Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 3, No. 3, 2023

Named Entity Recognition of Electronic Medical Records Based on Multi-Feature Fusion

Xiaoqin Tan*

School of Electrical Engineering, Southwest Minzu University, Chengdu, China

Abstract: Named entity recognition (NER) is a very basic task in natural language processing (NLP). The paper studies the problem of named entity recognition in Chinese electronic medical records, and proposes a method based on the Bert-Bi-LSTM-CRF model. In addition, the model incorporates the functionality of radical components and dictionaries to improve the recognition accuracy. The complexity of Chinese medical record entities, the ambiguity in language expression, and the lack of adequate labeled data make traditional rule-based or machine learning methods less effective. To address this problem, we adopt the Bert-Bi-LSTM-CRF model, which effectively captures contextual information and semantic relationships to improve entity recognition accuracy. Furthermore, to further enhance the model's performance, we introduce the functionality of radical components and dictionaries. Radical components are an important component of Chinese characters and can be used to assist in identifying entities and improve the model's generalization ability. We also utilize medical dictionaries to assist in entity recognition. These dictionaries contain rich medical terms and vocabulary, which can effectively help the model identify entities. The proposed method is evaluated on the public dataset CCKS2019, and the experimental results demonstrate that it outperforms traditional methods, achieving an F1 score improvement of nearly 7 percentage points, and achieves good experimental results.

Keywords: Chinese electronic medical records; Named entity recognition; Bert-Bi-LSTM-CRF model; Deep learning.

1. Introduction

Named Entity Recognition (NER) is the extraction of required entities from structured or unstructured text. In 1996, the MUC-6 conference first proposed the task of named entity recognition and specified the main task as identifying entities in the text to be processed. With the widespread application of Chinese electronic medical records, it has become increasingly important to automate the processing and mining of the rich medical information contained within them. Among them, named entity recognition is an important part of automated processing of electronic medical records, whose purpose is to automatically identify entities in text and label them as different categories, such as diseases, drugs, surgeries, examinations, and tests. However, named entity recognition in Chinese electronic medical records faces challenges such as numerous entity types, complex expressions, and ambiguity, and traditional methods have poor performance. Therefore, a named entity recognition method based on the Bert-Bi-LSTM-CRF model is proposed, which combines the functions of radical components and dictionaries to improve the accuracy and generalization ability of Chinese electronic medical record named entity recognition.

Specifically, this method first uses a pre-trained Bert model to obtain contextual information of the text, then uses a Bi-LSTM network for feature extraction and sequence labeling, and finally uses a CRF model to constrain the labeling results to obtain the final entity recognition results. The CCKS2019 Chinese electronic medical record dataset was used for experimental evaluation, and comparison and analysis were performed with traditional methods. The experimental results show that the proposed named entity recognition method achieves good performance, with significant improvements in precision, recall, and F1-score compared to other methods. The proposed method can not only effectively solve the problem of named entity recognition in Chinese electronic medical records but also achieve better results in practical applications, and has certain promotional and application value.

2. Related Work

Named entity recognition methods can be classified into three categories: rule-based methods, traditional machine learning methods, and deep learning-based methods. Among them, deep learning-based methods have become a research hotspot in recent years due to their ability to automatically capture input sentence features and achieve end-to-end named entity recognition. Some recent research work has applied deep learning-based methods to the named entity recognition task. Huang et al applied the bidirectional long short-term memory network (BiLSTM) and conditional random field network to named entity labeling. However, BiLSTM has limited encoding ability for long sequences and performs poorly in terms of computational speed. Strubell et al. used convolutional neural networks (CNN) for named entity recognition. Compared with recurrent neural networks like BiLSTM, CNN has faster computational speed. However, CNN may lose a large amount of global information despite its good local capturing ability. Previous research has improved the Transformer encoder by supplementing directional information, enhancing its encoding ability. However, shallow Transformer encoders have the disadvantage of insufficient structural ability at the encoding layer level.

Due to the unique characteristics of the Chinese language, Chinese named entity recognition is more challenging than English named entity recognition, even though it has been developed earlier. Yang et al. used a segmentation tool to segment Chinese sentences before labeling word sequences. However, word segmentation tools inevitably make errors in word segmentation, leading to errors in identifying entity boundaries. This is because, unlike in English where delimiters are used to identify word boundaries, Chinese words do not have natural boundaries, making segmentation

more difficult. Word-enhancement methods can reduce segmentation errors and increase Chinese semantic and boundary information, but Wu et al. argued that this method ignores the information in Chinese character structures and proposed a multi-dimensional data embedding method that combines Chinese character features and radical information for improvement. Studies have shown that character-level named entity recognition is more effective than word-level [9,10]. However, a clear disadvantage of character-based named entity recognition is the loss of rich information in words. Therefore, fully integrating dictionary information into character models is a major research focus in Chinese named entity recognition. The Lattice-LSTM model proposed in literature [11] improves model recognition capabilities by combining dictionary information with the model. Specifically, the model uses the gate mechanism of long short-term neural networks to automatically match each character in the sentence with the corresponding word and incorporates the word information that is most compatible with the sentence semantics into the sentence representation. Li et al [12]. proved Chinese glyph embedding by using the "Wubi" stroke coding method to represent Chinese character structural patterns and improve overall performance in Chinese named entity recognition. Xu et al [13]. Rgued that the feature information contained in Chinese character radicals can also help improve the ability to recognize named entities. This work proposes using three different levels of embedding, i.e., character, word, and radical, in the model to enrich the character representation in the sentence and validates the effectiveness of radical information. The success of these works also confirms the effectiveness of multi-level feature information in Chinese. To further improve the performance of the model, this paper introduces external information, i.e., dictionary and radical information, into the model.

3. Bert-Bi-Lstm-Crf Model

This paper proposes a multi-feature fusion model for Chinese electronic medical record named entity recognition, which consists of two parts. The first part is a BERT-Bi-LSTM-CRF model, and the second part incorporates radical components and domain dictionaries. The BERT-Bi-LSTM-CRF model is a deep learning-based text sequence labeling model that integrates BERT [14,15]. Brectional long short-term memory networks (Bi-LSTM), and conditional random fields (CRF), mainly used for named entity recognition (NER) tasks. The model structure is shown in Figure 1.

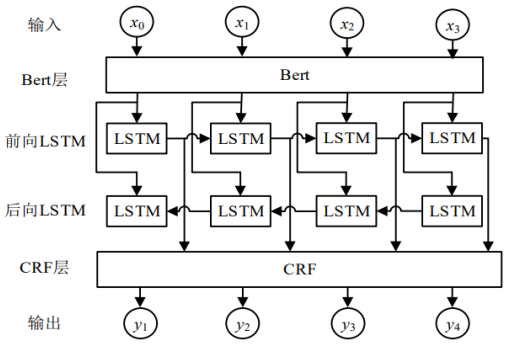

Fig.1 Named entity recognition framework based on BERT-Bi-LSTM-CRF

In the BERT-Bi-LSTM-CRF model, the input text sequence is first processed by the BERT model to obtain contextualized word embeddings for each word. These embeddings are then fed into the Bi-LSTM layer to learn feature representations that capture contextual information. Finally, the CRF layer labels the output of the Bi-LSTM layer to obtain the final output sequence. By using BERT as the input layer, the model can learn contextual representations and apply them to tasks such as text classification and named entity recognition. The Bi-LSTM acts as a feature extractor that captures context information, including the preceding and following words in a sentence. The CRF layer considers dependencies between labels, leading to more accurate label predictions.

Compared to other text sequence labeling models, the BERT-Bi-LSTM-CRF model effectively utilizes contextual information and label dependencies, resulting in improved predictive ability and accuracy. As a result, this model has achieved good performance in NER tasks and has been widely applied to various natural language processing tasks.

3.1. BERT Pre-trained Language Model

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model proposed by Google in 2018[14]. It is designed based on the Transformer model architecture and can use massive amounts of unlabeled text data for pre-training, followed by fine-tuning for various natural language processing tasks. The main contribution of the BERT model is its ability to utilize both context information and bidirectional processing of text, which makes it perform well in various natural language processing tasks [15]. The BERT model adopts two pre-training tasks, Masked Language Model (MLM) and Next Sentence Prediction (NSP). The MLM task randomly replaces some words in the input text with the [MASK] symbol, and then trains the model to predict the replaced words. The NSP task is to train the model to understand the relationship between sentences in the input text, by determining whether two sentences are consecutive. The BERT model has shown good performance in various natural language processing tasks, and in the medical field, it performs very well in electronic medical record named entity recognition. Nowadays, most studies on named entity recognition in electronic medical records use BERT model for word vector representation. The diagram of the BERT model is shown in Figure 2.

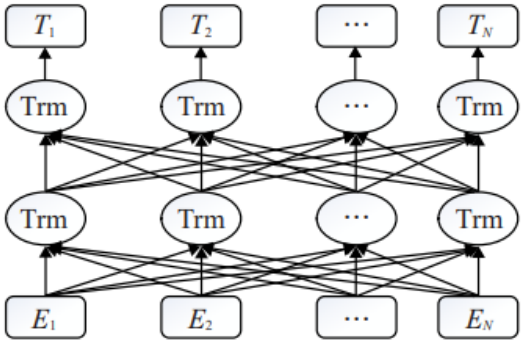

Fig. 2 Structure of BERT

3.2. Bi-LSTM Network

Long Short-Term Memory (LSTM) networks are an improved version of recurrent neural networks (RNNs) that were introduced in 1997 by Hochreiter et al. to address the issues of vanishing and exploding gradients in RNNs [16]. LSTM is a type of time-recursive neural network that

introduces "gates," including an input gate, forget gate, and output gate [17]. LSTM is an effective technique for solving the problem of long-term dependencies. As a variant of RNN, LSTM is widely used in the field of natural language processing (NLP). In the LSTM model, only unidirectional transmission exists, and it only considers the information content of the "previous context," but not the "next context." In reality, entity named recognition needs to consider all information content in all input orders. Therefore, the current approach to entity named recognition in electronic medical records typically uses bidirectional LSTM (Bi-LSTM) networks [18], as shown in Figure 3.

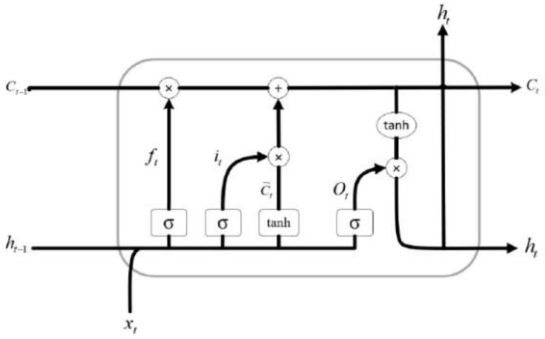

Fig. 3 LSTM structure

In Figure 3, xt is the input at time t; ft is the output for t+1 at time t; ot is the output at time t; ht is the hidden layer representing the output at time t; σ and tanh are the sigmoid function; ct is the cell state at time t. The gate unit calculation formulas in LSTM are shown below:

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

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (2)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (3)$$

$$\tilde{C}_t = tanh(W_c[h_{t-1}, x_t] + b_c) \quad (4)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (5)$$

$$h_t = o_t \odot tanh(C_t) \quad (6)$$

In the formula, Wi, Wf, and Wo are weight matrices that connect to the gate unit; bi, bf, and bo are the bias values, represents element-wise multiplication.

3.3. Radical Embedding

Unlike English, Chinese characters are ideographic, and most of the radical components still retain their original meanings, with rich intrinsic features. For example, "口" (mouth) often appears in symptom entities, such as vomiting; "月" is a simplified form of "肉" (meat), which is often related to body parts such as the heart and chest; "疒" often appears in symptom descriptions, such as pain. Therefore, measuring the similarity between entities to some extent through radical features is feasible [19]. In addition, radical embedding can enhance the semantic information of characters, such as improving the model's generalization ability for characters that appear only in the test set and not in the training set [20,21]. To fully utilize the potential intrinsic features in electronic medical record texts, a CNN-based radical extraction framework is designed. The CNN network is used to extract the implicit radical information inside characters and fine-tune during the training process. The network consists of three parts: radical embedding layer, convolutional layer, and max-pooling layer. All radicals in this paper are obtained from the corpus.

3.4. Dictionary embedding

Clinical texts contain many specialized terms, which are constructed into a term dictionary as features for the model [22]. Given a sentence $X = (x_1, x_2, \ldots, x_n)$ of length n, a feature vector d_i is constructed for each character x_i based on the dictionary D and the context. Dictionary D is constructed at the entity level, while the text sequence is based on character-level tagging, so different schemes are used to represent dictionary features. Through the construction steps of the dictionary feature vector, given a sentence X, the character embedding e_i and the feature vector di for each character x_i are obtained. In common models, the original Bi-LSTM-CRF model only takes e_i as input. Since dictionary features can provide valuable information for Chinese electronic medical record named entity recognition tasks, they are integrated into the Bert-Bi-LSTM-CRF model.

4. Experiments and Results

4.1. Experimental environment and experimental data set

The specific experimental environment is as follows: Python was chosen as the development language, the CPU model is AMD Ryzen 7 4800H, and PyCharm was chosen as the development tool.

In this study, the CCKS2019 dataset was used for experimentation. CCKS2019 (China Conference on Knowledge Graph and Semantic Computing) is an academic conference on knowledge graphs and semantic computing, as well as a Chinese-based natural language processing (NLP) competition. The named entity recognition (NER) part of the CCKS2019 dataset mainly includes annotated data of entities related to medical fields, including 6 types of entities, namely diseases and diagnoses, surgeries, drugs, anatomical sites, imaging examinations, and laboratory tests, with a total of 1379 data. The specific entity types are shown in Table 1.

Table 1. CCKS2019 dataset

CCKS2019	Disease and diagnosis	Imaging examination	Laboratory test	Operation	Drug	Anatomical site
Train	2116	222	318	765	456	1486
Test	682	91	193	140	263	447

4.2. Evaluation metrics

The precision (P), recall (R), and F1-score were used as the evaluation metrics in the experiments, and their formulas are as follows:

$$P = \frac{TP}{TP+FP} \quad (7)$$

$$R = \frac{TP}{TP+FN} \quad (8)$$

$$F1 = \frac{2 \times P \times R}{P+R} \quad (9)$$

In the formula, TP represents true positive which is the

number of positive instances that are correctly predicted as positive, FP represents false positive which is the number of negative instances that are wrongly predicted as positive, and FN represents false negative which is the number of positive instances that are wrongly predicted as negative. F1-score takes into account both precision and recall of a classification model, and is the harmonic mean of precision and recall.

4.3. Experimental result

An improved feature fusion method for electronic medical record named entity recognition based on Bert-Bi-LSTM-CRF is proposed by combining Chinese radical and domain dictionary features with deep learning models. By studying the distribution of radicals for different entities in the dataset, the basic features of Chinese characters are embedded into the model to enrich the semantic features. Three models are compared, and the Bert-Bi-LSTM-CRF model with fusion of radical embedding and domain dictionary has improved the performance of Chinese electronic medical record named entity recognition. On the CCKS2019 dataset, the F1 scores based on radical embedding and domain dictionary features have increased by 4.92% and 6.90%, respectively, compared to the Bert-Bi-LSTM-CRF model, indicating that adding radical embedding and dictionary features can effectively improve the performance of the model.

Table 2. Comparison of experimental results of different models

Model	P	R	F1
Bert-Bi-LSTM-CR	72.25%	71.69%	71.62%
Radical Embedding	78.36%	76.88%	76.54%
Dictionary embedding	81.28%	78.34%	78.52%

5. Conclusion

Based on the named entity recognition data set provided by CCKS2019 (Chinese Conference on Knowledge Graph and Semantic Computing), this paper proposes a Chinese named entity recognition method based on Bert-Bi-LSTM-CRF model, and introduces the function of side radical and dictionary to improve the recognition ability of the model. Specifically, the author adds the partial radicals and dictionary information to the input layer, and inputs them together with the Bert representation vector into the Bi-LSTM layer for feature extraction. Then, the model is trained and predicted using CRF layer, and the model is tested and evaluated in detail. The results show that the introduction of partial radicals and dictionary information can significantly improve the recognition ability of the model, and the combination of Bert-Bi-LSTM-CRF model can further improve the recognition accuracy. In conclusion, the proposed Chinese named entity recognition method based on Bert-Bi-LSTM-CRF model introduces the functions of side radicals and dictionaries at the same time, which can significantly improve the recognition ability of the model and has good practical value. Future research direction can further explore how to apply this method to more complex natural language processing tasks, such as text classification, relation extraction, etc. In addition, other techniques such as transfer learning and confrontation training can be combined to improve the performance and efficiency of the model.

References

[1] Chinchor N.MUC-6 Named Entity Task Definition (Version 2.1) [C]. Proceedings of the 6th Conference on Message Understanding, Columbia, Maryland, 1995:142-194.

[2] Xiang Xiaowen, Shi Xiaodong, Zeng Hualin A Chinese named entity recognition system that combines statistics and rules [J]. Computer Applications, 2005, 25 (10): 3.

[3] Zhang Chuanyan, Hong Xiaoguang, Peng Chaohui, et al Mesh entity activity extraction based on support vector machine and extended conditional random field [J]. Journal of Software Science, 2012, 23 (10): 16.

[4] Huang Z, Wei X, Kai Y. A bidirectional LSTM-CRF model for sequence labeling [J]. Computer Science, 2015.

[5] Strubelle E, Verga P, Belanger D, etc. Fast and accurate entity recognition using iterative expansion convolution [J]. two thousand and seventeen.

[6] Yan H, Deng B, Li X, et al. TENER: Adaptive transformer encoder for named entity recognition [J]. two thousand and nineteen.

[7] Yang J, Teng Z, Zhang M, et al. Combining discrete and neural features for sequence labeling [J]. two thousand and seventeen.

[8] Wu S, Song X, Feng Z. MECT: A Cross Transform Based on Multivariate Data Embedding for Chinese Named Entity Recognition [J]. two thousand and twenty-one.

[9] Liu Z, Zhu C, Zhao T. Chinese Named Entity Recognition Based on Sequence Marking Method: Based on Characters or Words? [J] . Springer Press, 2010.

[10] Yang X, Mao K. Using comprehensive knowledge to learn multi prototype word embedding from single prototype word embedding [J]. Expert Systems and Applications, 2016, 56 (September): 291-299.

[11] Zhang Y, Yang J. Net enrollment rate in China using Lattice LSTM [J]. 2018. Chung J, Gulcehre C, Cho K H, et al. Empirical Evaluation of Gated Recurrent Neural Networks for Sequence Modeling [J]//arXiv Preprint, 2014: arXiv: 1412.3555.

[12] Li Jie, Meng Ke. MFE-NER: Multi feature fusion embedding for Chinese named entity recognition [J]. two thousand and twenty-one.

[13] Devlin J, Chang M W, Lee K, et al. BERT: Pre training of deep bidirectional converters for language comprehension [J]. two thousand and eighteen.

[14] Yu Tongrui, Jin Ran, Han Xiaozhen, Li Jiahui, Yu Ting. Review of research on natural language processing pre training models [J]. Computer Engineering and Applications, 2020,56 (23): 12-22.

[15] Hochreiter S, Schmidhuber J, Long term and short-term memory [J]. Neurocomputing, 1997, 9 (8): 1735-1780.

[16] Wu Zongyou, Bai Kunlong, Yang Linrui, et al A review of research on text mining in electronic medical records [J]. Computer Research and Development, 2021, 58 (3): 15.

[17] Zhao R, Wang D, Yan R, et al. Machine Health Monitoring Based on Local Feature Gated Recursive Unit Networks [J]. IEEE Industrial Electronic Trading, 2018.

[18] Wu S, Song X, Feng Z. MECT: A Cross Transform Based on Multivariate Data Embedding for Chinese Named Entity Recognition [J]. two thousand and twenty-one.

[19] Xu C, Wang F, Han J, et al. Using Multiple Embedding Technology for Chinese Named Entity Recognition: 10.1145/33577384.3358117 [P]. 2019.

[20] Zhang Yunqiu, Wang Yang, Li Bocheng. Chinese electronic medical record named entity recognition based on Roberta WWM dynamic fusion model [J]. Data Analysis and Knowledge Discovery, 222,6 (Z1): 242-250.

[21] Zhang Y, Yang J. Net enrollment rate in China using Lattice LSTM [J]. two thousand and eighteen.

[22] Peng M, Ma R, Zhang Q, et al. Simplify the use of vocabulary in Chinese NER [J]. two thousand and nineteen.