ISSN: 2664-9640

Deep Learning-Based Distantly Supervision Relation Extraction

Guanyu Lai, Xu Zhang and Haozhuo Tong

School of Electronics and Information, Xijing University, Xi'an 710000, China

Abstract

Relation extraction(RE) is a key technology of information extraction, and its purpose is to mine the semantic relations existing between entities. Relation extraction is of great significance for the automatic construction of knowledge bases, question answering systems and other fields. Aiming at the high cost of data annotation of traditional RE methods and the weak ability of existing RE models to extract text semantic representation and features, a remote supervised relation extraction method is proposed to enhance text semantic representation and feature extraction capabilities. The model uses a BERT pretrained model as an embedding layer to enhance the semantic representation of text, and enhances feature extraction through a Bidirectional Long Short-Term Memory (bilstm) neural network. The experimental results show that the model has better performance when dealing with relation extraction tasks.

Keywords

Relationship extraction, BERT, BiLSTM.

1. Introduction

Information extraction is a subfield in the field of natural language processing, and its goal is to mine structured information from unstructured data. Relation extraction is a key technology of information extraction, and its purpose is to mine the semantic relations existing between entities. Relation extraction is of great significance for the automatic construction of knowledge bases, question answering systems and other fields. The relationship between entity pairs can be represented in the form of a triple <head entity, relationship, tail entity>, where the head entity and the tail entity are entity types, and the relationship is a relationship description. Relation extraction is to extract relation triples <head entity, relation, tail entity> from unstructured text information to extract text information. For example, "Martin·Eberhard is the co-founder of Tesla", the entities "Martin·Eberhard", "Tesla" and the relationship "Originator" can be extracted.

Since the RE task was first proposed at the MUC-7 conference, the current RE methods include supervised learning method, semi-supervised learning method and distantly supervised learning method. The RE task based on supervised learning can use the high-quality labeled training data to obtain a good extraction effect, but the high-quality labeled data requires high manpower and material resources, and it is difficult to achieve a great improvement at present; based on semi-supervised learning entity relation extraction only needs to iteratively train a small number of labeled samples and a large number of unlabeled samples to obtain a classification model, which can not only reduce the dependence on labeled corpus to a certain extent, but also obtain higher accuracy, but this method has semantics Elegance and other problems are also easily affected by the quality of the initial relationship seed; for the problem of corpus dependence, Mintz et al. [1] first proposed the use of remote supervision to achieve entity relationship extraction. The paper assumes that if two entities exist in a known knowledge base a relationship, then all sentences that contain both entities also express this relationship.

ISSN: 2664-9640

2. Relate Work

Although remote supervision alleviates the difficulty of manually annotating data to a certain extent, this method also brings two main problems: one is that sentences containing entity pairs may not express the relationships in the knowledge base; the other is that two An entity pair may have multiple relationships, and it is impossible to tell which one it is. Therefore, Riedel et al. [2] proposed the express-at-least-once hypothesis, which assumes that among all sentences containing the same entity pair, at least one of them explicitly expresses the relationship between them, to alleviate the current assumption with strong constraints. With the development of neural networks, deep learning has achieved relatively good results in many natural language processing tasks in recent years. Compared with classical extraction methods, the main advantage of deep learning-based relation extraction methods is that deep learning neural network models can automatically learn sentence features without complex feature engineering, resulting in better performance. Liu et al. [3] uses the input of a convolutional neural network by mapping each word to a low-dimensional word vector. Zhuo et al. [4] used a bidirectional long short-term memory network to solve the long-range dependency problem of sequences. After Google released the pretrained model BERT [5], the model has been shown to achieve excellent performance on most NLP tasks.

Therefore, based on the analysis of existing relation extraction methods, this paper introduces structural knowledge and semantic knowledge to generate knowledge-aware word embedding vectors, and proposes the BERT-BiLSTM network structure model, which can not only handle the identification of different relations, but also Effectively learn the information and structure of long-range semantics of texts. The main contributions of this paper are as follows:

(1) In order to improve the semantic representation and feature extraction ability of the text, it is proposed to use the BERT pre-training model as the embedding layer instead of Word2vec to train the initial word vector. BERT can comprehensively consider the context-related information in the sequence to obtain an initial word vector with stronger semantic representation, and use BiLSTM to implement context encoding to extract more effective feature text and comprehensively improve the joint extraction effect of model entity relationships.

(2) We evaluate our BERT-BiLSTM relation extraction model on the benchmark dataset ACE2005 benchmark dataset, and the experimental results show that our model significantly outperforms other methods on this dataset.

3. BERT-BiLSTM Relaiton Extraciton Model

The improvement of semantic representation and feature extraction ability is the key to improve the performance of relation extraction models. In the process of RE, if the model used is difficult to fully mine the semantic features of the text, the initial word vector obtained through the training of the embedding layer is usually fused with the artificially constructed feature vectors such as part of speech and grammar to improve the semantic representation ability of the word vector. To a certain extent, the effect of entity relation extraction is improved. However, this method increases the computational cost of the model training process and is less versatile.

To address the above issues, this paper investigates how to improve the semantic representation and feature extraction capabilities of text without adding other artificial features and without relying on any external NLP processing tools to extract lexical or syntactic features. And a relation extraction model with enhanced text semantic representation and feature extraction capability is proposed. The model structure is shown in Figure 1.

ISSN: 2664-9640

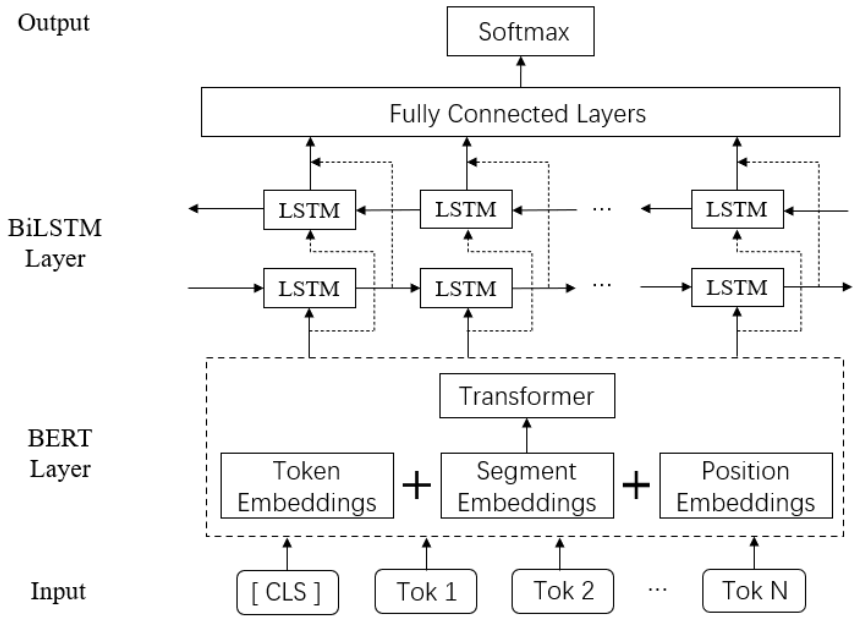

Figure 1: Framework of relation extraction model based on BERT-BiLSTM

The text semantic representation and feature extraction capability enhancement module consists of BERT layer and BiLSTM layer from bottom to top. The principle and function of each layer are described in detail below.

3.1. BERT Layer

BERT is a natural language processing pre-trained language representation model released by Google. It consists of multi-layer bidirectional Transformer coding units. Each coding unit contains a self-attention layer and a feedforward neural network layer. The Transformer coding unit structure is shown in the figure 2.

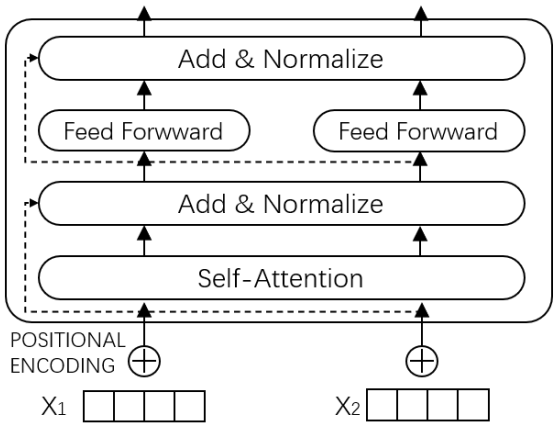

Figure 2: Transformer coding unit structure

Among them, the self-attention layer is the most important part of the Transform model. It considers the semantic and grammatical connection between each word by calculating the association between different words in the sentence. The calculation method is shown in equation (1).

ISSN: 2664-9640

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

First, self-attention randomly initializes three vectors Q, K, and V, and performs vector multiplication with its own Q value and the K value of each other word in the sentence; then, after dividing the multiplication result by a constant, perform softmax calculate to get the correlation of each word to the word at the current position; finally, multiply the value obtained by V and softmax to get the final score of self-attention at the current node.

In order to allow the model to learn relevant information in different representation subspaces, BERT uses Multi-head attention to perform multiple self-attention operations on each word in the input sequence. The specific calculation method is as shown in equation (2) and (3). The above is the main algorithm implementation of BERT.

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (2)$$

$$\text{Multihead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \quad (3)$$

3.2. BiLSTM Layer

After the input sequence is mapped to the initial word vector by the BERT embedding layer, it is contextually encoded by the BiLSTM layer to extract more useful contextual features.

BiLSTM is a combination of forward LSTM and backward LSTM. The LSTM model is an improved model proposed by Hochreiter in 1997 for the problem of gradient disappearance and gradient explosion of recurrent neural network(RNN). LSTM cleverly introduces the "gate" mechanism. Each "gate" structure contains a sigmoid network neural layer and a point-by-point multiplication operation to control whether information can pass through, thereby removing or enhancing information about the cell's state. LSTM consists of a series of repeated sequential modules, each module contains three "gates" and a memory unit, namely forget gate, input gate and output gate. The specific structure is shown in Figure 3.

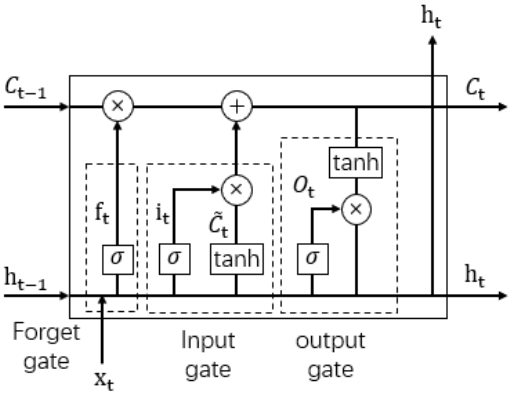

Figure 3: Schematic diagram of LSTM structure

The forget gate determines what information the cell will discard, reads h_{t-1} and x_t, and outputs a value between 0 and 1 for each state in the cell C_{t-1} middle.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (4)$$

ISSN: 2664-9640

The input gate decides what information to store in the cell state, and there are two parts here. First, a sigmoid neural network layer decides what values will be updated, called the "input gate layer". Then, a tanh layer creates a new vector of candidate values $\tilde{C}_t$ and adds it to the state.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{5}$$

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{6}$$

When the unit information is updated, the old state is multiplied by f_t, irrelevant information is discarded, and $i_t * \tilde{C}_t$ is added to form a new candidate value.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{7}$$

The output gate works by running a sigmoid layer to determine which part of the cell state will be output, then passes the cell state through the tanh function to get a value between -1 and 1, and multiplies it with the output of the sigmoid gate, Only the part that determines the output is finally output.

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{8}$$

$$h_t = o_t * \tanh(C_t) \tag{9}$$

Among them: tanh( ) represents the activation function, σ represents the sigmoid neural network layer, x_t is the unit state input at time t; f_t,i_t,O_t represent the settlement results of the forget gate, input gate, and output gate, respectively; W_f,W_i,W_o,W_c represents the forgetting gate, input gate, output gate and updated weight respectively; b_f,b_i,b_o,b_c are the corresponding offsets.

In the process of relation extraction, in order to make full use of the contextual information of the text, a bi-directional long short-term memory network BiLSTM will be used, that is, two LSTM models with opposite time series will be combined.

$$h_t = \text{LSTM}(h_{t-1,}W_t, c_{t-1}), t \in [1, T] \tag{10}$$

$$h_t = \text{LSTM}(h_{t+1,}W_t, c_{t+1}), t \in [T, 1] \tag{11}$$

$$H_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{12}$$

Among them, H_t is the text feature vector output by the BiLSTM model.

4. Experiment

4.1. Relation Extraction Datasets And Metrics

The model in this paper is tested on the benchmark dataset ACE2005 benchmark dataset. The ACE2005 corpus is various types of data composed of entity, relation and event annotations published by the Language Data Consortium (LDC), including English, Arabic and Chinese

ISSN: 2664-9640

training data, with the goal of developing automatic content extraction techniques that support automatic processing in text form human language. As shown in Table 1, the dataset is predefined with 7 entity types and 6 relation types.

Table 1: Data Annotation Entries and Interpretation

entity type	relationship type
FAC	ART
LOC	ORG-AFF
ORG	GEN-AFF
PER	PHYS
WEA	PERT-SOC
GPE	PART-WHOLE
VEH	

In order to evaluate the quality of the model, this paper uses the F1/E curve (F1 score/Epoch) as the evaluation indicators.

4.2. Experiment Parameters

In this experiment, the parameters of the relation extraction model are shown in Table 2.

Table 2: Experimental parameters

Parameter	Scheme 2
Batch_size	50
Embedding_dim	768
Label_num	7
Adam Learning_rate	0.02
Number of epoch	50
Droupout_rate	0.5
Max seentence length	300

4.3. Analysis Of Results

In order to verify the effectiveness of the BERT pre-training model in the NLP task and the extraction effect of the BERT-BiLSTM model in the relation extraction task, this paper builds a model based on the BERT and BERT-BiLSTM relation extraction models, and uses them in the ACE2005 relation extraction dataset. The models are compared and the experimental results are shown in Figure 4.

According to the analysis of the graph, the graph shows the F1 update of the above model in the first 30 epochs of training. As can be seen from the figure, BERT can quickly reach a relatively high level during the training process, and can remain relatively stable. Although the performance of the BERT-BiLSTM model is not as good as that of BERT in the early training stage, the BERT-BiLSTM model achieves better performance and remains stable after 15 epochs of training. It also means that our model outperforms the BERT model on relation extraction tasks.

ISSN: 2664-9640

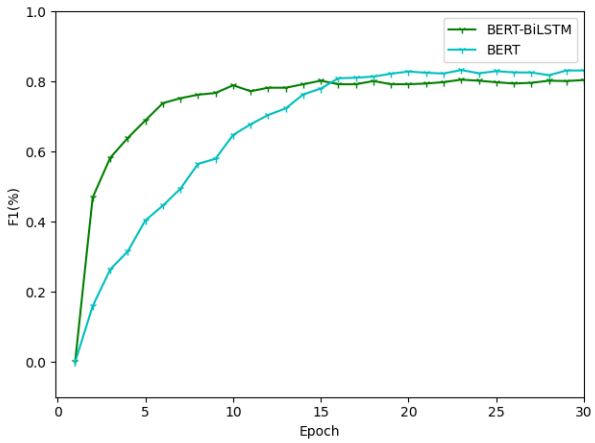

Figure 4: F1 update situation map

5. Concluding Remarks

Relation extraction is a key problem in natural language processing. After Google proposed the BERT pre-training model, the traditional language processing model in the relation extraction task could not effectively represent the contextual semantic information in the text, and could not handle the difference of ambiguity. The context problem is effectively resolved. This paper proposes a BERT-BiLSTM relation extraction model. Using the BERT-pretrained model as the embedding layer of the model, the input text is generated, and word vectors representing contextual semantic information are generated. At the same time, the BiLSTM bidirectional long short-term memory neural network can process the features of the generated vectors more effectively and improve the performance of text information extraction. The comparative experimental results show that the model has achieved excellent results on the ACE2005 corpus, and has achieved good results in comparative experiments with other models.

References

[1] Mintz M,Bills S,Snow R, Jurafsky D, et al. Distant supervision for relation extraction without labeled data, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, vol. 2 (2009), 1003-1011.

[2] Riedel S,Yao L,McCallum A, et al. Modeling Relations and Their Mentions without Labeled Text, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, vol. 6323 (2010), 148-163.

[3] CY Liu,WB sun,WH Chao, et al. Convolution Neural Network for Relation Extraction, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, vol. 8347 (2013), 231-242.

[4] Zhuo P,Shi W,Tian J, et al. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 2 (2016), 207-212.

[5] Devlin,Jacob,Ming W, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,Arxiv Preprint,2018.