Electric Circuits Course Knowledge Named Entity Recognition Based on Enhanced Word Embedding

Nan Wang(✉), Dong Liang, and Ruolin Dou

Wireless Signal Processing and Network Lab, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China
wangnan0123@foxmail.com

Abstract. This paper conducts named entity recognition research on electric circuits course knowledge, realizing entity extraction from unstructured text data in a specific field, aiming to achieve the effective utilization of subject knowledge data. This paper proposes a word embedding enhanced BiLSTM-CRF model. Feature vectors based on the keyword dictionary of electric circuits course are constructed to make better use of domain text information. Word embedding enhancement includes two aspects. One is static feature vectors, which are composed of static word embedding, POS feature vectors, and feature vectors combined with keyword dictionary. The other is context word embedding, which uses an attention mechanism to fuse static feature vectors and context word embedding. Through comparative experiments and result analysis, the F1 score of the model with enhanced word embedding has increased by 3.18%, proving the effectiveness of using static feature vectors and contextual word embedding.

Keywords: electric circuits course · named entity recognition · feature vector · enhanced word embedding

1 Introduction

The aim of named entity recognition (NER) task is identifying the entities contained in a given text data. It can effectively utilize unstructured text data, and its results can be used in various downstream tasks, such as text segmentation, question answering, relationship extraction, etc. Currently, the Chinese public dataset for NER mainly includes fields such as journalism, medicine and finance, while there is little research on other vertical fields. In order to achieve the goal of educational digitization and improve the quality of educational resources, NER of course knowledge text data can extract valuable information from it, thereby making better use of data in the field of education.

The early use of NER tasks was based on manually defined rules or statistics. Neural network models have become increasingly popular in named entity recognition tasks in recent years, leading to significant improvements in recognition performance. The commonly used models in research include CNN [1], RNN [2], LSTM [3], etc. Deep

P. Qi and Z. Chen (Eds.): ICBDIE 2023, AISR 178, pp. 679–686, 2024.
https://doi.org/10.2991/978-94-6463-238-5_89

learning excels in acquiring strong semantic analysis capabilities through the use of vector representation and neural network processing. Compared to linear models, models that integrate deep learning possess the capacity to extract more intricate features from data by leveraging nonlinear activation functions.

This paper mainly studies NER tasks in the field of electric circuits course. The main purpose is to identify specific nouns in the text data of electric circuits course. This paper proposes a word embedding enhanced BiLSTM-CRF model. For better utilization of domain text data, we construct feature vectors based on the keyword dictionary of electric circuits course. Additionally, static feature vectors are obtained by splicing static word embedding and POS feature vectors. Finally, the fusion of static feature vectors and word embedding containing context information use the attention model. Through experimental comparison, the effectiveness of this model has been verified, and the effect of the enhanced BiLSTM-CRF model with word embedding has been improved.

2 Methodology

This section uses some pre-trained models to vectorize the input text, using part of speech features, and also using domain corpus to construct a dictionary to extract some domain related information of the input sequences. In addition, attention mechanisms are used to fuse the above vector representation, then the modification of word embedding layer is completed. For the above steps, Fig. 1 shows its specific technical route.

2.1 Static Feature Vectors

Static Word Embedding

This paper utilizes the Word2Vec [4] model to process input text sequences and generate a vector representation of the input, resulting in lower dimensional vector representations. The Word2Vec model is first trained using Chinese electric circuits course corpus, and then the model is used to vectorize the Chinese input text sequences. The input sequence

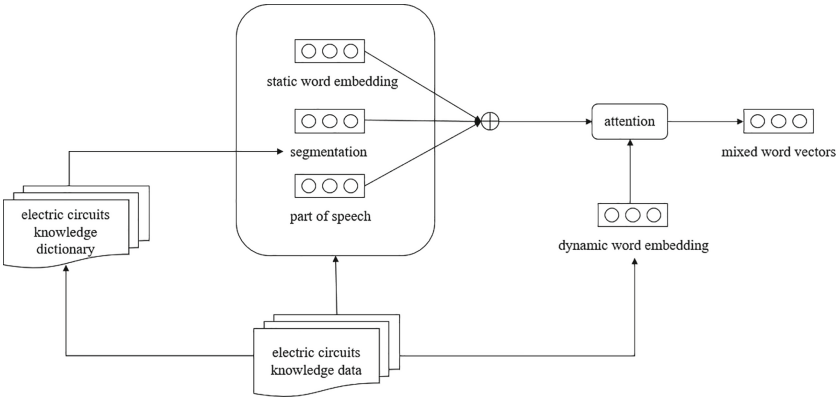

Fig. 1. Vectors fusion method.

is defined as $X = \{x_1, x_2, \ldots, x_n\}$, and the corresponding static word embedding is $X^w = \left\{x_1^w, x_2^w, \ldots, x_n^w\right\}$.

Part of Speech Feature Vectors
For named entity recognition tasks, part of speech (POS) tagging can label nouns in text, which is helpful for subsequent entity positioning and recognition. This paper uses the Jieba tool to extract POS features. Firstly, it segments the input text to obtain a set containing parts of speech and words, and then extracts the corresponding parts of speech sequence. The POS feature vector of the input sequence is $X^p = \left\{x_1^p, x_2^p, \ldots, x_n^p\right\}$.

Feature Vectors Combine with Dictionary
In order to better utilize the field text corpus of electric circuits course, a keyword dictionary is constructed based on the collected corpus. Combining the keyword dictionary, the entity type sequence of the input sequence is obtained through the bidirectional matching algorithm [5], and then the feature vector obtained by encoding it is $X^d = \left\{x_1^d, x_2^d, \ldots, x_n^d\right\}$.

2.2 Context Word Embedding

This paper uses the XLNet [6] pre-trained model to better represent the input text, which is essentially an autoregressive language model. In order to overcome the disadvantage that autoregressive language models cannot utilize both the above and the following information, XLNet proposes the Permutation Language Model (PLM). By using this strategy, it can obtain more comprehensive contextual information and thus have a further understanding of the semantics of the text. For the given input sequence, the context word embedding of input sequence is obtained through the XLNet pre-trained model as follows: $X^c = \left\{x_1^c, x_2^c, \ldots, x_n^c\right\}$.

2.3 Vectors Fusion

Static Feature Vectors Fusion
Splice the static word embedding representation, POS feature vectors, and entity type feature vectors combining dictionary to obtain a vector $x_i^f = x_i^w \oplus x_i^p \oplus x_i^d$.

Attention Model Fusion Vectors
By means of training, the attention model facilitates the identification of crucial segments within input vectors, simplifying the input process and updating the weight matrix of various input vectors. The attention mechanism assigns varied weights to distinct inputs, enabling the discovery of significant portions of input data while streamlining the input process. This paper uses feedforward attention model [7] to fuse static feature vectors x_i^f and context word embedding x_i^t.

The model determines the corresponding weights for the input vectors, and the final output vector is the weighted average of all input vectors, as demonstrated by the following Eqs. (1)–(3).

$$p\left(x_i^f\right) = \sigma\left(W_2^T \tanh\left(W_1^T x_i^f + b_1\right) + b_2\right) \tag{1}$$

$$p(x_i^t) = \sigma\left(W_2^T \tanh\left(W_1^T x_i^t + b_1\right) + b_2\right) \tag{2}$$

W_1 and W_2 are learnable weight coefficients. The vector representation of the position i character is represented as:

$$e_i = \sum_{i=1}^{n}\left(p\left(x_i^f\right) \times x_i^f + p\left(x_i^t\right) \times x_i^t\right) \tag{3}$$

2.4 Overall Model

The word embedding enhanced BiLSTM-CRF [8] model presented in this paper is composed of three primary components, as illustrated in Fig. 2. The embedding layer is utilized to convert the input text sequence into vector form. The BiLSTM layer serves to encode the input vectors and obtain richer semantic information. When decoding the output of the BiLSTM layer, this paper uses conditional random field (CRF) layer, which solve the conditional probability sequence of each location tag, and obtain the final text annotation output sequence.

For each position t, the BiLSTM layer calculates forward $\vec{h}_t$ and backward $\overleftarrow{h}_t$ representations of sequence contexts. The final representation is obtained by concatenating them: $h_t = \vec{h}_t \oplus \overleftarrow{h}_t$.

The output label sequence is modeled by CRF, which is represented as the product of conditional probabilities. During prediction, decoding algorithms such as Viterbi algorithm can be used to obtain the output sequence Y. The calculation equation for the output label sequence $Y = \{y_1, y_2, \ldots, y_n\}$ is as follows:

$$P(y_i|X, y_1, y_2, \ldots, y_n) = P(y_i|X, y_{i-1}, y_{i+1}) \tag{4}$$

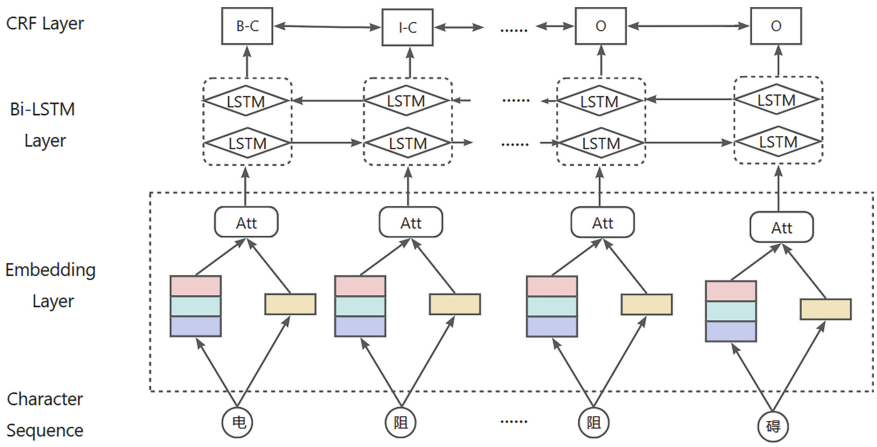

Fig. 2. Overall structure of the model.

$$P(Y|X) = \frac{1}{Z(x)} \exp\left\{ \sum_{i,k} \gamma_k f_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l h_l(y_i, x, i) \right\} \tag{5}$$

$$Z(X) = \sum_{y} \exp\left\{ \sum_{i,k} \gamma_k f_k(y_{i-1}, y_i, x, i) \right\} \tag{6}$$

f_k and h_l represent the feature function, γ_k and μ_l are the corresponding weights of the feature function. When the input sequence X is given, $P(Y|X)$ represents the conditional probability distribution of the output sequence Y, and Eq. (5) normalizes it through parameter $Z(X)$. Based on the conditional probability distribution $P(Y|X)$, using the maximum likelihood estimation, the label sequence $\hat{Y}$ can be obtained, which is shown in Eq. (7):

$$\hat{Y} = argmaxP(Y|X) \tag{7}$$

3 Experiment and Analysis

3.1 Dataset

Before conducting the experiment, it is necessary to construct an electric circuits course entity recognition dataset. This paper uses the open source data annotation tool doccano [9], which adopts two methods: manual annotation and semi-automatic annotation based on the entity keyword dictionary. The final dataset has a total of 565 entities and a total of 3265 pieces of data. The BIO (B-begin, I-inside, O-outside) labeling method is used. There are five types of entities, namely, element, variable, law, method, and term.

3.2 Ablation Experiments

Three metrics, namely precision (P), recall (R), and F1 score, are employed to evaluate the performance of named entity recognition models. The equations for computing these metrics are as follows:$P = \frac{TP}{TP+FP}$, $R = \frac{TP}{TP+FN}$ and $F_1 = 2\frac{PR}{P+R}$.

In the course of the experiment, the dataset is partitioned into ten segments, with a training set, validation set, and test set in a 6:2:2 ratio. The specific comparison modules are displayed in the first column of Table 1.The basic model is BiLSTM-CRF, "+feature" indicating that only feature vectors are used, "+cemb" indicating that only word vectors containing context information are used, and "+cemb + feature" indicating that all vectors are fused. The results of the ablation experiments, as presented in Table 1, indicate that the introduction of feature vectors enhances the F1 score of the basic model by 1.83%, while the introduction of all vectors improves the F1 score by 3.18%. In particular, the word vector with context information exhibits the most significant improvement, primarily due to the dynamic pre-trained model's capacity to augment the representation of input text. In addition, the use of the fused feature vectors in the paper also has a good promotion effect on named entity tasks, proving that POS and dictionary features can help enhance the static word embedding vector representation.

Table 1. Ablation experiment result

Models	P(%)	R(%)	F1(%)
BiLSTM-CRF	80.64	82.07	81.35
+feature	84.03	82.35	83.18
+cemb	83.86	82.89	83.37
+cemb+feature	85.63	83.46	84.53

3.3 Comparative Experiments

In this section, various models, including BiLSTM-CRF, IDCNN-CRF [10], and BERT-CRF [11], are selected for comparative experiments on the electric circuits course domain dataset constructed in this study. The objective is to confirm the efficacy of the model proposed in this paper.

IDCNN-CRF is a named entity recognition model that integrates an iterated dilated CNN with conditional random fields. The iterated dilated CNN allows the model to capture more contextual information compared to a basic CNN. On the other hand, BERT-CRF is a named entity recognition model that utilizes a pre-trained language representation model in combination with conditional random fields. The pre-trained BERT model provides a more comprehensive representation of input text and a deeper understanding of its semantics.

As shown in Table 2, the comparative experiments demonstrate that the word embedding enhanced BiLSTM-CRF model achieves superior performance over other models in the target dataset. Compared to the proposed model, the basic BiLSTM-CRF model has a 4.99% lower accuracy. Additionally, the BERT-CRF model performs better than the BiLSTM-CRF model, indicating that the mainstream BERT pre-trained language model still has some effect on the target dataset. This is mainly due to the fact that the pre-trained language model has a strong representation of input text and can learn more text features. The word embedding enhancement in this paper integrates feature vectors and context word embedding, and the effect is better than BERT-CRF. This indicates that the fusion of POS features and domain dictionary information still has an effect. The integration of feature vectors and contextual word embedding in the word embedding enhancement of this paper resulted in the best F1 score, demonstrating the effectiveness of the proposed model. What's more, the fusion of POS features and domain dictionary information had a positive impact.

Table 2. Comparative results with other models

Model	P(%)	R(%)	F1(%)
BiLSTM-CRF	80.64	82.07	81.35
IDCNN-CRF	81.58	82.51	82.04
BERT-CRF	84.63	83.89	84.26
Our Model	85.63	83.46	84.53

4 Conclusions

This paper carries out named entity recognition for electric circuits course knowledge and proposes a word embedding enhanced BiLSTM-CRF model. Word embedding enhancement includes two aspects. One is static feature vectors, which are composed of static word embedding, POS feature vectors, and feature vectors combined with dictionary information in the field of electric circuits course. The other is context word embedding, which uses an attention mechanism to fuse static feature vectors and context word embedding.

In this paper, a named entity recognition dataset was constructed in the domain of electric circuits course knowledge, using the BIO annotation method and including five types of entities. Subsequently, an experiment was conducted to evaluate named entity recognition on this dataset. The results demonstrate that the word embedding enhanced model outperforms other models, with an F1 score 3.18% higher than the basic BiLSTM-CRF model, thereby verifying the effectiveness of the word embedding enhancement.

References

1. Kong J, Zhang L, Jiang M, et al. Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition[J]. Journal of Biomedical Informatics, 2021, 116: 103737.
2. Moon S, Chung S, Chi S. Bridge damage recognition from inspection reports using NER based on recurrent neural network with active learning[J]. Journal of Performance of Constructed Facilities, 2020, 34(6): 04020119.
3. Jia C, Zhang Y. Multi-cell compositional LSTM for NER domain adaptation[C]//Proceedings of the 58th annual meeting of the association for computational linguistics. 2020: 5906–5917.
4. Bojanowski P, Grave E, Joulin A, et al. Enriching word vectors with subword information[J]. Transactions of the association for computational linguistics, 2017, 5: 135-146.
5. Gai R L, Gao F, Duan L M, et al. Bidirectional maximal matching word segmentation algorithm with rules[C]// Advanced materials research. Trans Tech Publications Ltd, 2014, 926: 3368-3372.
6. Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[J]. Advances in neural information processing systems, 2019, 32.
7. Raffel C, Ellis D P W. Feed-forward networks with attention can solve some long-term memory problems[J]. arXiv preprint arXiv:1512.08756, 2015.

8. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.
9. Daudert T. A web-based collaborative annotation and consolidation tool[C]// Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020: 7053–7059.
10. Strubell E, Verga P, Belanger D, et al. Fast and accurate entity recognition with iterated dilated convolutions[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 2670–2680.
11. Kenton J D M W C, Toutanova L K. BERT: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of NAACL-HLT. 2019: 4171–4186.