ZIBELINE INTERNATIONAL™ PUBLISHING
ISSN: 2616-5961 (Online)
CODEN: IMCSBZ

Information Management and Computer Science (IMCS)

DOI: http://doi.org/10.26480/imcs.01.2025.01.05

REVIEW ARTICLE

ROBERTA-BILSTM-CRF CHINESE NAMED ENTITY RECOGNITION BASED ON MULTI-HEAD ATTENTION

Yinfei Ruan, Shanli Ye

School of Science, Zhejiang University of Science and Technology, Hangzhou 310023, China.
****Corresponding author E-mail: slye@zust.edu.cn***

ARTICLE DETAILS

Article History:

Received 03 December 2024
Revised 07 January 2025
Accepted 20 January 2025
Available online 09 February 2025

ABSTRACT

In this paper, we propose a Chinese Named Entity Recognition (NER) model based on Multihead Attention Mechanism (MAM), RoBERTa-BiLSTM-MHA-CRF, which combines the deep semantic representation capability of RoBERTa, the global context modelling capability of MAM, and the advantage of BiLSTM-CRF in capturing sequence dependencies, to provide a novel solution for the Chinese NER task. , and the advantage of BiLSTM-CRF in capturing sequence dependencies, providing a novel solution for the Chinese NER task. The experimental results on the MSRA dataset show that the model outperforms the mainstream NER model in the key metric of F1 value, with an F1 value of 95.43%. The ablation experiments further validate that the semantic understanding capability of RoBERTa, the long-distance dependency modelling capability of the multi-attention mechanism, and the role of CRF for global label optimisation all have important contributions in performance improvement. Compared with traditional methods, the model not only significantly enhances the ability to recognise complex entity boundaries, but also improves the model's ability to comprehensively understand contextual information. The results show that the RoBERTa-BiLSTM-MHA-CRF model is able to effectively solve the semantic ambiguity and long-distance dependency problems in the Chinese NER task, and has high academic research value and application potential. Future work will focus on exploring the model's adaptability in specific domains and its performance in low-resource scenarios to extend its capabilities in practical applications.

KEYWORDS

Chinese Named Entity Recognition, RoBERTa, Multi-attention Mechanism,Bidirectional Long and Short-term Memory Networks, Conditional Random Fields

1. INTRODUCTION

Named Entity Recognition (NER) represents a critical research area within Natural Language Processing (NLP), focusing on identifying and extracting named entities, such as individuals' names, geographical locations, organizational names, and domain-specific terminologies from unstructured Chinese text. As one of the foundational technologies in NLP, NER holds extensive applications in fields like information extraction, question-answering systems, and semantic understanding.Traditional approaches to Chinese NER largely relied on statistical machine learning techniques, including Hidden Markov Models, Maximum Entropy Models, Support Vector Machines (SVMs), and Conditional Random Fields (CRFs). However, these methods often faced challenges in handling complex linguistic structures, capturing long-distance dependency features, and comprehending semantic nuances。

Recent advancements in deep learning have significantly revitalized the NER task, enabling more robust solutions. Researchers have integrated CRFs with linguistic structure features and dependency representations, achieving notable improvements in biological NER tasks (Peng, 2009). A Lattice LSTM model was introduced, which incorporates word and word sequence information to mitigate segmentation errors, achieving an F1 score of 93.18% on the MSRA dataset (Zhang et al., 2018).

The emergence of pre-trained language models has further enhanced NER performance. For instance, Some scholars developed the BERT-BiLSTM-CRF model, leveraging BERT's deep semantic representations to improve recognition accuracy to 95.91% on the People's Daily corpus (Shen et al., 2022). Luo and co-researchers optimized the BiLSTM-CRF model with an attention mechanism, achieving an F1 score of 91.14% on the BioCreative IV dataset (Luo et al., 2018). Additionally, Wu et al. introduced a joint segmentation and CNN-BiLSTM-CRF training framework, which enhanced boundary recognition using pseudo-tagged samples (Wu et al., 2019). Further modular innovations have pushed NER research forward. For example, the RoBERTa model was proposed, which improves pretraining strategies based on BERT, increases model capacity, and incorporates byte encoding to enhance lexical representation accuracy and efficiency (Liu et al., 2019). Qiu introduced the BERT-CNN-BiLSTM-CRF architecture, integrating CNN and Multi-Head Attention mechanisms to further elevate Chinese NER performance (Qiu, 2024). These contributions highlight the transformative role of deep learning and pre-trained models in advancing the field of NER.

Based on the above research, this paper proposes an innovative Chinese named entity recognition model integrating RoBERTa-BiLSTM-CRF with Multi-Head Attention mechanism, which combines the powerful representation capability of RoBERTa, the global information extraction capability of Multi-Head Attention mechanism, and the modelling advantage of BiLSTM-CRF on sequence dependency, providing an better solution for Chinese named entity recognition

2. MODELING

2.1 Overview of the Model

Quick Response Code	Access this article online	
	Website: www.theimcs.org	**DOI:** 10.26480/imcs.01.2025.01.05

In this paper, we establish the RoBERTa-BiLSTM-CRF model based on the Multi-Head Attention mechanism, which is an end-to-end language model that can better capture the syntactic and semantic features present in the text and automatically understand the contextual relevance, and the Multi-Head Attention mechanism can focus on multiple different parts of the input sequence at the same time to improve the model's ability to understand the context, and it also allows for a Different types of contextual information can be modelled at a finer granularity, making the model more flexible in processing diverse information. The model is mainly composed of four modules, namely, Multi-Head Attention, RoBERTa, BiLSTM and CRF. The model generates contextually relevant semantic representations by passing the input text sequences into the RoBERTa module, and then extracts the contextual dependencies by the BiLSTM module, followed by the Multi-Head Attention module in parallel, and then by the Multi-Head Attention module in parallel. and then the Multi-Head Attention module pays attention to different parts of the sequence in parallel to further enhance the understanding of the complex context, and finally decodes the labelled sequence through the CRF module to output the globally optimal sequence annotation results. The overall structure of the model is shown in Figure 1.

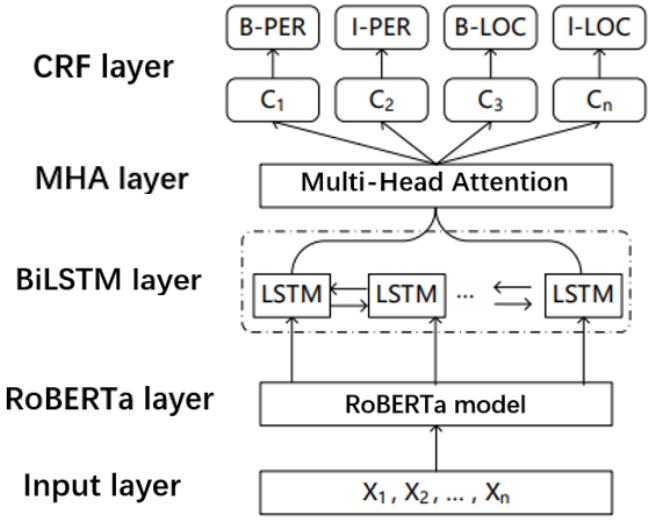

Figure 1: Structure of the RoBERTa-BiLSTM-MHA-CRF model

2.2 RoBERTa

2.2.1 The BERT model

Traditional language models, such as GloVe, Word2Vec and GPT, fail to solve the polysemy problem. For example, in In the phrases 'meeting' and 'opening the door', 'open' expresses completely different meanings , but in these models, the two instances of 'open' are regarded as having the same meaning, and their word vectors have exactly the same value. The BERT model introduced by the Google team in 2018 effectively addresses the challenge of word polysemy. It uses a bi-directional Transformer (Transformer) architecture as an encoder, enabling it to take contextual input into account when predicting subsequent characters.

In the BERT model, the input vector consists of three kinds of embeddings: token embedding, segment embedding, and position embedding. These embeddings work together to enable the model to capture the semantic, structural and sequential information of the input sequence. The token embedding represents the semantic information of each word or sub-word unit in the input sequence.BERT uses WordPiece segmentation technique to decompose words into smaller sub-word units. For example, the word 'unbelievable' is broken into 'un' and '##believable'. Each sub-word unit corresponds to an embedding vector that is stored in a pre-trained word list (Vocabulary) and reflects the semantic features of the sub-word unit. Segment embeddings are used to distinguish the origin of sentence pairs. For the input sentence pairs (Sentence A and Sentence B), BERT assigns different segment embedding vectors to the words of the two: all the words of Sentence A are assigned as Segment A embeddings all the words of Sentence B are assigned as Segment B embeddings. If the input contains only a single sentence, all words share the same embedding (as in Segment A). The positional embeddings provide information about the position of each word in the input sequence, since the Transformer architecture itself does not have the ability to model sequence order. Each position (e.g., 1st word, 2nd word, etc.) is mapped to a fixed embedding vector.BERT uses learnable positional embeddings that are automatically adapted during training to optimise the sequence representation. By integrating these three embeddings, BERT is able to efficiently understand and process the multidimensional information of the input text.An example of the BERT model input is shown in Figure 2.

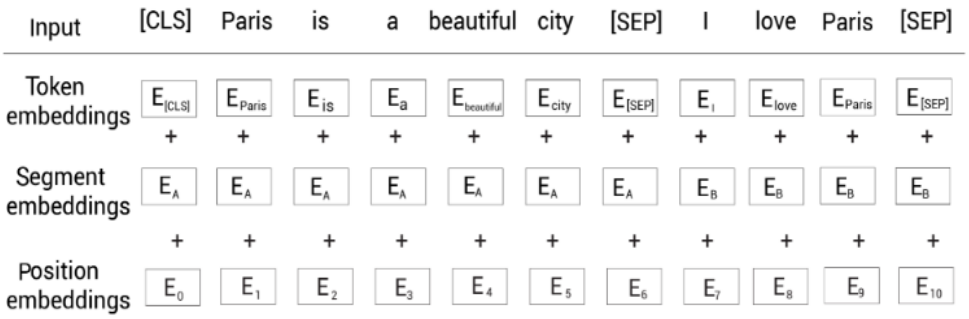

Figure 2: Representation of BERT model inputs

2.2.2 RoBERTa model

The Masked Language Model (MLM) task of the BERT model is one of the core aspects of its pre-training. In this task, BERT masks 15% of the words in the input sequence by randomly masking them (replacing them with special symbols [MASK]) and trains the model to predict these masked words. In contrast, RoBERTa introduces a dynamic masking strategy.

In BERT, the masked tokens of each sample sequence are fixed after the first masking, and a static masking approach is used. In contrast, RoBERTa generates multiple masked versions for each input sample through a dynamic masking strategy. For example, each sample sequence is copied 10 times, and each copy randomly generates different masked tokens. During the training process, sentences with different masks are input to the model for each epoch, and sentences with the same mask tag are seen by the model again only after 10 epochs of training. Taking sentence 1 as an example, its mask-tagged version will be used in epoch 1, epoch 11, epoch 21 and epoch 31. The application of dynamic mask avoids the limitation of static mask, which improves the training efficiency of the model.

In addition, BERT includes a Next Sentence Prediction (NSP) task in pre-training to determine the logical relationship between two sentences. However, studies have shown that the NSP task has limited performance improvement on downstream tasks. Therefore, RoBERTa abandons the NSP task and focuses on the more efficient MLM task. In terms of training data, BERT's pre-training dataset includes BOOKCORPUS and English Wikipedia, with a total size of 16GB, while RoBERTa uses larger datasets including Toronto BookCorpus, English Wikipedia, CC-News (Common Crawl-News), Open WebText, and Stories (Common Crawl-News). In terms of training parameters, BERT has a pre-training batch size of 256 with a total of 1M steps (1,000,000 steps), while RoBERTa uses a larger batch size (8000) with 300, 000 steps. 000 steps. With the same batch size, an additional version with 500,000 steps was trained. The larger batch size not only significantly increased the training speed, but also improved the model performance. These improvements allow RoBERTa to capture semantic information more efficiently and significantly improve its performance on downstream tasks.

2.3 BiLSTM layer

2.3.1 LSTM

LSTM is an improved RNN, but LSTM can only process information from one direction of the input sequence (usually left to right or right to left). This means that it can only rely on previous states to predict the current output, limiting its ability to capture future information.The rough internal structure of LSTM is shown below:

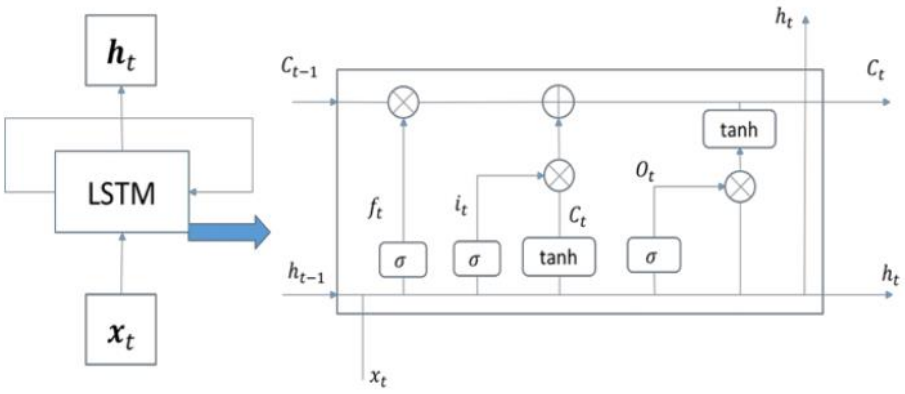

Figure 3: Structure of LSTM

Each LSTM cell consists of a combination of real vectors, including the input gate vector it, the forget gate vector f_t, the output gate vector O_t, the memory cell C_t, and the hidden state h_t.

The formulae within each variable are as follows:

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

$$i_t = \sigma(w_i \cdot [h_{t-1}, x_t] + b_i)$$

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

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

$$O_t = \sigma(w_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = O_t * tanh(C_t)$$

where x_t is the input of time step t, h_{t-1} is the hidden state of the previous time step, i_t , f_t, ot are the input gate, forgetting gate and output gate, respectively, σ is the Sigmoid function, and tanh is the hyperbolic tangent function. w is the weight matrix, b is the bias vector, and is the temporary state of the cell, $\tilde{C}_t$ is the state of the time t, and h_t is the time t. output

2.3.2 BiLSTM

The BiLSTM module used in this paper is a state-of-the-art recurrent neural network structure, BiLSTM is composed of two independent LSTMs that are responsible for processing the input sequence from two directions (forward and reverse). This allows the model to acquire information both before and after the current time step, which combines the advantages of LSTM models in capturing long-term dependencies with enhanced bidirectional perception of contextual information. This structure also effectively solves the problems of gradient vanishing and gradient explosion, thus enabling a global understanding of the entire sequence. The BiLSTM module structure is shown in Figure 4.

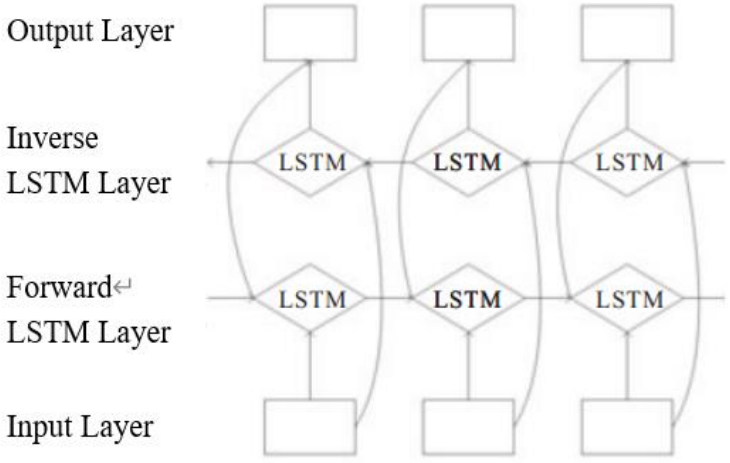

Figure 4: Structure of BiLSTM

2.4 Multi-Head Attention Mechanism

The Multi-Head Attention mechanism is a core component of the Transformer model, which processes input features (i.e., queries, keys, and values) by means of multiple independent attention modules (often referred to as 'heads') that run in parallel. Each attention head independently computes attention weights and generates corresponding weighted outputs, which are then merged by splicing or averaging to form a more complex and richer representation.

The Multi-Head Attention mechanisms offer significant advantages over traditional attention mechanisms. Conventional attention mechanisms can only compute a single weighted average, and their weights are entirely dependent on the correlation between inputs, making it difficult to

comprehensively capture the diverse information in the inputs. In contrast, the Multi-Head Attention mechanism enhances the diversity and expressiveness of the representation by learning multiple independent semantic features in parallel. In the BiLSTM model, although it can effectively capture contextual dependencies, it still has some limitations in dealing with long-distance dependencies. The introduction of the multi-attention mechanism can further enhance the interaction of contextual information between the BiLSTM and CRF layers, thus helping the model to understand complex semantic relationships more accurately. This mechanism provides more comprehensive information to the CRF layer by enriching the input features, which improves the accuracy of the sequence annotation task. In addition, the Multi-Head Attention mechanism is especially effective in capturing long-distance dependencies when dealing with long text and complex structures, thus significantly improving the model's ability to understand the text and the annotation accuracy, the structure of the Multi-Head Attention module is shown in the following Figure 5.

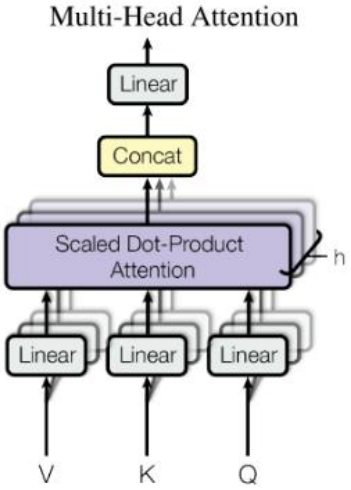

Figure 5: Structure of Multi-Head Attention module

Q, K, and V are obtained by multiplying the input vector X by the three coefficients W^q ,W^k,,W^v respectively, and W^q ,W^k,,W^v are trainable parameter matrices.

2.5 CRF Layer

Conditional Random Fields (CRF) is a statistical modelling tool commonly used in sequence prediction tasks, which is able to effectively take into account the interdependence between labels in a sequence. Therefore, it is particularly suitable for scenarios involving the annotation or classification of sequence data, such as Named Entity Recognition (NER) and Lexical Annotation (POS). In the entity recognition task of Chinese text, the Multi-Head Attention mechanism is unable to directly model the dependencies between tags, although it is good at capturing contextual semantic information. For example, the 'I-ORG' tag should not directly follow the 'B-PER' tag. Such dependencies need to be modelled through the CRF layer to impose reasonable constraints on the predicted tags.

The CRF ensures that the sum of probabilities of all possible tag sequences is equal to 1 by constructing a model of the tag sequences and implementing global normalisation; in this paper, the CRF is embedded into the BiLSTM module and used to process the output of BiLSTM. For a sentence instance Instance={X_1,X_2,...,X_n} where X_i is a word in the sentence, it is put into the recognition framework and after BiLSTM we get each word denoted as hi. Subsequently, the output of BiLSTM is used as an input to the CRF. By learning the dependencies between the labels, the CRF assigns label y_i (e.g., 'entity' or 'non-entity') to each word. In addition, the order and interrelationships of the labels play a key role in the final annotation result. For each possible tag sequence y={y_1,y_2,...,y_n}, the CRF module defines its probability distribution as follows:

$$h_i = BiLSTM(x_i)$$

$$P(y1, y2, \ldots, yn \mid h1, h2, \ldots, hn) = \frac{\exp\left(\sum_{i=1}^{n} A_{y_i, y_{i-1}} + B_{yi} h_i\right)}{\sum_{y' \in Y} \exp\left(\sum_{i=1}^{n} A_{y_i, y_{i-1}} + B_{yi} h_i\right)}$$

where: h_i is the hidden state representation of the BiLSTM output, capturing the contextual information of each word. $A_{yi,yi-1}$ is the transfer matrix in the CRF module, representing the transfer probability between labels (e.g., the probability of transferring from label y_{i-1} to label y_i), and B_{yi} is the score corresponding to the output label y_i,usually is the scalar product of BiLSTM outputs.Y is the sequence of all possible labels.

3. Experimental Results and Analyses

3.1 Data set labelling and assessment indicators

This study uses the publicly available Named Entity Recognition (NER) dataset from Microsoft Research Asia (MSRA). This dataset contains rich entity types, including person names (PER), place names (LOC), and organisation names (ORG). Its high annotation quality and data accuracy make it suitable for training and evaluating NER models. The commonly used annotation systems for named entity recognition include BIO, BIOE and BIOES, and in this paper, we adopt the BIO system, which has 7 types of labels, namely 'O' (non-entity), 'B-PER', 'I-PER', "B-ORG", "I-ORG", "B-LOC" and 'I-LOC'.

Compared with other datasets, the MSRA dataset has significant advantages in terms of corpus size, entity coverage and linguistic complexity, and is a reliable benchmark for Chinese information extraction and processing. In this experiment, the training and test sets are divided in the ratio of 8:2, and the focus is on identifying and evaluating the annotation performance of three types of entities, namely, personal names (PER), place names (LOC) and organisational names (ORG). Compared with other datasets, the MSRA dataset has significant advantages in terms of corpus size, entity coverage and linguistic complexity, and is a reliable benchmark for Chinese information extraction and processing. In this experiment, the training and test sets are divided in the ratio of 8:2, and the focus is on identifying and evaluating the annotation performance of three types of entities, namely, personal names (PER), geographical names (LOC) and organisational names (ORG).

Table 1: MSRA corpus segmentation information (unit: units)

	Number of sentences	Number of characters	Number of entities
Training set	45000	2171573	75059
Test set	3442	172601	6190

The specific proportions of the corpus are shown in the table: in this paper, we use recall R, precision P and F1 value to judge the performance of the model, and each evaluation index is calculated as follows:

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

$$R = \frac{TP}{T_p + F_N} \times 100\%$$

$$F_1 = \frac{2 \times P \times R}{P + R} \times 100\%$$

where T_p is the number of correctly identified entities, ;F_p is the number of non-entities whose predictions result in real bodies, and F_N is the amount of unpredicted entities.

3.2 Experimental environment as well as parameters

This experiment is based on the PyTorch framework to build neural network models, and the running environment is Linux operating system. The hardware configuration includes Intel Xeon Platinum 8352V processor as CPU and NVIDIA GeForce RTX 4090 graphics card (with 24GB of video memory) as GPU.The Python version is 3.8.10 and the deep learning framework is PyTorch 2.0.In order to alleviate the problem of parameter overfitting, this experiment adopts AdamW optimiser with the learning rate set to 0.00004. In the added Multi-Head Attention module, the number of attention heads is 8, the dimensionality of the hidden layer is 1024, and the number of transformer layers is 12. In addition, in order to further reduce the effect of overfitting, a random deactivation (Dropout) mechanism is introduced and the deactivation probability is set to 0.1.

Table 2: Model hyperparameter settings

Parameters	Value
Transformer layers	12
Epoach	5
Learning rate	0.00004
The number of the Multi-Head Attention	8
Batch size	32
Dropout	0.1

3.3 Experimental results and analyses

In order to validate the effectiveness of the model proposed in this study, the model is compared with the current mainstream named entity recognition model, and the running results of the experiments after testing on the MSRA dataset are shown in Table 3.

Table 3: Comparative results of model testing on MSRA corpus (unit: %)

Models	P	R	F1
ALBERT-CRF	88.89	90.46	89.67
BERT-IDCNN-CRF	94.86	93.97	94.41
ERNIE-BiGRU-CRF	94.40	93.26	93.84
Lattice-LSTM-CRF	93.57	92.79	93.18
Word2vec-BiLSTM-CRF	85.37	84.12	84.72
RoBERT-BiLSTM-MHA-CRF	95.10	95.77	95.43

The above results show that in the Chinese named entity recognition (NER) task, the RoBERTa model demonstrates stronger representational capabilities and is able to provide richer word vector information for the subsequent BiLSTM layer, thus effectively improving the accuracy of entity recognition. Compared to other models, such as ALBERT-CRF (F1 = 89.67) and Word2vec-BiLSTM-CRF (F1 = 84.72) the RoBERTa module significantly outperforms the Word2vec module in terms of feature representation and information capture. Therefore, the RoBERTa-BiLSTM-MHA-CRF model based on RoBERTa exhibits superior performance in terms of F1 values. The results of the comparison experiments show that the RoBERTa-BiLSTM-MHA-CRF model achieves an F1 value of 95.43, which is higher than that of BERT-IDCNN-CRF (F1 = 94.41) and ERNIE-BiGRU-CRF (F1 = 93.84). Although large models with high complexity such as BERT and ERNIE are already competitive in terms of performance, RoBERTa-BiLSTM-MHA-CRF not only significantly improves the model's semantic comprehension ability, but also captures entity boundaries and contextual information more accurately through the introduction of the Multi-Head Attention mechanism and optimisation based on the RoBERTa model, thereby achieving the best F1 value in this task. In summary, compared with the mainstream module combination models in the current NER task, the RoBERTa-BiLSTM-MHA-CRF model exhibits optimal performance, which fully proves its important research value in improving the named entity recognition accuracy.

In order to further analyse in depth the extent to which the modules in the RoBERTa-BiLSTM-MAH-CRF model affect the performance of the model, six sets of ablation experiments are conducted in this paper, namely RoBERTa-CRF, RoBERTa-BiLSTM, BERT-BiLSTM-CRF, RoBERTa-BiLSTM-CRF, BERT- BiLSTM-MAH-CRF and RoBERTa-BiLSTM-MAH-CRF. The experimental results of each model are shown in Table 4.

Table 4: Experimental results of model ablation in MSRA corpus test (Unit: %)

Models	P	R	F1
RoBERTa - CRF	90.34	92.18	91.25
RoBERTa -BiLSTM	91.28	93.52	92.39
BERT-BiLSTM-CRF	94.42	95.39	94.90
RoBERT-BiLSTM-CRF	94.91	95.79	95.35
BERT-BiLSTM-MHA-CRF	94.55	95.35	94.95
RoBERT-BiLSTM-MHA-CRF	95.10	95.77	95.43

The experimental results show that when replacing the BERT module with the RoBERTa module, the F1 value is improved by 0.45% over the baseline model BERT-BiLSTM-CRF. This improvement is attributed to RoBERTa's advantages in contextual semantic capture and textual representation capabilities, which enhance the quality of the input features, thus improving the overall performance of the model. The further introduction of the Multi-Head Attention mechanism improves the F1 value of the RoBERTa-BiLSTM-MAH-CRF model by another 0.08%, which is attributed to the ability of the Multi-Head Attention mechanism to capture word-to-word dependencies from multiple perspectives, enhancing the potential semantic relevance of the current information to different heads, which helps to identify difficult entities and improve detection accuracy. After adding the CRF layer, the F1 value of the RoBERTa-BiLSTM-CRF model improves from 92.39 to 95.35, indicating that the CRF layer optimises the overall score of sequence annotation by capturing the global dependencies between labels and helps the model to obtain the globally optimal label sequences, which significantly improves the accuracy of entity recognition. In the RoBERTa-CRF model, the further introduction of the BiLSTM module leads to a 4.1% improvement in the F1 value, as the bidirectional sequence modelling capability of BiLSTM better captures fine-grained contextual information and complements the global features of RoBERTa. In addition, the enriched feature representation enhances the accuracy of the CRF module in entity boundary determination and label prediction, which significantly improves the F1 value for the named entity recognition task.The RoBERTa-CRF model outperforms the ALBERT-CRF model in terms of F1 value, with a 1.58% improvement in F1 value. This result indicates that RoBERTa has gained stronger context modelling and language understanding through large-scale data and long training time, which provides the CRF module with more accurate feature representations and helps it to perform more accurate entity decoding. Although ALBERT effectively reduces the model size in terms of parameter optimisation, it does not perform as well as RoBERTa when dealing with complex tasks.Therefore, RoBERTa's deeper semantic understanding and larger model capacity enable it to perform better in the NER task, especially in terms of F1 value enhancement. In summary, the experimental results show that the RoBERTa-BiLSTM-MAH-CRF model proposed in this paper performs well in the Chinese named entity recognition task, with a final F1 value of 95.43%.

4. Conclusion

In this paper, a novel entity recognition model is proposed, which improves the recognition performance in the Chinese named entity recognition task by integrating the Multi-Head Attention mechanism in the RoBERTa-BiLSTM-CRF model with an F1 value of 95.43%, and the comparison with some mainstream models also verifies the unique superiority of the RoBERTa-BiLSTM-MHA-CRF model in the Chinese named The next step is to apply the model to specific domains to complete entity recognition and extraction, and on the other hand, to explore how to give full play to the performance of the model in a training environment with limited samples.

References

Chunyan, P., Hui, Z., and Lingyu, B., 2009. Biological nomenclature entity recognition based on conditional random domain [J]. Computer Engineering, 35 (22), Pp. 197-199.

Li, Q.L., Yan, S.J., Chen, Q., and Zhang, K., 2024, December. Research on Chinese Named Entity Recognition Based on BERT-CNN-BiLSTM-CRF Model with Fusion Multi-Head Attention Mechanism. In 2024 14th International Conference on Information Science and Technology (ICIST), Pp. 583-588.

Liu, Y., 2019. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 364.

Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., and Wang, J., 2018. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics, 34 (8), Pp. 1381-1388.

Shen, T.P., Yu, L., Jin, L., 2022. Research on Chinese entity recognition based on BERT-BiLSTM-CRF model [J]. Journal of Qiqihar University (Natural Science Edition), 38(1), Pp. 26-32.

Wu, F., Liu, J., Wu, C., Huang, Y., and Xie, X., 2019. Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation. In The World Wide Web Conference, Pp. 3342-3348.

Zhang, Y., Yang J., 2018. Chinese NER using lattice LSTM[J]. arxiv preprint arxiv:1805.02023.