Mathematical Subject Information Entity Recognition Method Based on BiLSTM-CRF

Haoze Li[1,2], Tianwei Xu[2,3](✉), and Juxiang Zhou[2,3]

[1] School of Information Science and Technology, Yunnan Normal University, Kunming, China
[2] Key Laboratory of Education Informalization for Nationalities,
Yunnan Normal University, Kunming, China
xutianwei@ynnu.edu.cn
[3] Yunnan Key Laboratory of Smart Education, Yunnan Normal University,
Kunming 650500, China

Abstract. Combining language conditional random field (CRF) and bidirectional long-term and short-term memory (BiLSTM) networks, a mathematical subject information entity recognition method based on BiLSTM-CRF is constructed to extract entity information in mathematical language. Experimental results show that compared with BiLSTM, BiLSTM-CRF improves the recall rate by nearly 5%, the accuracy rate by nearly 2%, and the F1 value by nearly 4%. The results of the BERT-CRF model are also significantly better than other models.

Keywords: Mathematical subject · Entity recognition · Deep learning

1 Introduction

Natural language processing is an interdisciplinary subject that combines computer science, artificial intelligence, and linguistics. In natural language processing, deep neural network methods are used in tasks such as named entity recognition [1], text classification [2], information extraction [3], machine translation [4], sentiment analysis [5], and question answering system [6]. Very good results have been achieved.

Mathematics is a very important subject in subject education. Mathematical language contains many theorems, conclusions and methods. It hides important information, but it is often overlooked, which makes our study of theoretical concepts not deep enough or the understanding of the topic deviate. For example, the classic zero point theorem:y = f(x) is a continuous function on the interval a, b and f(a) and f(b) are different signs, then there is at least one f(x) makes f(c) = 0 hold. Continuous functions, different signs, zero points and other information can make us quickly find a breakthrough in solving problems when doing problems about this theorem, and then complete the problem. Therefore, in the field of mathematics, the recognition of mathematical information entities has very important research value and significance.

The name entity of the mathematics subject makes its name abstract and logically rigorous due to the characteristics of its own subject, which also makes the name recognition of mathematical information unique and complex. Compared with traditional

X. Chen et al. (Eds.): ML4CS 2020, LNCS 12488, pp. 259–268, 2020.
https://doi.org/10.1007/978-3-030-62463-7_24

recognition methods, deep neural networks are driven by data, which can automatically extract effective feature parts from them, and have obvious advantages when applied to unstructured, variable and unknown data. This paper presents a method of named entity recognition based on deep neural networks to identify and extract logical concepts such as objectively existing entities in mathematics and methods and theorems in mathematical information.

2 Ralated Work

2.1 CRF Model

The CRF model is a conditional random field model proposed by Lafferty et al. [7] to solve sequence labeling problems. Conditional random fields are widely used in natural language processing, especially in named entity recognition tasks. Results. In general, the conditional random field model can solve the long-term dependence problem between sequences when dealing with sequence problems and can fully learn the context information in the text. This probabilistic graphical model can solve the problem of labeling bias. Effect, so the current natural language processing field generally uses linear chain conditional random field model to solve the sequence labeling problem.

The principle of linear chain conditional random field application in named entity recognition task is shown in Fig. 1.

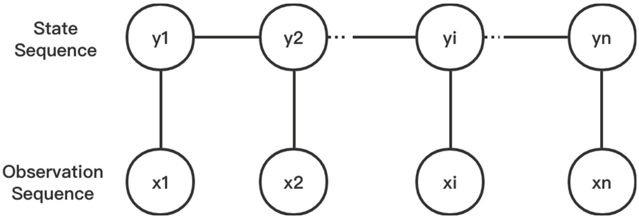

Fig. 1. Linear chain conditional random field structure

It can be seen from the figure that the observation sequence $\mathrm{x} = (\mathrm{x}_1, \mathrm{x}_2, \cdots \mathrm{x}_\mathrm{n})$ and the state sequence $\mathrm{y} = \left(\mathrm{y}_1, \mathrm{y}_2, \cdots, \mathrm{y}_\mathrm{n}\right)$, let P(y|x) be a linear chain conditional random field, then P(y|x) is a linear chain conditional random field, then the form of (y|x) is defined as:

$$P(y|x) = \frac{1}{Z(x)} exp\left\{\sum\nolimits_{i,k} \lambda_k f_k(y_{i-1}, y_i, x, i) + \sum\nolimits_{i,l} \mu_j h_j(y_i, x, i)\right\} \quad (1)$$

among them,

$$z(x) = exp\left\{\sum\nolimits_{i,k} \lambda_k f_k(y_{i-1}, y_i, x, i) + \sum\nolimits_{i,l} \mu_j h_j(y_i, x, i)\right\} \quad (2)$$

where λ_k and μ_j are corresponding weights, f_k and h_j are characteristic functions, and $z(x)$ is a normalization function.

Named entity recognition is actually to treat a sentence as a sequence to be observed, take each word in the sentence as a symbol, and set a state for each symbol, and finally maximize the λ_k and μ_j parameters through the training set training to find To meet the conditional probability to complete the sequence labeling.

Given a sequence of input states, the optimal state sequence can be obtained as shown in formula (3):

$$y^* = \arg maxP(y|x) \tag{3}$$

2.2 LSTM Model

Long short-term memory network (LSTM) [8] is to solve the recurrent neural network (RNN) [9] In the actual training process, due to the problem of gradient disappearance, it is often impossible to use information that is too far away. Generally speaking, the memory ability of RNN after layer 7 tends to zero, and the long-term and short-term memory network emerged to solve this problem [10]. Hochreiter introduces long- and short-term memory cells. The main idea is to store information in a memory cell. Update, attenuation, input and output in the memory cell will be controlled by multiple gates. Parameters to decide whether to save or forget information in the memory unit. The LSTM cell diagram is shown in Fig. 2.

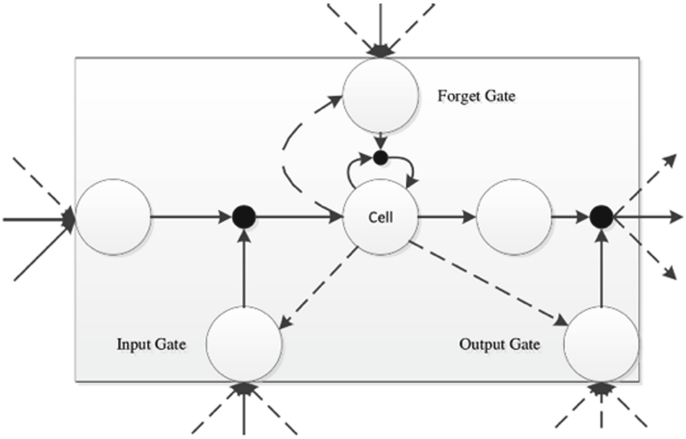

Fig. 2. LSTM structure

Let the input of LSTM at time t be $\mathrm{i_t}$, the hidden layer and memory unit at time $\mathrm{t-1}$ are $\mathrm{h_{t-1}}$ and $\mathrm{c_{t-1}}$, respectively, and output the hidden layer $\mathrm{h_t}$ and memory unit $\mathrm{c_t}$ at time t.

- Calculation gate information

$$\text{Input gate: } \mathrm{i_t} = \sigma(\mathrm{w_{xi}x_t} + \mathrm{w_{hi}h_{t-1}} + \mathrm{w_{ci}c_{t-1}} + \mathrm{b_i}) \tag{4}$$

$$\text{Forgotten door}: \mathrm{f_t} = \sigma(\mathrm{w_{xf}x_t} + \mathrm{w_{hf}h_{t-1}} + \mathrm{w_{cf}c_{t-1}} + \mathrm{b_f}) \tag{5}$$

$$\text{Output gate}: \mathrm{o_t} = \sigma(\mathrm{w_{xo}x_t} + \mathrm{w_{ho}h_{t-1}} + \mathrm{w_{co}c_t} + \mathrm{b_o}) \tag{6}$$

- Calculate the value of the memory unit

$$\mathrm{c_t} = \mathrm{f_t c_{t-1}} + \mathrm{i_t tanh}(\mathrm{w_{xc}x_t} + \mathrm{w_{hc}h_{t-1}} + \mathrm{b_c}) \tag{7}$$

- Calculate the value of the hidden layer at time *t*.

$$\mathrm{h_t} = \mathrm{o_t}\,\mathrm{tanh}(\mathrm{c_t}) \tag{8}$$

where *w* and *b* both represent parameters, and σ generally takes the sigmod function [11].

It can be seen from Fig. 2 and the above calculation formula that the input cell and the value of the memory cell without the gate are multiplied to input the input information to the memory cell. Forgetting the gate and multiplying the value at time t − 1 will get the attenuation of the memory unit. The output gate and the memory unit at time t are multiplied to output the information in the memory unit to the hidden layer, which affects the output of each gate at time t + 1.

2.3 BERT Model

BERT is a multi-layer bidirectional Transformer encoder based on fine-tuning. This model training requires massive data and powerful computing power to achieve. Google has open sourced two versions of the BERT model. This article uses Google to train Chinese corpus BERT Base version. The BERT model has two main tasks, namely input representation and pre-training tasks.

Input Indication

For different tasks, the model can represent a single text sentence or a pair of texts in a sequence of words. For a given word, the input representation can be composed by a three-part Embedding summation. The visual representation of Embedding is shown in Fig. 3.

Token Embeddings represent word vectors. In Chinese processing, they can be word vectors or word vectors. In this paper, word vectors are used in experiments, because word vectors conform to Chinese characteristics. Segment Embeddings is used to distinguish between two sentences when doing a classification task that takes two sentences as input. Position Embeddings is the position information obtained through model learning.

Pre-training Tasks

The BERT model uses two new unsupervised prediction tasks to preprocess BERT, which are Masked LM and next sentence prediction. The goal of pre-training is to build a language model. The BERT model uses a two-way Transformer. In order to train the deep two-way Transformer representation, a simple method is adopted: randomly cover some input words, and then predict those masked words. During the training process, the

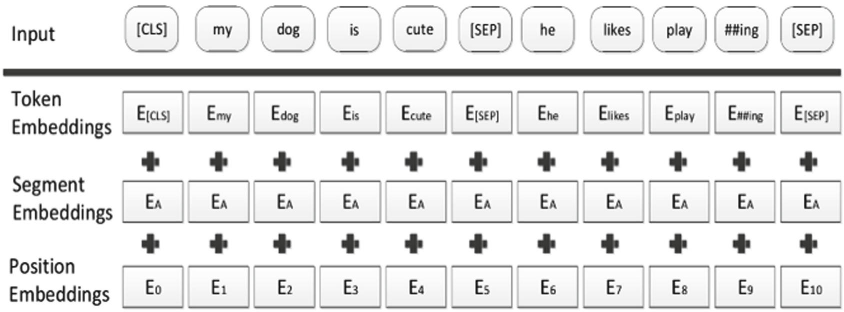

Fig. 3. BERT model input representation

original vocabulary of the word is randomly covered. It is different from the left-to-right language model pre-training. Masked LM randomly hides some words from the input. Its goal is to predict the original vocabulary of the masked word based on the context. It is different from the left-to-right language model pre-training. The representation learned by Masked LM can merge the left and right contexts. The bidirectional Transformer in the model does not know which words it will be required to predict, or which have been replaced by random words, so it must maintain a distributed contextual representation of each input word. In addition, since random replacement occurs only 1.5% of all words, it will not affect the model's understanding of the language.

2.4 BiLSTM-CRF Model

It can be seen that the long-short-term memory network basically solves the problem of the disappearance of gradients in the recurrent neural network, but when it is actually applied to natural language processing tasks, it will still be found that the long-short-term memory network model can only use the historical information of the foregoing, without considering The following text is also very important for the impact of the previous text, and for sequence labeling tasks, if the context information is not fully utilized, the prediction results will also have an impact. In response to this problem, scholars have proposed a bidirectional long-short-term memory network (Bidirectional LSTM, Bi LSTM) model. BiLSTM makes it possible to use contextual information at the same time, that is, information of the entire sequence.

The structure of BiLSTM is shown in Fig. 4. BiLSTM is composed of a forward LSTM network and a reverse LSTM network. Calculating the input sequence in two directions can make full use of the context information of the input sequence. The results of the calculation are simultaneously passed to the output layer for output.

3 The Proposed Model

In mathematics, the description of logical concepts such as methods and theorems usually has obvious hints, such as "XXXX method" and "XXXX theorem". Therefore, when judging whether a text sequence contains annotated entities, the start word can play

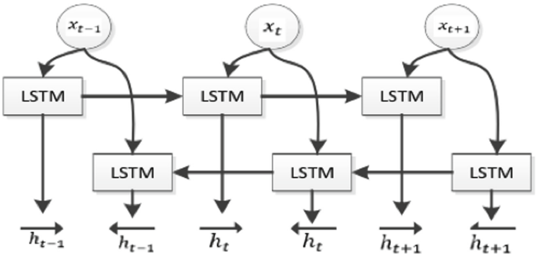
x_{t-1}
x_t
x_{t+1}
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
$\overrightarrow{h_{t-1}}$
$\overleftarrow{h_{t-1}}$
$\overrightarrow{h_t}$
$\overleftarrow{h_t}$
$\overrightarrow{h_{t+1}}$
$\overleftarrow{h_{t+1}}$

Fig. 4 BiLSTM structure

a very important role and can obtain the strong dependency relationship between the before and after text. The BiLSTM model can not only capture the dynamic information of the time series, but also use the context information of the current word, and finally obtain a better dependency relationship.

Combining BiLSTM with CRF, so that BiLSTM can be used to extract the context information in the text sequence, and the accuracy of annotation can also be improved by CRF annotation information at the entire sentence level. The structure of the BiLSTM-CRF model is shown in Fig. 5.

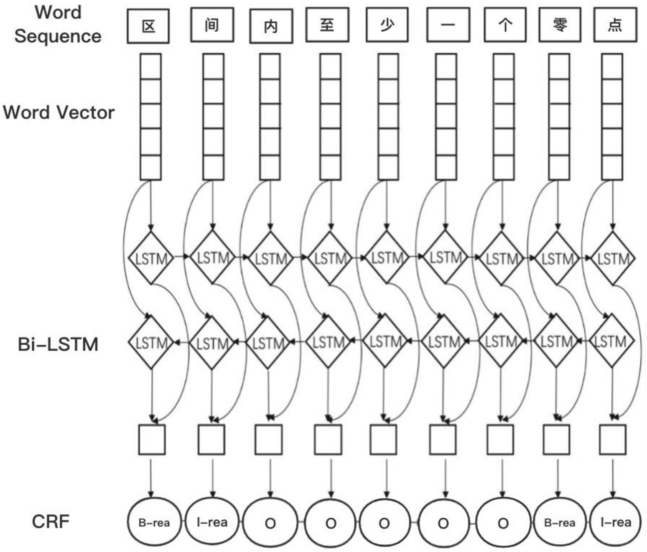

Fig. 5. BiLSTM-CRF model structure

The structure of BiLSTM-CRF model mainly includes word vector layer, BiLSTM network layer, and neural network language model CRF layer. The input of the model is sequence text, input according to each character, and the output is the label of each character, which represents whether it is part of the required entity. Each character in the input sequence is input into BiLSTM once through the word vector expression, and then the bidirectional expression of the text sequence containing the context information is established through the BiLSTM network. After the bidirectional expression of the BiLSTM neural network is obtained, it is merged and the combined expression After a layer of implicit exposure, input it into the CRF, and then calculate the label of each character in the sequence text through the CRF, compare it with the label to obtain the log likelihood of the input sequence, and then define it as the loss of the overall model. And in order to prevent overfitting, this experiment added a dropout layer to the model.

The important and complex information in mathematical information is mainly a large number of logical concepts such as the entities that actually exist in the subject and theorems and methods. In this paper, the BiLSTM-CRF model is applied to the information extraction of mathematics, the process is shown in Fig. 6. First, crawl relevant information from websites such as People's Education Textbooks and Baidu Encyclopedia, normalize each text and remove abnormal characters. Then, the entire text is cut through punctuation and paragraph control to form many small texts. The BiLSTM-CRF model is used to calculate the label sequence corresponding to each small text sentence. Finally, the final entity is found by combining the far text according to the label sequence. The algorithm of using BiLSTM-CRF model to extract information in mathematics is shown in Algorithm 1.

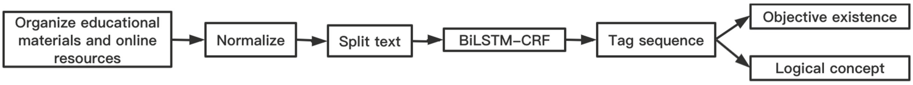

Fig. 6. Mathematics subject information extraction process

Algorithm 1 : Mathematics subject information extraction algorithm

```
Input: Enter the content of each mathematics document source
Output: Objectively existing set of reality, set of logical concepts in mathematical discipline
1. reality = {}, theoretical = {}; / * initialize the collection * /
2. source = normalize (source); / * Normalize mathematics discipline files * /
3. clips = token (source); / * Cut punctuation marks and paragraphs of the processed mathematics files to form text fragments and save them in the list clips * /
4. for clip in clips do; / * For each segment in clips * /
5. temp_reality, temp_theoretical = get_entity (clip); / * Enter each clips into the BiLSTM-CRF model to get the objectively existing entities and logical conceptual entities in the mathematical information * /
6. reality temp_reality; theoretical temp_theoretical; / * Put the objective knowledge points and logical concepts obtained in step 5 into reality, theoretical * /
7. end for;
8.return reality，theoretical；
```

4 Experimental Results and Analysis

4.1 Experimental Data

This experiment obtained a total of 10232 effective documents from the knowledge data shared by the People's Education Press electronic textbooks published by the People's Education Press, Mathematics Subject Network, Mathematics Resource Network, Mathematics China Network, Baidu Encyclopedia, and Wang Jianing, which basically covered primary and junior high schools, High school and advanced mathematics knowledge points, of which entity 4139, for these data, this article randomly selected according to the ratio of 7: 3 to form a training set and a test set.

4.2 Experimental Environment and Parameter Settings

In this experiments, the CPU is Intel 9750H, the memory is 16G, and the GPU is RTX1660Ti. The models used in this article are all built using Tensorflow, which is a deep learning framework developed by Google and widely used in the implementation of various machine learning algorithms. Some parameters are set as follows: dropout is 0.5, seq_length is 128, training learning rate is 3e−5, and Epochs is 30.

4.3 Experimental Results and Analysis

In the mathematics data set, HMM, CRF, BiLSTM, BiLSTM-CRF, BERT-CRF, BERT-softmax models were used for performance analysis. The experimental results are shown in Table 1 and Table 2.

Table 1. Comparison of the recognition results of the mathematical objective existence entities

	Recall	Precision	F1-score
HMM	82.15%	83.63%	83.38%
CRF	86.85%	89.70%	88.24%
BiLSTM	95.11%	96.44%	95.32%
BERT-softmax	38.37%	59.26%	46.58%
BERT-CRF	96.42%	95.73%	96.07%
BiLSTM-CRF	96.52%	94.79%	95.56%

Table 2. Comparison of the recognition results of the mathematical logic concept entities

	Recall	Precision	F1-score
HMM	82.38%	74.82%	78.41%
CRF	81.69%	88.05%	84.65%
BiLSTM	74.13%	88.39%	79.29%
BERT-softmax	27.16%	66.67%	38.60%
BERT-CRF	83.95%	89.47%	86.62%
BiLSTM-CRF	82.20%	91.11%	86.32%

It can be seen from Table 1 and Table 2 that:

- The method based on recurrent neural network (such as BiLSTM) is generally better than the method based on HMM, because the model based on RNN can effectively extract sequence features.
- Comparing the combination of CRF and deep neural network, it can be seen that the recall rate of BiLSTM-CRF is increased by nearly 5% compared to BiLSTM, the accuracy is increased by nearly 2%, and the F1 value is increased by nearly 4%. Compared with BERT-softmax, BERT-CRF has also been significantly improved. It can be seen that the addition of deep neural networks can provide better sequence characteristics for the model, thereby improving the model's effect.

5 Conclusion

In this paper, deep neural networks are applied to the recognition of information entities in mathematics, and a BiLSTM-CRF model is constructed to identify the objectively existing entities and logical concept entities in mathematical information. The Chinese text is vectorized word by word, and the bi-directional semantic features of the pretext and posttext are obtained using the BiLSTM network. Experimental results show that the BiLSTM-CRF model has a better recognition effect than HMM, CRF, and BiLSTM. In the future, we will try to construct an entity relationship extraction method for mathematics information.

Acknowledgments. We thank the anonymous reviewers. This work is supported by Doctoral Research launch project of Yunnan Normal University (No. 2019XJLK21), and Program for innovative research team (in Science and Technology) in University of Yunnan Province.

References

1. Cho, M., Ha, J., Park, C., Park, S.: Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognitio. J. Biomed. Inform. **103**, 103381 (2020)

2. Celardo, L., Everett, M.G.: Network text analysis: a two-way classification approach. Int. J. Inf. Manage. **51**, 102009 (2020)
3. Haihong, E., Xiao, S., Song, M.: A text-generated method to joint extraction of entities and relations. Appl. Sci. **9**(18), 3795 (2019)
4. Farhan, W.: Unsupervised dialectal neural machine translation. Inf. Process. Manage. **57**(3), 102181 (2020)
5. Huang, M., Xie, H., Rao, Y., Feng, J., Wang, F.L.: Sentiment strength detection with a context-dependent lexicon-based convolutional neural network. Inf. Sci. **520**, 389–399 (2020)
6. Kodra, L., Meçe, E.K.: Question answering systems: a review on present developments, challenges and trends. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **8**, 217–224 (2017)
7. Graves, A.: Long short-term memory. In: Supervised Sequence Labelling with Recurrent Neural Networks, pp. 1735–1780. Springer, Berlin (2012). https://doi.org/10.1007/978-3-642-24797-2_2
8. Mirza, A.H., Kerpicci, M., Kozat, S.S.: Efficient online learning with improved LSTM neural networks. Digital Sig. Process. **102**, 102742 (2020)
9. Lafferty, J.D., Mccallum, A., Pereira, F.C.N.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., pp. 282–289 (2001)
10. Hocreiters, S.J.: Long short-termmemory. Neural Comput. **9**(8), 1735–1780 (1997)
11. Graves, A., Jaitly, N., Mohamed, A.R.: Hybrid speech recognition with deep bidirectional LSTM. In: Proceedings of 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 273–278. IEEE Press, Washington, D. C. (2013)