Research on Domain Specific Chinese Named Entity Recognition Based on RTBC Algorithm

Xiaohua Ke(✉), Xiaobo Wu, Zexian Ou, and Binglong Li

School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou 510006, China
carrieke@gdufs.edu.cn

Abstract. In the task of Chinese named entity recognition, how to enhance the recognition ability of the model for the boundary between characters and words and how to process the common polysemy of words is a hot issue that many scholars are working on it. In this paper, we propose a Chinese entity recognition model incorporating language model, RoBERTa-WWM-TextCNN-BiGRU-CRF model, which uses RoBERTa model pretrained on large-scale corpus to dynamically generate word vector sequence according to its input context, and then uses BiGRU and TextCNN combined model to further extract sentence features and capture word boundary information, and finally input the sequences of feature vectors into the final prediction results are achieved by inputting some constraints into the CRF model. Experiments were performed on the resume dataset and a customized dataset foreign affair, and the precision, recall, and F1 values were all improved compared to current mainstream named entity recognition models.

Keywords: Named Entity Recognition · RoBERTa · TextCNN-BiGRU · Foreign Affair

1 Introduction

Chinese Named Entity Recognition (CNER) is a foundational task in the field of natural language processing. Its purpose is to identify named entities from Chinese text and classify them into predefined categories such as the names of people, places, organizations, currencies, and certain proper nouns. The recognition of named entities is both a necessary technical foundation for the construction of knowledge graphs and plays an important role in natural language processing applications such as information retrieval [1, 2], question-answering systems [3], and machine translation [4], where accurate NER makes a positive contribution to these Information extraction tasks. Some route entities pose two problems: on the one hand, the entity words are ambiguous, e.g. “Guangdong University of Foreign Studies” is synonymous with “Guang wai”. On the other hand, there are difficulties in identifying entity word boundaries, as in the case of “Guangdong University of Foreign Studies”, which may be both a place and an institutional entity; Therefore, the identification of named entities in this domain is much more complex

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 574, pp. 118–129, 2024.
https://doi.org/10.1007/978-3-031-65123-6_9

than the identification of named entities in general, and no research has yet been conducted specifically for named entity identification in foreign affair areas. To address the above issues, this study focuses on NER in foreign-related domains, constructs a dataset for recognizing named entities in foreign-related domains, transposes text features learned from a large corpus of general domains to foreign affair areas, proposes a NER method combined with a Chinese pretraining model, compares the existing methods. The results are compared with existing methods, and better results are obtained on both the constructed corpus and the public dataset.

2 Related Work

There are three common methods for Chinese NER, as follows: dictionary-based and rule-based pattern-matching methods, statistical machine learning techniques, and deep learning methods based on neural networks [5]. Methods for recognizing named entities based on rules and lexicons include: building rule templates or lexicons using text analysis methods and then using these to perform string matching [6]. This method is less portable and can only be used in a single domain. Inspired by statistical machine learning, the NER task is converted to a sequence labeling problem using the concepts of probability and statistics, where the model makes probability predictions about possible sequences. Examples include Hidden Markov Model (HMM), Conditional Random Field (CRF), where CRFs have been used in combination with other models to perform well on NER tasks and are used extensively in the medical, agricultural, and military domains. In the last few years, with the rapid expansion of the deep learning field, deep neural network methods are able to independently extract text features and perform better with stronger generalization a new language model, Bidirectional Encoder Representations from Transformers (BERT) [7], obtained SOTA results in a variety of common Natural Language Processing tasks. Language models pretrained on Incorporating BERT in BiLSTM-CRF models have become a common tool in the field of NER nowadays.

Due to the lack of partial information on word granularity in Chinese texts compared to English named entity identification tasks, and the existence of multiple senses in Chinese make it difficult to perform named entity identification tasks in Chinese. To address the above issues, we propose the NER algorithm RoBERTa-TextCNN-BiGRU-CRF using RoBERTa to pretrain the language model in order to get dynamic word vector information and to be able to deal with multiple senses of a word. The Bidirectional Gating Recurrent Unit (BiGRU) model has a simpler structure and is faster to compute, and the feature information [8] is obtained from the vector representation layer and the sequence modeling layer, and semantic information is more adequately utilized, which can effectively address the problem of boundary ambiguity in CNER. In contrast, crawled data from the website of Confucius Institute of Guangdong University of Foreign Studies were preprocessed and annotated in order to build a corresponding dataset of foreign affair related corpora containing 450k words of text. The paper was finally validated with both the foreign corpus dataset and the resume dataset. This study shows that the current model provides better solution than the benchmark model mentioned earlier.

3 Methods

3.1 Overall Framework of the Model

In this paper, we propose a RoBERTa-TextCNN-BiGRU-CRF model to address the aforementioned issues. There are three parts in this model: RoBERTa as the vector representation layer, which addresses the problem of multiple senses by obtaining dynamic word vectors; TextCNN-BiGRU as a sequence modeling layer, which uses Text Convolutional Neural Networks (TextCNN) for extracting textual features from multiple viewpoints and BiGRU for extracting contextual information to address the ambiguous boundary problem; and the last layer is a labelling layer composed of CRF The final layer is the label layer composed of CRF, which increases the accuracy of the results by adding some constraints, The overall structure diagram is shown in Fig. 1.

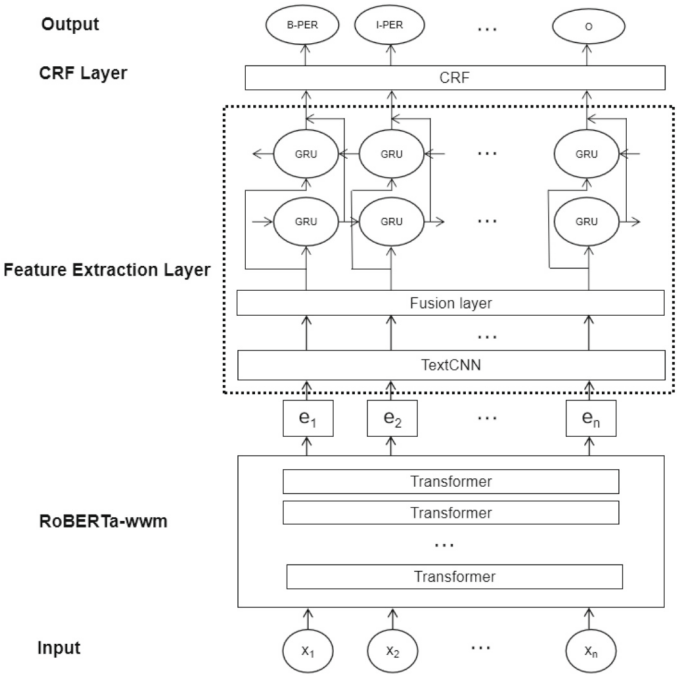

Fig. 1. RoBERTa-WWM-TextCNN-BiGRU-CRF model

3.2 Pretrained Language Model

This experiment uses the Sino-RoBERTa-Whole Word Masking (WWM) pretrained language mode [9] introduced by the Xunfei Joint Lab at HIT. The BERT language model was built on a multilayer bidirectional fine-tuned Transformer encoder that exploits the mutual positional relationship between words for deeper retrieval of text information.

To improve the model, we use the context fusion language model, which improves upon the commonly used bidirectional language model, involving the masking of randomly selected characters (called tokens) from the text rather than the simple concatenation of left-to-right or right-to-left phrase encodings, and the loss function counts only the tokens masked during training, allowing the model to learn more effectively and to produce more accurate predictions. Figure 2 shows the structure of RoBERT-WWM model.

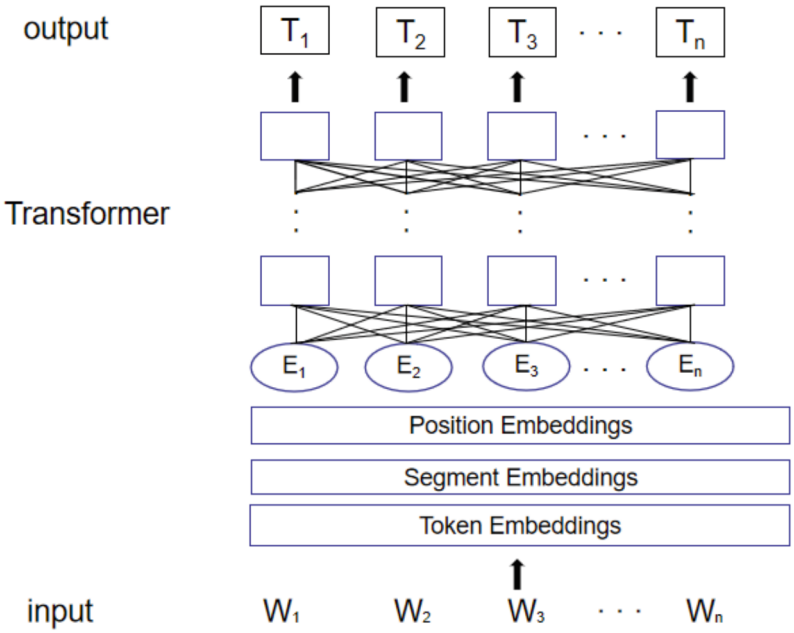

Fig. 2. RoBERTa-WWM model

(1) **Whole Word Coverage**: RoBERTa-WWM adopts a whole-word masking approach in comparison to the BERT model which masks some characters in some proportion, the RoBERTa-WWM model directly masks a certain pattern word, which is more in line with the semantic layout of the Chinese language and may better capture semantic information. In comparison to the BERT model, which masks some characters to a certain extent, the RoBERTa-WWM model directly masks a certain model sentence, which is more consistent with the Chinese semantic layout and can capture semantic information better.

(2) **Dynamic Mask:** The training of the pre-trained BERT language model consists of a random selection of 15% of words within the corpus to be masked using Mask tokens, which do not vary. The algorithm uses a dynamic masking strategy, where a new set of tokens is selected for masking in each iteration cycle [10].

(3) **Simple Removal of Next Sentence Prediction and A Richer Training Process:** As distinguishing between positive and negative samples is not challenging for neural

networks during pre-training, the simple removal of next sentence prediction task was removed while improving model efficiency. On the other hand, the RoBERTa-WWM training process uses more data and more adequate training epochs [11].

3.3 Feature Extraction Layer Model

Convolutional Neural Networks (CNN) are derived from the development of artificial deep neural networks, originally designed for image recognition and specifically for the analysis of pixel data. CNN convolutional layers are different from normal neural networks. The network traverses each vector and matrix dimension to classify the images, allowing the CNN to be more resilient to raster data. TextCNN is an improved CNN-based model for text features, which is preferable for acquiring local information and capturing deep information [12]. The structure of the model is shown in Fig. 3, which consists of the following three major parts: the convolutional layer, the pooling layer, and the fully connected layer. Each filters can obtain a feature, so the convolutional layer can realize the extraction of local features of the text, that is, to obtain different feature expressions. The calculation formula is as follows:

$$a_i = f(W \times T_{i:i+n-1} + b) \tag{1}$$

$$\mathbf{A} = \left[a_1, a_2, \ldots, a_{n-h+1}\right] \tag{2}$$

where w is the weight, b is the bias, n is the size of the convolution kernel, and f is a nonlinear function that maintains the nonlinearity of the model, $T_{i:i+n-1}$ represents word vectors at different locations in the text. The resulting ai is the I^{th} feature extracted from a convolution kernel. The convolutional layer extracts key input feature information in a neural network, and it consists of multiple convolutional units, each with parameters optimized using a back-propagation algorithm [13]. In practice, different convolutional kernels are selected to obtain better results according to the situation. A maxpooling approach is used in the pooling layer, which serves to reduce the dimensionality of the output of the convolutional layer in an attempt to avoid overfitting situations, and the final output is obtained with the same dimensionality of the features. In the fully-connected layer, the output vectors of the pooling layer are stitched together and the acquired local features are merged into global features to obtain a fixed length feature vector [14].

Gate Recurrent Unit [15] is a type of Recurrent Neural Network (RNN) [16], which can solve the problems of non-long-term memory and gradient in backpropagation in RNN. Compared to a one-way GRU model, in this paper we use a bidirectional GRU model, which combines forward and reverse implicit layers, incorporating context information about the text. The structure of the system is shown in Fig. 4.

x_t is used as input at time t, h_{t-1} as the hidden state at the previous time t, h_t as the hidden state at that time t, r_t is the resetting gate, z_t is the updating gate and σ is the sigmoid activation function. The update parameters appear in Eqs. (3)–(6) like this:

$$r_t = \sigma(w_r x_t + u_r h_{t-1}) \tag{3}$$

$$z_t = \sigma(w_z x_t + u_z h_{t-1}) \tag{4}$$

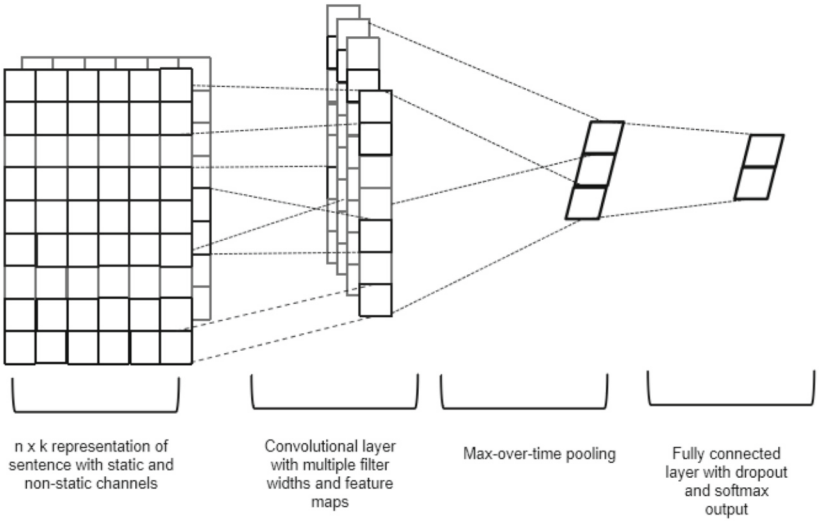

Fig. 3. TextCNN model

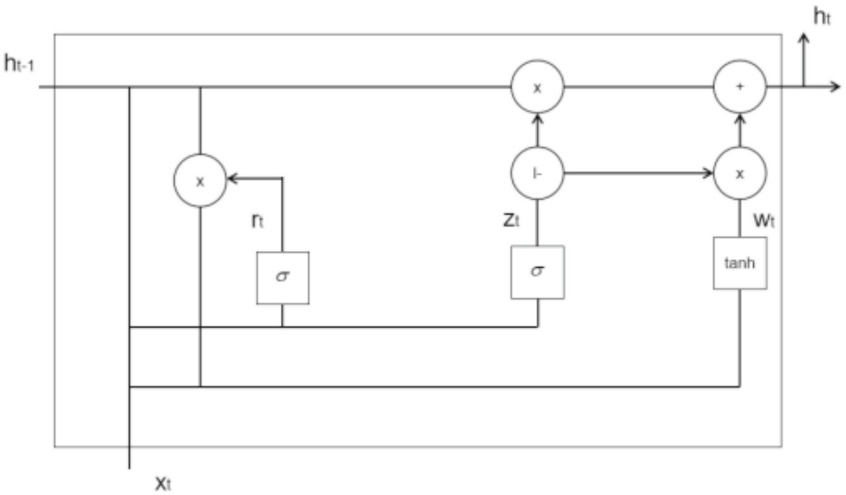

Fig. 4. BiGRU model

$$w_t = \tanh(wx_t + r_t \odot uh_{t-1}) \tag{5}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot w_t \tag{6}$$

3.4 CRF Layer

Conditional Random Field can take full account of dependencies among individual labels, and may reduce the probability of outputting incorrect labels by adding certain constraints. As a result, a CRF layer is added to the BiGRU layer, and decoding using the CRF can achieve the globally optimal label sequence.

In CRF, for an input sequence $X = \{x_1, x_2, \ldots, x_n\}$, and the corresponding predicted labels $Y = \{y_1, y_2, \ldots, y_n\}$, n is the length of the sequence, the evaluation score of the CRF model is shown in Eq. (7). In this section, we present the results of the proposed method. The state transfer matrix is given by where, A denotes the state transfer matrix, $W_{yi,\, yi+1}$ is the probability score of the transfer of label y_i to label y_{i+1}, $P_{i,\, yi}$ is the probability score of the i^{th} character being labeled as label y_i, and finally the scores for each tag sequence are normalized to obtain the likelihood, where the sequence of tags with the highest likelihood is the final annotated sentence sequence.

$$\text{Score}(X, y) = \sum_{i=1}^{n} W_{y_i, y_i+1} + \sum_{i=1}^{n} P_{i, y_i} \tag{7}$$

Equation (8) shows the normalization process. During decoding, the Viterbi algorithm is used, and the method finds the sequence of labels with the highest probability y*, as can be seen from Eq. (9).

$$p(y|s) = \frac{e^{\text{score}(x,y)}}{\sum_{\tilde{y} \in y_{y_i}} \text{score}(x, \tilde{y})} \tag{8}$$

$$y^* = \arg\max_{\tilde{y} \in Y_X} \text{Score}(X, \tilde{y}) \tag{9}$$

4 Experiments

4.1 Datasets

In this experiment, we use two datasets, the resume dataset and the self-constructed foreign affair dataset. The NER dataset resume was generated through a process of filtering and manual annotation from the summary data of the top managers of listed companies in Sina Finance. Because there is no publicly available dataset for the identification of named entities in foreign-related fields, 450K texts were crawled from the website of the Confucius Institute of Guangdong University of Foreign Studies according to the keywords "foreign-related" and "foreign-exchange" in order to test the accuracy of the model. Based on the analysis, this textual information consisted of the following five entity types: title (POS), role (PER), location (LOC), organization (ORG), and event (RES). We manually annotate the five entity types using the BIO annotation strategy, with B (Begin) corresponding to the entity's start position in the character string, I (Intermediate) corresponding to the midpoint or end position of the entity in the sequence of characters, O (Other) corresponding to the non-entity character in the character sequence,

location labels including (B-LOC, I-LOC), the labels of people include (B-PER, I-PER), the labels of event keywords include (B-RES, I-RES), the labels of related organization keywords include (B-POS, I-POS), and the labels of non-entity characters are (O), and finally 111397 entities are obtained from the experimental data. Lastly, the data was split into a training datase, a test datase, and a validation datase based on the 8:1:1 ratio, and experiments were performed in four different models. The Foreign affair dataset is as follows (Table 1):

Table 1. Data entity label statistics

Data	POS	PER	RES	LOC	ORG	Total
Training dataset	18235	25302	28480	9445	8845	90307
Test dataset	2401	4503	1068	1093	3419	12484
Validation dataset	1833	3658	790	720	1605	8606

4.2 Experimental Environment and Parameters

The Pytorch deep learning framework was used in the construction of the experimental model, and the experimental environment was set up as shown in Table 2. Key model parameters included the RoBERTa-WWM model, the parameters of the BiGRU and TextCNN models, and each model was tuned to achieve the optimum parameters. The RoBERTa-WWM model uses a twelve-layer Transformer; the TextCNN model takes as input the dynamic word vector obtained from the pre-training of the RoBERTa model, the number of layers in the network is 1, the size of the hidden layer is 128, the activation function is chosen to be Relu, the pooling strategy is max_pooling, and the dropout ratio is set to 0.1. The MLP model was added to the TextCNN model in order to adjust the dimensionality of the output to correspond to the input of the BiGRU model, using a hidden layer BiGRU unit of 128, dropout of 0.1, batch size of 64, maximum sentence length of 64, initial learning rate of 5e-5, and an Adam optimizer, we were able to achieve a performance improvement by using the following methods.

Table 2. Experimental environment configuration

Operating system	windows
CPU	24 vCPU AMD EPYC 7642 48-Core Processor
GPU	RTX 3090 (24 GB)
Python	3.9.12

4.3 Evaluation Indicators

The metrics used to evaluate the effectiveness of this set of models are precision (P), recall (R), and f1-score (F1), which are most commonly used in the field of NER, The calculation equations are shown in (10)–(11):

$$P = \frac{T_p}{T_p + F_p} \times 100\% \tag{10}$$

$$R = \frac{T_p}{T_p + F_n} \times 100\% \tag{11}$$

$$F1 = \frac{2P \cdot R}{P + R} \times 100\% \tag{12}$$

4.4 Analysis of Experimental Results

In order to verify the effectiveness and robustness of the proposed model in recognition, experiments were conducted on the mainstream models BiLSTM-CRF, BERT-CRF and BERT-BiLSTM-CRF on the foreign and resume datasets, and the experimental results are shown in Tables 3 and Table 4. The F1 value of ours' is 1.58% higher than that of the BERT-BiLSTM-CRF model, which is a common benchmark model, while comparing the F1 values of BiLSTM-CRF and BERT-CRF models also shows an improvement of 12.83% and 2.96% respectively, there are two main reasons for this: One is the effect of pre-trained language model. Using RoBERTa-WWM as a pre-trained language model, NER tasks can get better results than BERT. On the other hand, since the feature extractor composed of TextCNN-BiGRU can extract text features from multiple angles, it can better extract context information, thus improving the performance of the whole model.

Table 3. Recognition rate of each entity in foreign-related datasets

Model	Precision/%	Recall/%	F1score/%
BILSTM-CRF	79.72	77.41	78.55
BERT-CRF	88.48	88.88	88.42
BERT-BILSTM-CRF	89.78	89.86	89.80
ours	**91.65**	**91.35**	**91.38**

Table 4 shows that the RoBERTa-WWM-TextCNN-BiGRU-CRF entity recognition model outperformed the other models on the entity recognition task on the resume dataset, which indicates that the model is appropriate for different datasets and is robust.

Table 4. Named entity identification results for Resume datasets

Model	Precision/%	Recall/%	F1score/%
Lattice-LSTM [17]	94.81	94.10	94.46
LR-CNN [18]	95.37	94.84	95.11
BERT [19]	94.2	95.5	95.0
softLexicon LSTM [20]	95.30	95.77	95.53
FLAT [21]	-	-	95.45
LGN [22]	95.28	95.46	95.37
ZEN2 [23]	95.34	96.17	95.75
ours	**97.82**	**97.82**	**97.81**

5 Conclusion

We focus on the task of NER in the domain of foreign affairs, by building a text-annotated dataset for NER in the domain of foreign affairs, and after studying and learning from the previously proposed models, an appropriate RoBERTa-WWM-TextCNN-BiGRU-CRF model for CNER is given. RoBERTa-WWM has richer training content and training methods, The sequence modeling layer of TextCNN-BiGRU can not only capture the local information of each part, but can also extract the global information from the context and gain a deeper understanding of the semantics of the context as well as determine the boundary between the entities. The final step is to use the CRF model to add constraints to make the output more reasonable. The precision of the model is 0.916, the recall rate is 0.913, and the F1 is 0.913, which are improved over the general-purpose BERT-BiLSTM-CRF by performing experiments on a foreign affair, self-constructed dataset. Furthermore, the model proposed in this paper also performs better solution compared to the baseline model and validates the robustness of the resume dataset under the open dataset proposed in this study.

Acknowledgements. This work was supported by the Center for Language Education and Cooperation Commissioning Projects (22YHXZ1011), the Construction Project of Teaching Quality and Teaching Reform Project for Undergraduate Universities in Guangdong Province.

References

1. Li, J., Wang, P.: Methods of Chinese named entity recognition. J. Comput. Age (04), 18–21 (2021). https://doi.org/10.16644/j.cnki.cn33-1094/tp.2021.04.005
2. Xie, R., Liu, Z., Jia, J., et al.: Representation learning of knowledge graphs with entity descriptions. In: Proceedings of the AAAI Conference on Artificial Intelligence (2016)
3. Peng, S.: Personalized course resource recommendation algorithm based on deep learning in the intelligent question answering robot environment. Int. J. Inf. Technol. Syst. Approach (IJITSA) **16**(3), 1–13 (2023)

4. Li, Y.: Study and implementation on key techniques for an example based machine translation system. In: 2010 Second IITA International Conference on Geoscience and Remote Sensing (2023). https://doi.org/10.1109/IITA-GRS.2010.5604108
5. Huang, C., Zhao, H.: Word formation by characters: a new method of Chinese word segmentation. In: Proceedings of the 25th Anniversary Academic Conference of Chinese Language Information Society of China, pp. 53–56. Chinese Language Information Society of China, Beijing (2006)
6. Peters, M.E., Neiamm, M., Iyyer, M.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 2227–2237. Association for Computational Linguistics, Stroudsburg (2018)
7. Devlin, J., Chang, M., Lee, K.: BERT: pretraining of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 4171–4186. Association for Computational Linguistics, Stroudsburg (2019)
8. Feng, X., Liu, X.: Sentiment Classification of Reviews Based on BiGRU Neural Network and Fine-grained Attention (2023)
9. Cui, Y., Che, W., Liu, T., et al.: Revisiting Pre-Trained Models for Chinese Natural Language Processing (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.58
10. Zhao, Z., Wang, H.: MaskGEC: improving neural grammatical error correction via dynamic masking. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 1, pp. 1226–1233 (2020). https://doi.org/10.1609/aaai.v34i01.5476
11. Liu, Y., Ott, M., Goyal, N.: RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019). https://doi.org/10.48550/arXiv.1907.11692
12. Kim, Y.: Convolutional Neural Networks for Sentence Classification. Eprint Arxiv (2014). https://doi.org/10.3115/v1/D14-1181
13. Lecun, Y., Bottou, L.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998). https://doi.org/10.1109/5.726791
14. Zhou, F., Jin, L., Dong, J.: Review of convolutional neural networks. Chin. J. Comput. **40**(06), 1229–1251 (2017)
15. Chung, J., Gulcehre, C., Cho, K.H.: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. Eprint Arxiv (2014). https://doi.org/10.48550/arXiv.1412.3555
16. Lipton, Z.C., Berkowitz, J., Elkan, C.: A Critical Review of Recurrent Neural Networks for Sequence Learning. Computer Science (2015). https://doi.org/10.48550/arXiv.1506.00019
17. Zhang, Y., Yang, J.: Chinese NER using lattice LSTM. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1554–1564 (2018)
18. Gui, T., Ma, R., Zhang, Q.: CNN-based Chinese NER with lexicon rethinking. In: International Joint Conference on Artificial Intelligence (IJCAI), pp. 4982–4988 (2019)
19. Devlin, J., Chang, M.W., Lee, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018). https://doi.org/10.48550/arXiv.1810.04805
20. Proceedings of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg (2019)
21. Peng, M., Ma, R., Zhang, Q., et al.: Simplify the usage of lexicon in Chinese NER. In: Proceedings of the 58 Annual Meeting of the Association for Computational Linguistics, pp. 5951–5960. Association for Computational Linguistics, Stroudsburg (2020)
22. Li, X.N., Yan, H., Qiu, X.P.: FLAT: Chinese NER using flat-lattice transformer. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1–8. Association for Computational Linguistics, Philadelphia (2020)

23. Gui, T., Zou, Y.C., Zhang, Q.: A lexicon-based graph neural network for Chinese NER. In: Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 1039–1049. Association for Computational Linguistics, Philadelphia (2019)