是母亲送给我童年最奢侈的零食。
 到了秋季, 母亲上山捡柴, 回来时, 衣兜里总是揣几粒酸枣, 或者山楂.....那就是母亲送给我童年最奢侈的零食。

Figure 10. Instance of attention matrix visualization 1.

孙进友家长着瓦松的青瓦房挡在五爷家南面;与东邻王恩家只一截矮墙之隔。
 孙进友家长着瓦松的青瓦房挡在五爷家南面;与东邻王恩家只一截矮墙之隔。
 孙进友家长着瓦松的青瓦房挡在五爷家南面;与东邻王恩家只一截矮墙之隔。
 孙进友家长着瓦松的青瓦房挡在五爷家南面;与东邻王恩家只一截矮墙之隔。
 孙进友家长着瓦松的青瓦房挡在五爷家南面;与东邻王恩家只一截矮墙之隔。
 孙进友家长着瓦松的青瓦房挡在五爷家南面;与东邻王恩家只一截矮墙之隔。
 孙进友家长着瓦松的青瓦房挡在五爷家南面;与东邻王恩家只一截矮墙之隔。
 孙进友家长着瓦松的青瓦房挡在五爷家南面;与东邻王恩家只一截矮墙之隔。
 孙进友家长着瓦松的青瓦房挡在五爷家南面;与东邻王恩家只一截矮墙之隔。
 孙进友家长着瓦松的青瓦房挡在五爷家南面;与东邻王恩家只一截矮墙之隔。

Figure 11. Instance of attention matrix visualization 2.

It can be seen from the above examples that the hierarchical attention matrix tends to focus on semantic words that have a more direct influence on the classification of the entity relationship. This is similar to the way a human being pays attention to associated words with a given entity to judge a relationship.

In order to discuss the influence of the hyperparameter L value on the relation extraction performance, the effect of the L value on the performance of the relation extraction model is studied using the same features. The results are shown in Table 7.

Table 7. Results of different L.

L	5	10	30	50	70	100	200
F_1	71.3	71.5	73.1	73.6	72.7	71.9	70.3

It can be seen that when the L value is less than 50, the performance of the model increases with the increase in the L value, and extracting more information of the sentence helps the performance of the relation extraction task, but when L is greater than 50, the performance of the model decreases with the increase in the L value, which may be due to the excessive redundancy of the sentence information. To verify this conjecture, this paper makes statistics on the sentence length of relationship instance in the data set, as shown in Figure 12.

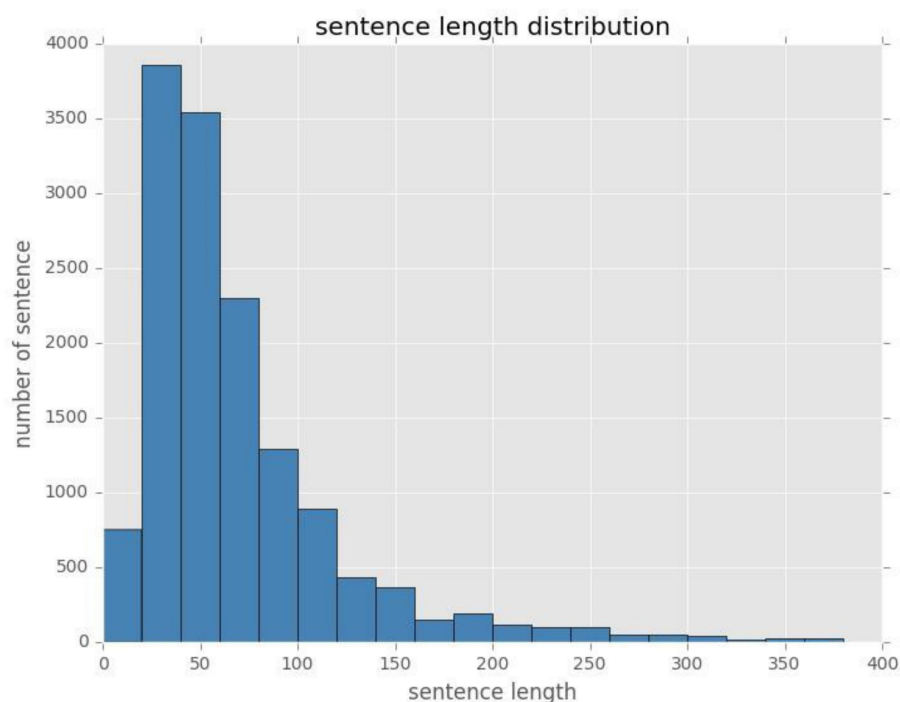


Figure 12. Instance of attention matrix visualization 2.

It can be seen from Figure 12 that the sentence length of the data set is mostly concentrated in the interval [30, 70], which matches the experimental results of the model performance when the L values are 30, 50, 70 in Table 7. This verifies that the influence of L value on experimental performance depends on the distribution of the sentence length in the task. This paper also pays attention to the influence of relative entity position, entity type feature and Chinese character stroke order feature on model performance, and the experimental results are shown in Table 8. It can be seen from the experimental results that the entity type feature is the most obvious factor for the improvement of the performance of the model, and the relative position of the word is the second. The experimental result is consistent with the linguistic law. The relative position of the word and the entity can reflect the importance of the word to the entity's classification of the relationship to a certain degree. The effect of the entity-type feature on relationship classification is more direct. For example, the relationship type of entity pair of <Thing, Location> may be located or near, but it may not be social, family, etc., because some types of entity pairs have limited the possible relationship classification results. To verify this conjecture, this paper makes statistics on the entity pair type and its corresponding relationship classification of relationship instances in the data set, and the result is shown in Figure 13.

Table 8. Results of relation extraction.

Feature	Word Embeddings	+Relative Position	+Entity Type	+Stroke Embeddings	+All
F_1	67.4	69.8 (+2.4)	71.4 (+4.0)	67.9 (+0.5)	73.6

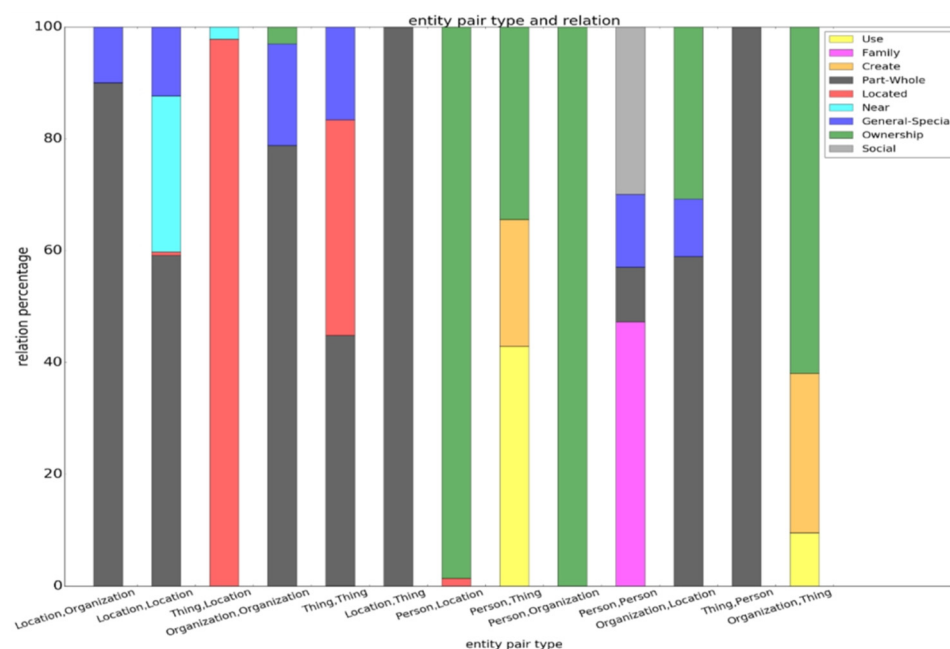


Figure 13. Distribution of entity pair type and relationship.

It can be clearly seen from the figure that the distribution of the entity pair type and its corresponding relationship type is completely uneven; some entity pair types only correspond to some of the relationship types. It can be seen that the entity pair type has a strong constraint on the relationship classification result, and its performance improvement on the relation extraction model is more obvious than other features.

In general, the neural network model that introduces the self-attention mechanism improves the performance of the named entity recognition and relation extraction model, and joint learning can significantly improve the performance of the relation extraction task. Joint training for the performance improvement of the named entity recognition task is not significant, because the underlying representation of the model in relation extraction is more concerned with information that affects the semantic relationship of a given entity, and this part of the information is not significant for the search for entities in the named entity recognition task. The joint training has a significant performance improvement for the relation extraction task because the way of sharing the underlying parameters with the named entity recognition model is beneficial for the relation extraction task to have more underlying representation information of named entities in the sentence, which contributes to the extraction of abstract representation of the sentence, thereby improving the performance of the relation extraction task.

5. Conclusions

In view of the shortcomings of current named entity recognition and relation extraction research, this paper proposes a Chinese named entity recognition model based on BiLSTM-Self-Attention-CRF, a relation extraction model based on BiLSTM-Multilevel-Attention and a joint learning method to share the underlying representations of the two models. By introducing the self-attention mechanism into the Chinese named entity recognition and relation extraction task, and sharing the underlying representation, we can improve the performance of the models. In order to verify the effect of the model, we carried out experiments in the Chinese data set with the largest task of both named entity recognition and relation extraction. The results show that the named entity recognition model proposed in this paper is superior to the current mainstream models and data set baseline standards. The performance of the named entity recognition module in the joint learning model is also slightly improved, but it is not significant. The F_1 value of the relation extraction model proposed in this paper is 7.7 higher than the SR-BRCNN model [29], which is currently

optimal in the data set relation extraction task. The F_1 value is greatly improved without using any NLP tools to introduce features. The effectiveness of the relation extraction model is verified. In the joint learning model, the F_1 value of the relation extraction module is improved by 5.6 compared to the separately trained model, so the way of sharing the underlying parameters with the named entity recognition model is beneficial for the relation extraction task to have more underlying representation information of the named entities in the sentence, thereby improving the performance of the relation extraction task. Future research work will continue in the following aspects:

1. Exploring the capability of using artificial feature functions to improve the performance of named entity recognition, such as complex organizational names;
2. Optimize the characteristics of the Chinese and English corpora in other fields to improve the universality of the model.

Author Contributions: Methodology, L.-X.L., L.L. and E.L.; Software, W.-S.W. and G.-Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been partly supported by Project U1711264 of National Natural Science Foundation of China Project.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aone, C.; Ramos-Santacruz, M. REES: A large-scale relation and event extraction system. In Proceedings of the Sixth Conference on Applied Natural Language Processing, Washington, DC, USA, 29 April–4 May 2000; pp. 76–83.
2. Bikel, D.M.; Schwartz, R.; Weischedel, R.M. An algorithm that learns what's in a name. *Mach. Learn.* **1999**, *34*, 211–231. [\[CrossRef\]](#)
3. Bender, O.; Och, F.J.; Ney, H. Maximum entropy models for named entity recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, Edmonton, AB, Canada, 31 May–1 June 2003; Volume 4, pp. 148–151.
4. McCallum, A.; Li, W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, Edmonton, AB, Canada, 31 May–1 June 2003; Volume 4, pp. 188–191.
5. Jiang, J.; Zhai, C.X. A systematic exploration of the feature space for relation extraction. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference, Rochester, NY, USA, 22–27 April 2007*; Omnipress Inc.: Madison, WI, USA, 2007; pp. 113–120.
6. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
7. Chiu, J.P.C.; Nichols, E. Named entity recognition with bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 357–370. [\[CrossRef\]](#)
8. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural Architectures for Named Entity Recognition. In Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 260–270.
9. Ma, X.; Hovy, E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 1064–1074.
10. Rei, M.; Crichton, G.K.; Pyysalo, S. Attending to characters in neural sequence labeling models. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics, Osaka, Japan, 11–16 December 2016; pp. 309–318.
11. Bharadwaj, A.; Mortensen, D.R.; Dyer, C.; Carbonell, J.G. Phonologically aware neural model for named entity recognition in low resource transfer settings. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 1462–1472.
12. Peng, N.; Dredze, M. Named entity recognition for chinese social media with jointly trained embeddings. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 548–554.
13. Peng, N.; Dredze, M. Improving named entity recognition for chinese social media with word segmentation representation learning. In Proceedings of the Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 149–155.
14. Chen, X.; Qiu, X.; Zhu, C.; Liu, P.; Huang, X.J. Long short-term memory neural networks for chinese word segmentation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1197–1206.
15. Zhang, Q.; Qian, J.; Guo, Y.; Zhou, Y.; Huang, X.J. Generating Abbreviations for Chinese Named Entities Using Recurrent Neural Network with Dynamic Dictionary. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 721–730.

16. Dong, C.; Zhang, J.; Zong, C.; Hattori, M.; Di, H. Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition. In Proceedings of the International Conference on Computer Processing of Oriental Languages National CCF Conference on Natural Language Processing and Chinese Computing, Kunming, China, 2–6 December 2016; pp. 239–250.
17. Ni, J.; Florian, R. Improving Multilingual Named Entity Recognition with Wikipedia Entity Type Mapping. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 1275–1284.
18. Nguyen, T.H.; Grishman, R. Relation extraction: Perspective from convolutional neural networks. In Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, Denver, CO, USA, 5 June 2015; pp. 39–48.
19. Santos, C.N.; Xiang, B.; Zhou, B. Classifying relations by ranking with convolutional neural networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015; Volume 1, pp. 626–634.
20. Xu, Y.; Mou, L.; Li, G.; Chen, Y.; Peng, H.; Jin, Z. Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Paths. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1785–1794.
21. Cai, R.; Zhang, X.; Wang, H. Bidirectional recurrent convolutional neural network for relation classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; Volume 1, pp. 756–765.
22. Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; Xu, B. Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; Volume 2, pp. 207–212.
23. Wang, L.; Cao, Z.; De Melo, G.; Liu, Z. Relation classification via Multilevel-Attention cnns. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; Volume 1, pp. 1298–1307.
24. Zeng, D.; Liu, K.; Chen, Y.; Zhao, J. Distant supervision for relation extraction via piecewise convolutional neural networks. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1753–1762.
25. Miwa, M.; Bansal, M. 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, Berlin, Germany, 7–12 August 2016; pp. 1105–1116.
26. Zheng, S.; Hao, Y.; Lu, D.; Bao, H.; Xu, J.; Hao, H.; Xu, B. Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing* **2017**, *257*, 59–66. [[CrossRef](#)]
27. Zheng, S.; Wang, F.; Bao, H.; Hao, Y.; Zhou, P.; Xu, B. 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, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1227–1236.
28. Xu, J.; Wen, J.; Sun, X.; Su, Q. A Discourse-Level Named Entity Recognition and Relation Extraction Dataset for Chinese Literature Text. *arXiv* **2017**, arXiv:1711.07010.
29. Wen, J.; Sun, X.; Ren, X.; Su, Q. Structure Regularized Neural Network for Entity Relation Classification for Chinese Literature Text. *arXiv* **2018**, arXiv:1803.05662.