Research on Chinese Medical Named Entity Recognition Based on ALBERT and IDCNN

Ziyue Zhang, Li Jin(✉), Yan Huang, and Weilin Li

School of Software, Xi'an Jiao-tong University, Xi'an, Shaanxi, China
{zyzxjtu,hy127338316}@stu.xjtu.edu.cn, jin_li@xjtu.edu.cn

Abstract. BERT (Bidirectional Encoder Representations from Transformers) as a pre-training model has been widely used in the field of natural language processing, of course, it also covers the field of Chinese medical text. In the process of actually dealing with Chinese tasks, BERT also has its own shortcomings, including the lack of Chinese word segmentation. This is because BERT is segmented based on the granularity of words. In addition, the amount of pre-training parameters of the BERT model is too large, which will also cause some problem of poor model performance caused by excessive computing power requirements, long training time, and excessive parameters. To solve the above problems, this paper proposes a Chinese medical named entity recognition model based on ALBERT and IDCNN. Experiments show that the ALBERT-IDCNN-CRF model constructed in this paper has a good performance on the Chinese electronic medical record named entity recognition task, and effectively solves the problems of polysemy and word recognition completion in Chinese electronic medical record named entity recognition. On the CCKS 2017 dataset the model effect F1 value reached 94.51%, and on the CCKS 2019 dataset, the model effect F1 value reached 88.61%.

Keywords: Electronic Medical Record · Named Entity Recognition · Deep Learning

1 Introduction

Named Entity Recognition (NER) refers to identifying entities with specific meanings from free text [8], such as person names, place names, proper nouns. It is an important branch in the field of natural language processing. Different from named entity recognition in the general field, Clinical Named Entity Recognition (CNER) is mainly for the given electronic medical record text information, to identify, extract and effectively classify medical clinical related entities. In defining categories, these entity categories typically include symptoms, drugs, diseases, and treatments. Electronic medical record [14] refers to medical personnel in the process of medical activities, using the digital information generated by the information system of medical institutions, and can achieve storage, management, transmission and reproduction of medical records. In other words, electronic medical records are descriptive textual content recorded around a patient's medical needs and service activities. For the Chinese electronic medical record

D. Qiu et al. (Eds.): ICBBEM 2022, AHIS 5, pp. 977–986, 2023.
https://doi.org/10.2991/978-94-6463-030-5_96

named entity recognition task, the difficulty lies in the lack of Chinese word breakers. Several words are usually formed into words to express meaning. At the same time, the composition of Chinese medical named entities is relatively it is complex, chaotic and unstructured. Usually, it is complicated to recognize entity combinations that are combined with words such as Chinese characters, numbers, symbols. Therefore, the Chinese medical named entity recognition task has certain theoretical research value and practical application value.

This paper proposes a Chinese medical named entity recognition method based on ALBERT-IDCNN-CRF, which performs fine-grained and lightweight named entity recognition for Chinese medical text data. The model consists of three parts. The first part is a lightweight pre-trained model ALBERT to extract text data to convert each character representation from a vector to a word embedding, and the second part extracts features from the input vector by dilating the convolutional neural network IDCNN to quickly obtain global information, and obtaining the contextual information of the text, the third part CRF labels each character in a sentence and models the transition behavior of every two different labels.

Finally, the model is applied to the CCKS 2017 dataset and the CCKS 2019 dataset, and the experimental results outperform other methods, proving the effectiveness of the model, which can be used to identify named entity recognition tasks in Chinese medical-related fields, and can be used for knowledge graphs in this field, knowledge question answering system and clinical decision expert system to provide corresponding technical support for entity recognition and extraction.

The remaining chapters of this paper are organized as follows. Section 2 presents related work on this task and summarizes the strengths and weaknesses of existing work. Section 3 introduces the medical named entity recognition method based on ALBERT and IDCNN, and Sect. 4 introduces the experimental results and result analysis. Section 5 provides the main conclusions and points out directions for future work.

2 Related Works

There are four main types of NER methods: dictionary-based, rule-based, statistical learning-based, and deep learning-based. These traditional methods include Hidden Markov Model (HMM), Support Vector Machine (SVM), and Conditional Random Field (CRF). The principle of Hidden Markov Model is to obtain the co-occurrence probability by directly modeling the transfer probability and the performance probability. Bikel [2] et al. Proposed a hidden Markov model, which can learn and recognize the name, date, time and number of classifications. For the named entity recognition task, the SVM model [16] usually has a higher accuracy rate than the hidden Markov model, but the computational cost is slightly larger and the recognition speed is slower. The CRF model counts the global probability, and considers the global distribution of the data when re-normalizing, not just normalizing locally. In traditional machine learning, CRF is regarded as the mainstream model for named entity recognition [3], and the key is to use internal and contextual feature information to label a location. Some scholars have also made some improvements to the CRF model. Scholars such as Minkov [12] optimize the model by fine-tuning the weights of features, and Culotta et al. [6] rank and

filter entity recognition by calculating the confidence score of phrases. In 2017, Gridach et al. [7]. Applied the CRF model to the task of medical named entity recognition. Yang et al. [13] applied the CRF model to the CCKS 2018 Chinese medical named entity recognition dataset in 2018. In these traditional machine learning methods, NER needs a large-scale corpus to learn and label the model, and professionals need to participate in feature engineering. However, the performance of final entity recognition is affected by the annotation quality of corpus data set, and then affects the generalization performance of the model.

Compared with the previous traditional methods, the method based on deep learning has the characteristics of not relying on rule design and feature engineering. The method based on deep learning does not rely on a lot of specific field experience, knowledge and human resources, and does not require manual extraction. Feature. A neural network-based named entity recognition method was first proposed by Collobert [5] and other scholars. The shortcoming of this method is that it fails to consider the effective information between long-distance words. To overcome this limitation, Chiu and Nichkols [4] propose a bidirectional LSTM-CNNs architecture that automatically detects word and character level features. BiLSTM solves the problem of recurrent neural network (RNN), long short-term memory network (LSTM), gated recurrent unit (GRU) These structures cannot adapt to longer sequences and cannot extract bidirectional information. Ma et al. [17] designed an end-to-end model based on BiLSTM, CNN and CRF to achieve satisfactory performance. However, CNN will lose some hidden information in the network due to its fixed size window, and cannot fully obtain global information. Will lead to huge computational complexity. For medical text named entity recognition tasks, Maryam H et al. [11], Topaz M et al. [15] also made some improvements based on the BiLSTM-CRF model to improve the performance of the model. In essence, convolutional neural networks are not suitable for named entity recognition tasks. When dealing with tasks in the text field, they are ineffective due to their disorder, while deep network structures such as recurrent neural networks and long short-term memory networks have better performance. Effect. In order to capture local information at the same time and obtain long-distance features without reducing the training speed, an expanded convolutional neural network is used to solve the above shortcomings.

Adding attention mechanism to the structure based on neural network is also the current mainstream research direction. L. Jinhyuk [1] and other scholars combined the BERT word embedding model on the basis of the traditional BiLSTM-CRF model, and the improvement of this improved a number of performance indicators. Transformer-based pre-training models represented by BERT can obtain prior knowledge from a large amount of unlabeled text to enhance semantics. BERT uses a masked language model to achieve pre-trained deep bidirectional representations, pre-trained deep bidirectional representations from unlabeled text by jointly conditioning the left and right contexts in all layers, and performs relatively well on a large number of sentence-level and token-level tasks good performance, outperforming many task-specific architectures. There are also many improvements to the BERT model. BioBERT [10], for biomedical text mining tasks. Hakala et al. applied multilingual BERT to medical named entity recognition.

The main parameters of the BERT network structure are token embedding, encoder feedforward, and multi-head self-attention, and the magnitude of these parameters is

huge. The parameters of the basic version of BERT will occupy 110 M, and the parameters BERT-large will occupy 340 M. Some improved models of BERT can reduce the number of parameters to 1.8 M, such as DualTrain [9], which will overcome BERT for GPU/TPU video memory requirement is too large. Tsai et al. utilize knowledge distillation to run a small BERT for NER on a single CPU.

There are many improvements to this, including ALBERT. ALBERT [18] is a streamlined pre-training model. Compared with BERT, it uses cross-layer parameter sharing and embedding vector parameterization to reduce parameters, and adopts sentence order prediction loss instead of next sentence prediction (NSP). The difference is that sentence order prediction applies two consecutive segments in the same document as positive examples, and the same two consecutive segments but the order is swapped for negative examples, rather than next sentence prediction combines topic prediction and consistency prediction in a single task, this will make ALBERT outperform BERT on downstream tasks.

3 Method

Based on literature research to deal with the task of Chinese medical named entity recognition, and to improve model performance on the premise of reducing model parameters to speed up training, this paper introduces lightweight pre-training models ALBERT and IDCNN to achieve the above optimization purposes. The model can enhance the semantic representation by acquiring prior semantic knowledge from a large amount of unlabeled text, and acquire word-level semantic representation during pre-training. At the same time, in order to consider the context information of a wider scale, solve the semantic connection between the long-distance contexts of text sentences and solve the problem that the CNN model loses hidden data due to the fixed window size in the process of named entity recognition, the CRF model is finally used information to improve the accuracy of named entity recognition. Therefore, considering the above, this paper proposes a method for Chinese Clinical medical record named entity recognition based on ALBERT-IDCNN-CRF, which is divided into three modules: ALBERT, IDCNN and CRF as shown in the figure. The main working characteristics of this model are as follows:

- Using the lightweight pre-training model ALBERT adopts methods such as embedded parameter decomposition and cross-layer parameter sharing to effectively reduce calculation parameters, speed up training, and reduce the impact of training data quality on Chinese named entity recognition, making it more suitable for Chinese Electronic Medical Record Named Entity Recognition Task.
- After acquiring the semantic representation, IDCNN is used to simultaneously learn global and local features, capture sequence feature information and implicit information, and use conditional random field CRF to limit the sequence relationship between labels to complete the task of named entity recognition.
- The model is verified on the real Chinese electronic medical record dataset, and the experimental results show that the method can effectively improve the recognition accuracy of Chinese electronic medical record named entities.

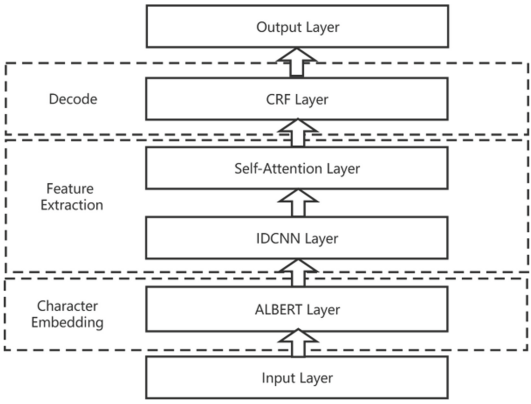

Fig. 1. The framework of ALBERT-IDCNN-CRF model

This paper proposes a Chinese medical named entity recognition model based on ALBERT and IDCNN. The model includes a total of three modules, namely ALBERT, IDCNN, and CRF, as shown in Fig. 1.

3.1 ALBERT Layer

The ALBERT pretrained model is used in this paper to obtain the vector representation of the input information. Compared with BERT, ALBERT uses Transformer encoder and GELU nonlinear activation function in its backbone network. From a modeling perspective, word embeddings learn context-independent representations of words, while hidden layers learn context-dependent representations. A large part of BERT's representational power comes from using context to provide context-dependent representational information during the learning process. ALBERT factorizes the word embedding parameters, which mainly depend on the word embedding size E, hidden layer size H, and dictionary size V, and dictionary size, so that the word embedding parameters are reduced from $O(V \times H)$ to $O(V \times E + E \times H))$. At the same time, ALBERT also uses a cross-layer parameter sharing mechanism to further improve parameter efficiency. All parameters are shared in all layers. Experiments show that parameter sharing can make model parameters more stable. In addition, ALBERT uses a loss function based on language coherence. Compared with the next sentence prediction loss used by BERT, it only focuses on the coherence between modeling sentences, which can significantly improve the performance of downstream multi-sentence encoding tasks.

3.2 Iterated Dilated CNN Layer

For the sequence labeling task, after the convolution of CNN, the last layer of neurons may obtain a small part of the information in the original input data. For named entity recognition, each word in the entire input sentence has the potential to affect the labeling of the current position. If you need to add more convolutional layers in order to cover all the input information, this will lead to more and more layers. In order to prevent

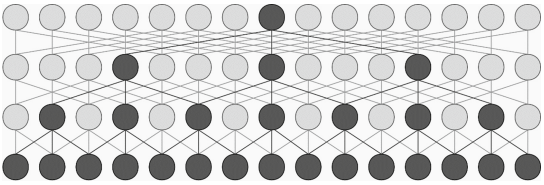

Fig. 2. A dilated CNN block with maximum dilation width 4 and filter width 3. Neurons contributing to a single highlighted neuron in the last layer are also highlighted

overfitting, it will also bring more hyperparameters, and the entire model will become huge and difficult to train. For which a dilated convolution is required. In IDCNN, the purpose of sharing parameters is achieved by reusing the modules of the self-expanding stable to prevent overfitting. At the same time, the Viterbi algorithm can also be used to predict the sequence labels for the input of IDCNN (Fig. 2).

3.3 CRF Layer

In named entity recognition tasks, adjacent tags have a certain order relationship, such as $I-BODY$ tags must appear $B-BODY$ after. The feature extraction layer outputs the hidden state context feature vector h, which is expressed as $h = (h_1, h_2, \cdots, h_t)$, this vector only considers the context information in the electronic medical record, but does not consider the dependencies between labels. Therefore, this paper adds a CRF layer for label decoding to label the global optimal sequence, and converts the hidden state sequence $h = (h_1, h_2, \cdots, h_t)$ into the optimal label sequence $y = (y_1, y_2, \cdots, y_t)$, and defines its probability as (1)

$$p(y|s) = \frac{exp\left(\sum_i \left(O_{i,y_i} + T_{y_{i-1},y_i}\right)\right)}{\sum_y exp\left(\sum_i \left(O_{i,y_i} + T_{y_{i-1},y_i}\right)\right)} \tag{1}$$

T denotes the transition matrix, denotes T_{y_{i-1},y_i} the transition score O_{i,y_i} from label y_{i-1} to label, y_i denotes the score at which a character x_i is predicted to be a label y_i, y denotes all possible sequence of labels. When decoding, the Viterbi algorithm is used to find the tag sequence with the highest score y^*, which is calculated as (2)

$$y^* = argmax \sum_i \left(O_{i,y_i} + T_{y_{i-1},y_i}\right) \tag{2}$$

4 Experimental Verification Analysis

4.1 Dataset Acquisition and Annotation

The experiment uses the Chinese medical data set CCKS 2017 and the Chinese medical data set CCKS 2019 for training and evaluation. CCKS 2017 is real and desensitized clinical Chinese electronic medical record data, with a total of 5 entity types annotated, including disease diagnosis, symptoms and signs, examination, treatment and body parts,

Table 1. CCKS 2017 NER Dataset Entity Distribution

All	Body	Disease	Symptom	Test	Treatment
5363	10719	722	7831	9546	1048

Table 2. CCKS 2019 NER Dataset Entity Distribution

All	Disease	Operation	Drug	Anatomy	Image inspection	Laboratory inspection
5363	2116	765	456	486	222	318

a total of 300 electronic medical records, divided into 1200. The number of each entity type in the training set is shown in Table 1.

The CCKS 2019 dataset annotates a total of 6 entity types, namely disease diagnosis, surgery, medicine, anatomy, imaging examination and laboratory test, and contains 1379 pieces of data, including 1000 training data and 379 testing data. The training data is divided into two parts according to 7:3, which are used as training set and validation set respectively. The number of each entity type in the training set is shown in Table 2.

4.2 Experimental Environment Settings

In this experiment, using Python 3.7 and Tensorflow 1.14.0 training framework, applying BERT-base and ALBERT-base, the basic architecture is a bidirectional Transformer stacked with 12 layers, and the hidden layer dimension is 768. Adam is used as the optimizer in the training process, and the dropout strategy is adopted to avoid overfitting. The CPU is Intel i7-11800 H, and the GPU is NVIDIA GeForce GTX 1080 Ti, during the training process, fine-tune the pre-trained model, the learning rate is 5e−5, the learning rate of the downlink structure is 1e−4, the learning rate is updated by a fixed step decay method, and the batch size is 32.

4.3 Evaluation Metrics

In Chinese named entity recognition tasks, models are usually evaluated using precision P recall R and F_1 value. Precision is a relatively intuitive evaluation that represents the percentage of correct results in all results. The higher the evaluation index, the better the model effect.

$$Precision = \frac{T_p}{T_p + F_p} \times 100\% \tag{3}$$

$$Recall = \frac{T_p}{T_p + F_n} \times 100\% \tag{4}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100\% \tag{5}$$

T_p represents the number of correctly recognized entities, F_p represents the number of misrecognized entities from other texts, and represents the number of entities F_n that were not recognized as entities.

4.4 Results and Discussion

The effectiveness and robustness of the ALBERT-IDCNN-CRF model on the Chinese medical dataset CCKS 2017, a comparative analysis with other named entity recognition methods is carried out, and the experimental results are shown in Table 3 and 4.

In order to demonstrate the effectiveness of ALBERT, we conducted a comparative experiment between BERT-BiLSTM-CRF and ALBERT-BiLSTM-CRF, and the results showed that on the CCKS 2017 dataset, the values of P, R and F_1 were improved by 0.51%, 0.06% and 0.61%. On the CCKS 2019 dataset, the values of P, R and F_1 are increased by 0.94%, 1.48% and 2.22% respectively, indicating that the ALBERT model makes the model more capable of extracting speech information and can form the effect of downstream tasks promote.

Also from the results, we can find that each layer of the composite model focuses on different features. BiLSTM-an additional CRF layer for CRF whose main purpose is to learn the constraints of sentences and reduce mis-predicted sequences. The introduction of BERT improves the accuracy, and sometimes the IDCNN-CRF model performs worse than the BiLSTM-CRF without the introduction of BERT.

The experimental results show that ALBERT can effectively improve the effect of pre-training. In feature extraction, IDCNN can obtain more local information than BiLSTM, which makes up for the defects of CNN. On the whole, IDCNN not only reduces the training parameters, but also expands the field of perception, and improves

Table 3. Comparison of NER Results on CCKS 2017 dataset (%)

Model	Precision	Recall	F1-score
BiLSTM-CRF	86.62	87.82	87.19
IDCNN-CRF	84.62	88.47	89.39
BERT-BiLSTM-CRF	93.11	94.27	93.66
BERT-IDCNN-CRF	93.24	93.68	93.90
ALBERT-BiLSTM-CRF	93.64	94.07	94.38
ALBERT-IDCNN-CRF	93.75	93.74	94.51

Table 4. Comparison of NER Results on CCKS 2019 dataset (%)

Model	Precision	Recall	F1-score
BiLSTM-CRF	82.68	82.37	82.41
IDCNN-CRF	81.79	82.36	82.85
BERT-BiLSTM-CRF	87.36	86.81	86.57
BERT-IDCNN-CRF	87.82	87.47	86.39
ALBERT-BiLSTM-CRF	88.57	89.24	88.67
ALBERT-IDCNN-CRF	88.76	88.95	88.61

the effect of the final NER task. In summary, the named entity recognition model based on ALBERT-IDCNN-CRF is superior to the above models, and the effectiveness of the model is verified by experiments.

5 Conclusion

Aiming at the task of Chinese medical named entity recognition, this paper proposes an entity recognition model based on ALBERT-IDCNN-CRF to recognize the entity information in the medical field. The model uses ALBERT to extract the raw vector representation of the input information and input it into DCNN-CRF for training, overcoming the shortcomings of BERT and BiLSTM methods. Experiments are carried out on the CCKS 2017 and CCKS 2019 datasets, and the experimental results both demonstrate the effectiveness of the model in this paper.

How to further improve the effect of the model in the Chinese medical named entity recognition task will be the focus of the next research. At the same time, related research and work will be carried out on the downstream tasks of named entity recognition, such as the construction of knowledge graphs in the medical field and the implementation of automated question answering systems in the medical field.

References

1. Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In COLING, pages 1638–1649, 2018.
2. Bikel Dm, Schwartz R, Weischedel R M. An algorithm that learns what's in a name [J]. Machine learning, 1999, 34(1):211–231.
3. Chen Shudong, Ouyang Xiaoye. Overview of Named Entity Recognition Technology [J]. Radio Communications Technology, 2020, 46(3): 251–260.
4. Chiu Jpc, Nicholse. Named Entity Recognition with Bidirectional LSTM-CNNs[J] Transactions of the Association for Computational Linguistics, 2016, 4: 357–370.
5. Collobert R, Weston J, Bottou L, et al. Natural Language Processing (almost) from Scratch[J]. Journal of Machine Learning Research, 2011, 12 (Aug): 2493–2537.
6. Culotta A, Mccallum A. Confidence Estimation for Information Extraction [C] Proceedings of HLT-NAACL 2004: Short Papers, 2004: 109–112
7. Gridach M. Character-level neural network for biomedical named entity recognition[J]. Journal of Biomedical Informatics, 2017, 70: 85–91.
8. Grishman R, Sundheim B. Message understanding conference 6: a brief history[C] Proceedings of the 16th Conference on Computational Linguistics-Volume 1, 1996.
9. Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazha gan, Xin Li, and Amelia Archer. Small and practical BERT models for sequence labeling. In EMNLP-IJCNLP, pages 3632–3636, 2019.
10. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 2019. Kai Hakala and Sampo Pyysal. Biomedical named entity recognition with multilingual BERT. In BioNLP Open Shared Tasks@EMNLP, pages 56–61, 2019.
11. Maryam H, Leon W, Mariana N, David LW, Ulf L. Deep learning with word embeddings improves biomedical named entity recognition[J]. Bioinformatics (Oxford, England), 2017, 33(14).

12. Minkov E, Wang R C, Tomasic A, et al. NER systems that Suit User's Preferences: Adjusting the Recall-precision Trade-off for Entity Extraction [C] Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. 2006: 93–96.
13. Rules Based Approach CCKS 2018 Yang X, Huang W. A conditional random fields approach to clinical name entity recognition[C] Proceedings of the Evaluation Tasks at the China Conference on Knowledge Graph and Semantic Computing (CCKS 2018). Tianjin, China: CCKS, 2018: 1–6.
14. Shen L, Li Q, Wang W, Zhu LJ, Zhao Q Z. Treatment patterns and direct medical costs of metastatic colorectal cancer patients: a retrospective study of electronic medical records from urban China [J]. Journal of Medical Economics, 2020, 23(5).
15. Topaz M, Murga L, Katherine M, Margaret V, McDonald, BarBachar O, Goldberg Y, Kathryn H, Bowles. Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches[J]. Journal of Biomedical Informatics, 2019, 90.
16. Vapnik VN and Lerner A.Y, 1963. Recognition of patterns with help of generalized portraits. Avtomat. I Telemekh, 24(6), pp.774–780.
17. X. Ma, E. Hovy, End-to-end sequence labeling via Bidirectional LSTM-CNNs-CRF, in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 1064–1074.
18. Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In International Conference on Learning Representations, 2020.