Journal of Information Hiding and Multimedia Signal Processing ©2025 ISSN 2073-4212
Ubiquitous International Volume **16**, Number **3**, September **2025**

Chinese Named Entity Recognition Based on Lexicon Knowledge Enhancement

Chengqiong Ye*

College of Computing and Information Technologies, National University
Manila 1008, Philippines
School of Big Data and Artificial Intelligence, Anhui Xinhua University
Hefei 230088, P. R. China
yechengqiong@axhu.edu.cn

Alexander A. Hernandez

College of Computing and Information Technologies, National University
Manila 1008, Philippines
aahernandez@national-u.edu.ph

Mideth Abisado

College of Computing and Information Technologies, National University
Manila 1008, Philippines
mbabisado@national-u.edu.ph

*Corresponding author: Chengqiong Ye
Received April 15, 2025, revised May 23, 2025, accepted May 24, 2025.

Abstract. *This study explores named entity recognition (NER) and suggests a Chinese NER model based on lexical knowledge enhancement, integrating external lexicon knowledge into the Chinese BERT pre-training model to improve model performance. First, using external lexicon matching, the input sequence is converted into a character-word pair sequence. Subsequently, by devising an attention mechanism-based deep interaction module for characters and words, the fusion of character vectors and matched word vectors takes place among the underlying Transformers of the BERT model. Finally, the globally optimal prediction result is obtained through the sequence decoding layer. In contrast to models operating at a single character level, the present model incorporates entity-level Chinese word granularity information during training, achieving a thorough integration of both character and word granularity information, this significantly improves the model's accuracy of Chinese entity boundary division and results in superior performance on the task of Chinese NER. The Micro-F1 scores of this model on the Note4, MSRA, Weibo, and Resume datasets reached 92.04%, 95.70%, 70.35%, and 96.07%, respectively.*

Keywords: Named entity recognition; Knowledge enhancement; BERT; Attention mechanism; Chinese word granularity

1. **Introduction.** As the internet increasingly permeates our lives, generating massive volumes of information, the majority of this data is available in natural language format [1]. In automating the processing and analysis of such data to extract critical information, Named Entity Recognition (NER) technology has emerged and gained widespread attention. Based on its characteristics, it can be categorized into three types: rule-based matching methods, machine learning-based methods, and deep learning-based methods.

Rule-based matching methods involve using predefined rules by experts to extract entities from text. The construction of these rules requires reference to linguistic knowledge such as grammar and morphology, as well as domain knowledge including domain-specific abbreviations, specialized vocabulary, and special grammar, etc. [2]. Appelt et al. [3] created the first named entity recognition system based on rules in 1995. That same year, researchers like Morgan et al. [4] further integrated linguistic features and rules into the systems, enhancing recognition reliability.

Machine learning-based methods involve using machine learning models to determine sequences and label them, then apply labeling rules to label and extract named entities from text [5]. Machine learning-based models include linear classification models like Support Vector Machines [6], HMM models [7], and linear-chain CRF [8, 9], etc. Most of these entity recognition models were proposed for English corpora and are not suitable for Chinese. Some models have been improved for Chinese characteristics with poor effect, for instance, multi-layer conditional random fields [10] and stacked Markov models [11], etc. Although machine learning methods reduce the heavy workload associated with manually designing numerous rules, they still require complex feature engineering.

Researchers are concentrating more and more on deep learning-based techniques to accomplish autonomous feature learning [12, 13]. Three benefits can be summed up regarding the use of deep learning-based techniques for NER tasks: (1) to reduce the dependence on feature engineering; (2) non-linear information can be extracted through non-linear activation functions; (3) the training method of deep learning NER models are based on gradient descent [14], which make it possible to design more complex models. The first deep learning NER model was proposed by Collobert in 2012 [15], which performs convolution operations on the length of a fixed window size on a sequence [16] and uses max-pooling to extract features, mapping the results through the ReLU [17] activation function to the labeling space. In 2018, researchers at Google, including Devlin et al. [18], brought forth the Transformer-based BERT language model [19], and subsequently, NER models fine-tuned from BERT have achieved excellent performance.

However, earlier models were mostly character-level, and for Chinese tasks, obtaining word granularity information is also crucial. Adding lexicon information to character-based models is one way to get word granularity information. In order for LSTM [20] to encode Chinese letters and words and choose the most pertinent characters and words from sentences, Zhang et al. [21] developed a lattice structure. Experimental results showed its performance was better than character-based models, however, the complex lattice structure limited its application in the industrial field. In order to import lexicon information, Ma et al. [22] presented a solution that avoids complex structures by merging the word lexicon into character representations with just minor alterations needed to the character representation layer. Subsequently, Li et al. [23] transformed the lattice structure into a shallow Flat-Lattice Transformer structure to integrate character and word information, further improving model performance. There have also been explorations using lexicons to guide model pre-training; Baidu's ERNIE model [24] integrates knowledge into BERT's pre-training process using entity-level full word masking; Jia et al. [25] further pre-trained BERT for NER tasks using professional domain Chinese datasets; Diao et al. [26] designed a multi-layer N-gram encoder to enhance the Chinese BERT model. Xiao et al. [27] proposed the Multi-View Transformer (MVT) method, which significantly improves the performance of Chinese named entity recognition by constructing the visibility matrix of multiple viewpoints and the viewpoint-aware attention mechanism to efficiently capture the different interaction information between character and word. Zhang et al. [28] proposed the MGBERT-Pointer model, which combines the multi-granularity BERT adapter with the efficient global pointer network, to effectively enhance the processing

capability of complex semantics and fuzzy boundaries and nested structures in Chinese named entity recognition. Liu et al. [29] propose the Sequential Lexicon Enhanced BERT (SLEBERT) method, which effectively reduces the problem of noisy words and vocabulary conflicts by constructing a sequential lexicon and introducing the positional encoding and adaptive attention mechanism.

Although many studies have attempted to combine word granularity information with pre-trained models, most of the existing methods only superficially superimpose word vectors in the input or output phase of the model, failing to effectively integrate word granularity features into the internal structure of the model, resulting in the failure to fully utilize the potential of word granularity information. To address this problem, this paper proposes a Chinese named entity recognition method based on lexicon knowledge enhancement. The method firstly transforms the original input into a sequence of character-word pairs through external lexicon matching; subsequently, a character-word interaction module based on the attention mechanism is introduced between the underlying Transformer layers of the BERT model to realize the in-depth interaction and fusion of the character-level representations with the word-level representations, so as to fine-tune the injection of entity-level Chinese word-granularity knowledge into the internal structure of the model, which significantly enhances the modeling capability of the entity boundary and the semantic information.

2. **Materials and Method.**

2.1. **Overall Model Design.**

This paper proposes a Chinese named entity recognition model based on external lexicon knowledge enhancement, which consists of two parts: the lexicon enhancement BERT model and the decoding layer, and its overall structure is shown in Figure 1.

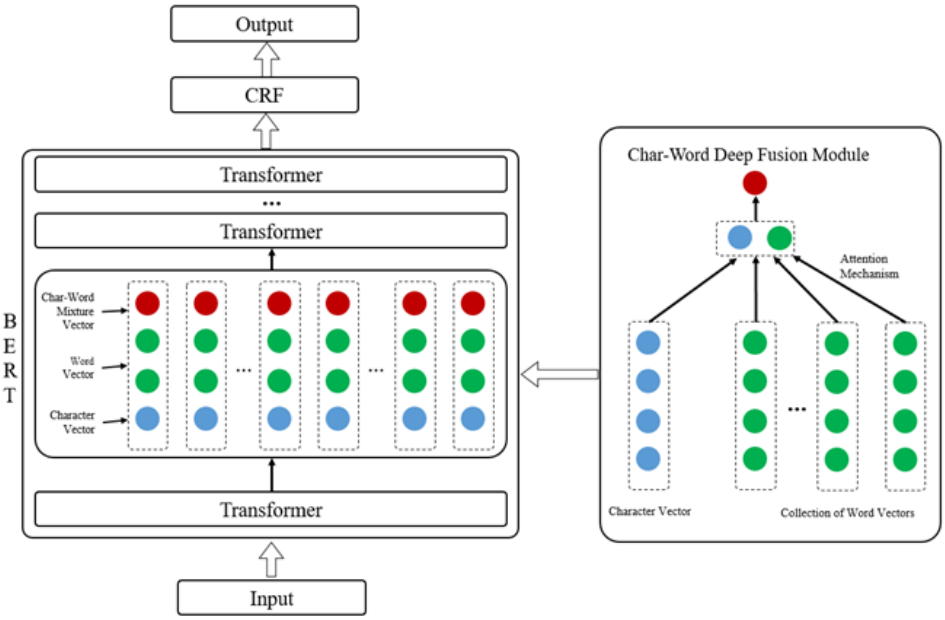

FIGURE 1. Overall structure of the model proposed in this paper

Currently, the more common approach to fusing BERT models with word granularity information is to perform character-word fusion at the Embedding level. Specifically, such approaches first use BERT to model character sequences to capture the dependencies between characters, and then fuse the character features output from BERT with lexical

features, which are finally input into the neural network annotation model. Although such methods can introduce lexical features to a certain extent, they fail to take full advantage of the internal sequence modeling of BERT because the interaction of character-word features only occurs in the shallow network at the end of BERT. In contrast, the lexical knowledge enhancement-based approach proposed in this paper directly incorporates Chinese lexical features in the Transformer layer inside BERT, thus more fully combining character granularity and word granularity information, as shown in Figure 2.

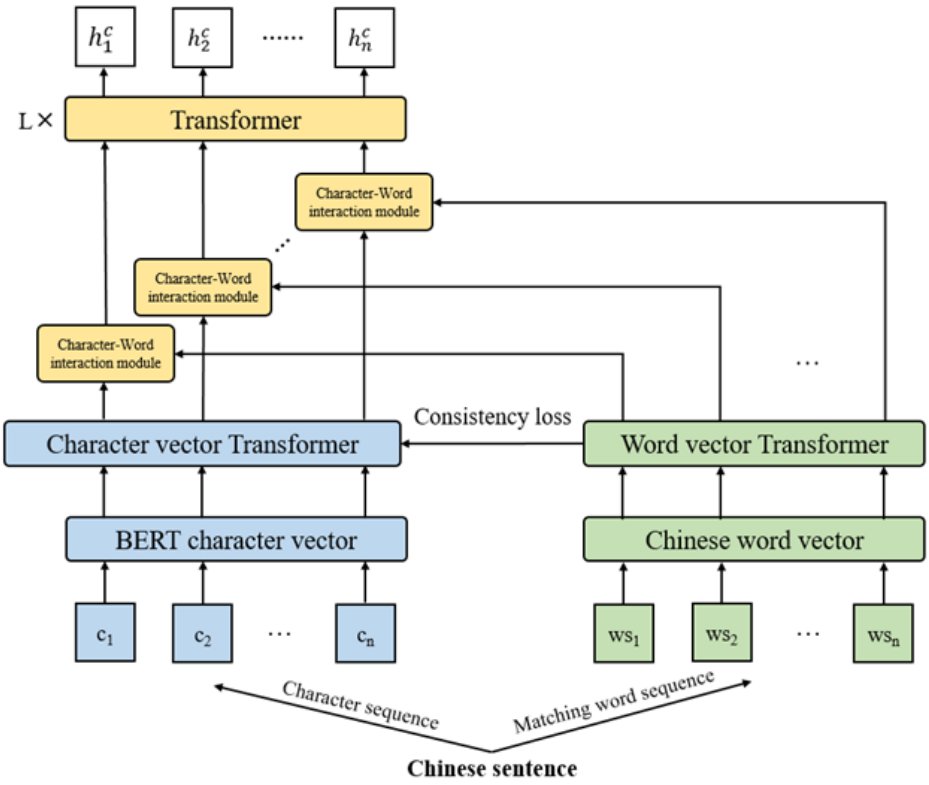

FIGURE 2. BERT model design based on lexicon knowledge enhancement

Specifically, the design of the model in this paper is improved in the following two aspects:

- It utilizes an external Chinese lexicon to transform the input features into character-word pair sequences.
- It designs a gate mechanism-based multi-granularity interaction module for characters and words to thoroughly integrate Chinese character features with word features.

2.2. **Character-word Pair Sequence.** In English, English words are the smallest unit of granularity in text and also the minimum unit bearing semantic meaning. However, in Chinese, Chinese characters are the smallest unit of granularity, but semantic information is often contained within both characters and words. If the model input granularity unit is based on Chinese characters, the model only learns character-level features, resulting in a language model thus trained that lacks word granularity information.

To integrate Chinese word granularity information, the model input is modified by expanding the original Chinese sequence into a sequence represented in the form of character-word pairs. Specifically, an external Chinese lexicon Dict is used to traverse a sentence sequence containing n Chinese characters, $s_c = \{c_1, c_2, c_3, \cdots, c_n\}$, matching it with the lexicon to form a set of character-word pairs. Taking the phrase "上海市长江路" as an

example, six different words can be segmented: “上海”, “上海市”, “市长”, “长江”, “长江路” and “江路”. Subsequently, these are assigned to the corresponding sets for each character contained within the words. As shown in Figure 3, where <PAD> represents a padding token used to standardize the length of inconsistent character-word pairs. Finally, each character is paired with its assigned word set, transforming the representation of a Chinese sequence input into a sequence of character-word pairs, that is, $s_{cw} = \{(c_1, ws_1), (c_2, ws_2), \cdots, (c_n, ws_n)\}$ where c_i indicates the i-th character in the sequence and ws_i represents the set of words matched with character c_i.

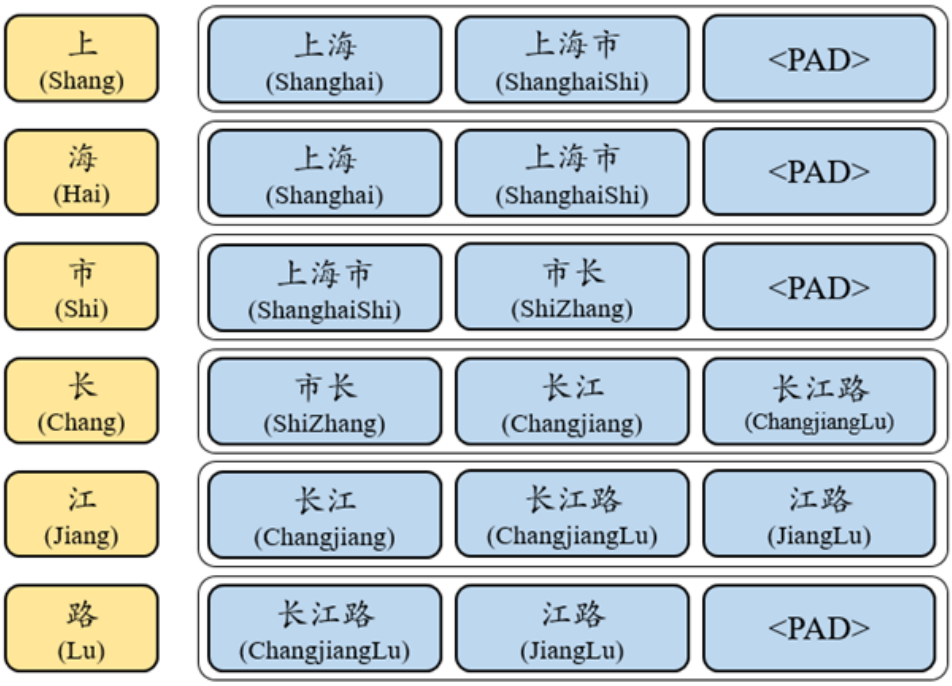

FIGURE 3. Example of character-word pairs

2.3. **Consistency Loss.** To improve the model's capacity to represent the semantics of characters and words, the BERT character vectors and Chinese word vectors must be effectively fused. But the character vectors and word vectors are from distinct models; the word vectors are from other open-source models, and the character vectors are from the BERT pre-trained model. Therefore, the character vectors and word vectors possess completely different semantics, and directly adaptive fusion of them will limit the effectiveness of the model.

To address this issue, this paper models the original character and word vectors obtained through the Transformer to achieve mapping of the original vectors and improve information interaction between them. Considering that both character vectors and word vectors are derived from the same Chinese sentence and possess similar semantic information, a consistency loss function L_{sim} is designed, with the specific formula as follows:

$$L_{\text{sim}} = \text{Cosine}(E_{\text{char}}, E_{\text{word}}) \tag{1}$$

In Equation (1), Cosine$(\cdot)$ represents the cosine distance between vectors, E_{char} is the representation of the Chinese character vector after passing through the Transformer layer, and E_{word} is the representation of the Chinese word vector after passing through the Transformer layer. After the word and character vectors that have gone through the Transformer are averaged, respectively, their similarity is shown by the cosine distance

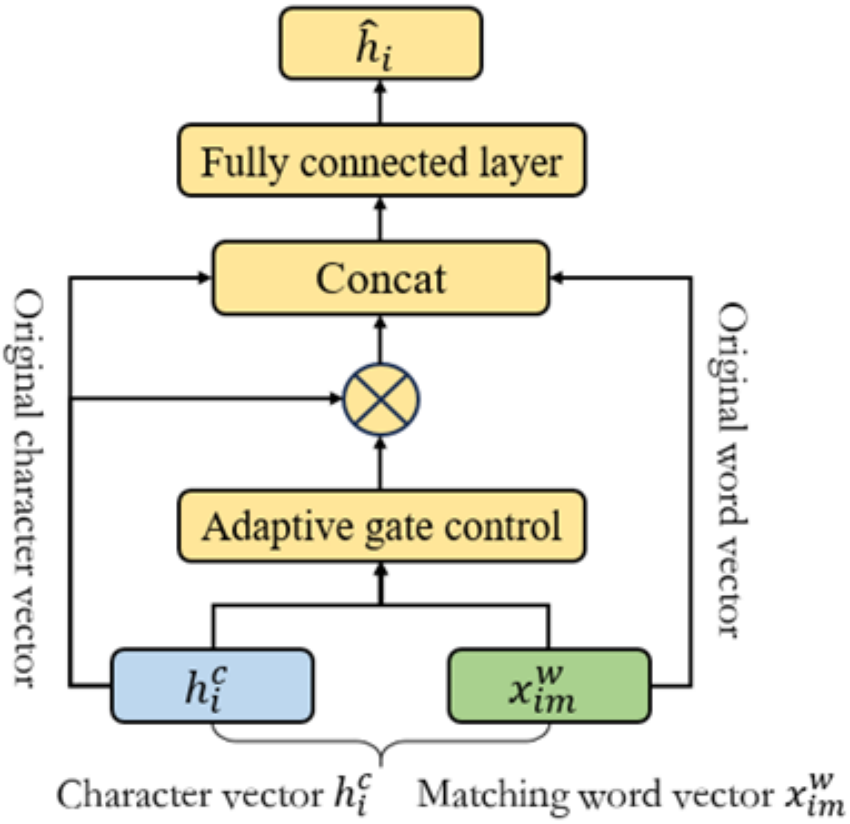

FIGURE 4. Character-word deep interaction module

between the two vectors. Losses are inversely correlated with similarity: higher similarity = smaller loss.

2.4. **Character-word Interaction Module.** To fully integrate Chinese word granularity information, the word information is deeply integrated with the BERT model itself. Inspired by the principle of gating mechanism, this paper designs a novel character-word deep interaction module, as shown in Figure 4. It can directly inject lexicon information into the internal of the BERT model, achieving a thorough fusion of both character and word granularity information in Chinese.

This module receives two parts of input: a character-level vector representation and its matching set of word vectors. For the i-th position in the character-word sequence, the input can be represented as (h_i^c, x_i^w). Here, h_i^c represents the character vector at the i-th position in the sequence, specifically the output vector from the previous Transformer layer in BERT. $x_i^w = \{x_{i1}^w, x_{i2}^w, \ldots, x_{im}^w\}$ denotes a set of pre-trained Chinese word vectors, where m represents the size of the corresponding word set for the current character, and x_{i1}^w is the word vector for the first corresponding word of the current character. The j-th word in x_i^w is represented as shown in (2):

$$x_{ij}^w = e^w \, w_{ij} \tag{2}$$

Where e^w represents the pre-trained word vector lookup table, and w_{ij} indicates the j-th word in x_i^w. To align the character and word vector representations of different dimensions, a non-linear transformation is also required for the corresponding word vectors, as shown in (3). W_1 is a $d_c \times d_w$ matrix, and the result of multiplying it with the word vector

is processed through a tanh activation function. d_w and d_c represent the dimensions of the word vector and character vector, respectively, where the dimension of the character vector is also the hidden layer dimension of BERT. W_2 is a $d_c \times d_c$ matrix used to transform the dimensions of character vectors and word vectors to make them consistent, and b_1 and b_2 are bias parameters.

$$v_{ij}^w = W_2\big(\tanh(W_1 x_{ij}^w + b_1)\big) + b_2 \tag{3}$$

The set dimension for the word vector set of all matching words corresponding to the i-th letter is $m \times d_c$, where m is the size of the matching word set. This word vector set is represented as $X_i = (x_{i1}^w, \ldots, x_{im}^w)$. The word vector for the i-th character's j-th matched word is represented by x_{ij}^w. The character-level vector corresponding to the i-th character is concatenated with the word vector information of the associated word set and transmitted through a gating network, taking into account that the current character has different focuses on each matched word. Following character-word gating fusion, the gating network's output is weighted to produce a representation that is concatenated with both the original word vector representation and the character vector representation. Finally, a fully connected layer is used to obtain the final representation vector after character-word fusion. The gating network weight calculation formula for each word vector in the matching word set and the current character vector is shown in (4), where W_{gate} is the fully linked layer's weight matrix in the gating network. After obtaining the weights, a weighted sum can be performed as shown in (5), that is, calculating the weighted sum of the word and character vectors using the gating network weights.

$$g_i = \text{Sigmoid}(W_{\text{gate}}\,[\,h_i^c,\, x_i^w\,]) \tag{4}$$

$$z_i^w = \sum_{j=1}^{m}\big(g_i\, x_i^w + (1 - g_i)\, h_i^c\big) \tag{5}$$

Finally, by concatenating the word-granularity information z_i^w obtained through weighted sum in its gating form with the original character vector and word vector, the representation vector after character-word fusion can be obtained as shown in (6):

$$\hat{h}_i = [\,h_i^c,\, x_i^w,\, z_i^w\,] \tag{6}$$

2.5. **Lexicon-Enhanced BERT Model.** Methods of some current researches that incorporate word-granularity information mostly perform interaction and fusion of characters and words at the embedding layer outside the model. However, this method cannot fully integrate character and word information.

The model adopted in this paper is a mode of character-word interaction within the model, adding the operation of word vector and character vector interaction after the first Transformer layer inside BERT. This approach better injects external lexicon knowledge into the model, allowing the information of both granularities to be fully integrated with the training of different layers of the model. The approach of integrating word information at the embedding layer outside the model is depicted in the left half of Figure 5, whereas the method used in this research is displayed in the right half.

Given a Chinese sentence sequence $s_c = \{c_1, c_2, c_3, \cdots, c_n\}$ containing n characters, the corresponding character-word pair sequence $s_{cw} = \{(c_1, ws_1), (c_2, ws_2), \cdots, (c_n, ws_n)\}$ is first constructed according to the above method. Then, the sequence $\{c_1, c_2, \cdots, c_n\}$ is sent to the pre-processing embedding layer of the BERT model input. Token vector, sentence pair vector, and position vector are added, and the result is the input representation $E = \{e_1, e_2, \cdots, e_n\}$. The vector E is then entered into the BERT model's Transformer

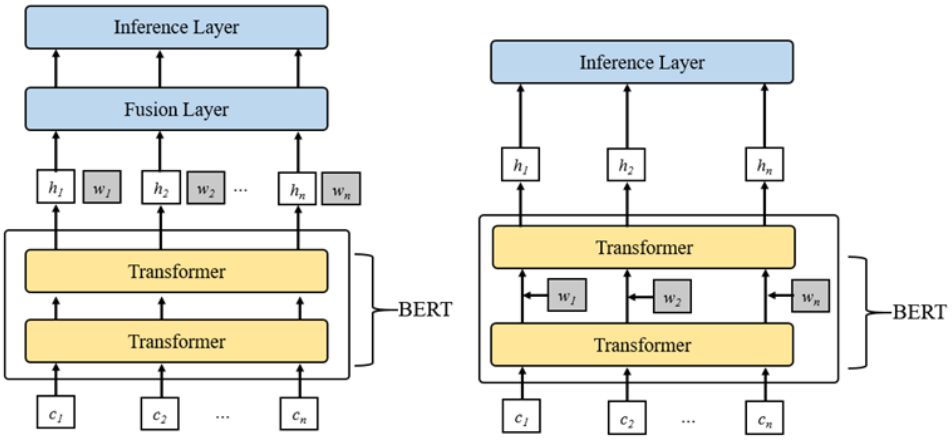

FIGURE 5. Comparison between external and internal model fusion

encoding layer. The Transformer encoder serves as the foundation for the BERT model's structure, with each Transformer layer's function illustrated in (7) and (8):

$$G = \text{LayerNorm}\big(H^{(L-1)} + \text{MHAttn}(H^{(L-1)})\big) \tag{7}$$

$$H^L = \text{LayerNorm}\big(G + \text{FFN}(G)\big) \tag{8}$$

Among them, the output of the L-th layer of the BERT model is represented by $H^L = \{h_1^l, h_2^l, \ldots, h_n^l\}$. *LayerNorm* is layer normalization processing, *MHAttn* is the multi-head attention mechanism, and *FFN* is a two-layer feedforward fully connected network with ReLU as the activation function.

To inject word-granularity information between the k-th and $(k+1)$-th Transformer layers inside the model, the model first obtains the output $H^c = \{h_1^c, h_2^c, \ldots, h_n^c\}$ after k consecutive Transformer layers. Then, each character-word pair (h_i^c, x_i^w) is processed using the character-word deep interaction module, denoted here as LA. As shown in (9). Following this, the fusion vector containing both character and word granularities is computed based on the attention mechanism.

$$\hat{h}_i = \text{LA}(h_i^c, x_i^w) \tag{9}$$

There are a total of L Transformer layers in the structure of the Chinese BERT model. After integrating the information of both character and word granularities, $\hat{H} = \{\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_n\}$ is fed into the next Transformer layer for continued learning. Finally, the output H from the model's final layer, the L-th Transformer layer, this semantic vector can be combined with softmax layer or Conditional Random Field decoding to obtain the predicted result of the sequence.

2.6. **Sequence Decoding Layer.** CRF is the named entity recognition model's top-level module. It accomplishes sequence decoding by calculating the probability values of potential sequences and choosing the path with the highest probability for output. The score of potential labels at each point in the sequence is the deep model's output. For instance, following a deep model's encoding of the input sequence $s_c = \{c_1, c_2, \ldots, c_n\}$, an output set reflecting the scores corresponding to the relevant candidate labels is obtained. The conditional random field layer receives this output after that for decoding. Figure 6 depicts the process of CRF decoding.

3. Results and Discussion.

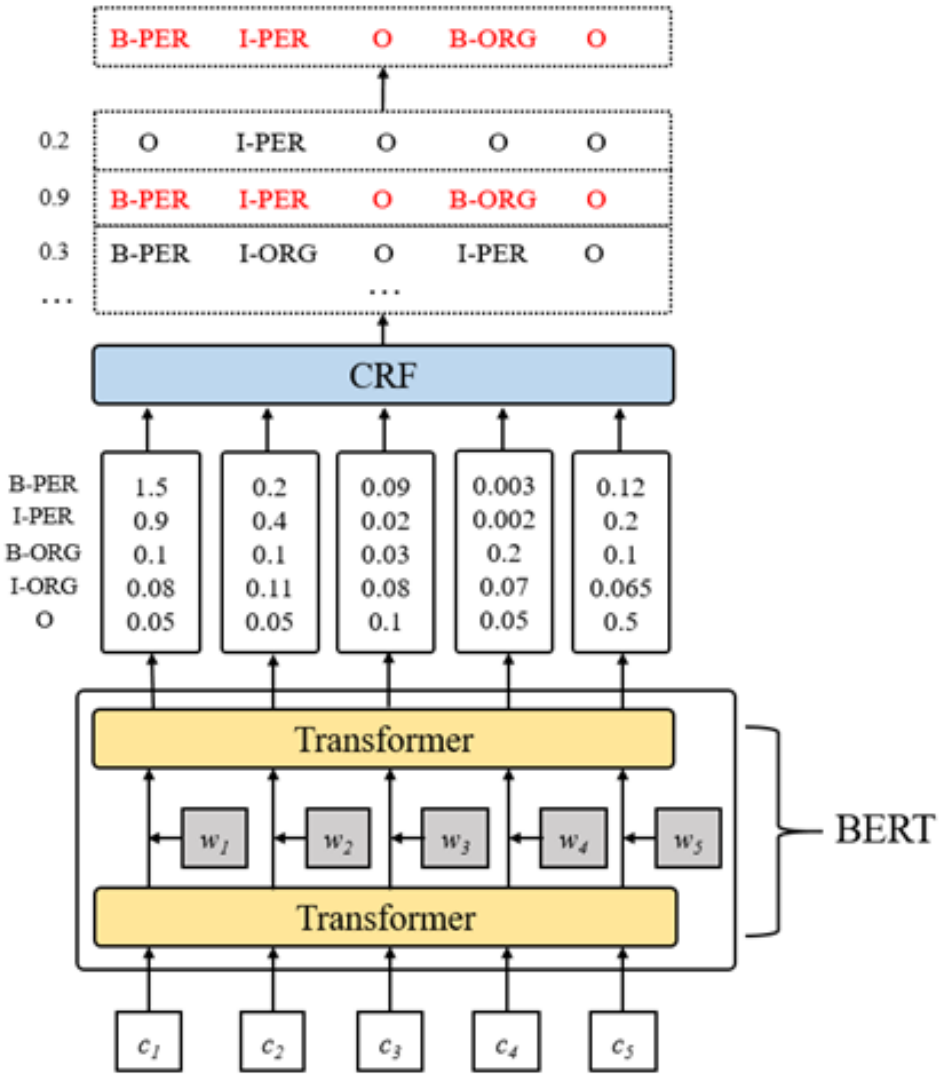

FIGURE 6. CRF sequence decoding

TABLE 1. Data set statistics table

Dataset	Type	Train	Dev	Test
Note4	Sentence	15.7k	4.3k	4.3k
	Character	491.9k	200.5k	208.1k
MSRA	Sentence	46.4k	–	4.4k
	Character	2169.9k	–	172.6k
Weibo	Sentence	1.4k	0.27k	0.27k
	Character	73.8k	14.5k	14.8k
Resume	Sentence	3.8k	0.46k	0.48k
	Character	124.k	13.9k	15.1k

3.1. **Data Sets.** In this paper, four Chinese Named Entity Recognition (NER) datasets are selected for model performance evaluation, namely Note4, MSRA, Weibo and Resume datasets. Each dataset is divided according to the training set, validation set and test set, and the experimental results are reported based on the performance of the test set, and the statistical information of the specific datasets is detailed in Table 1.

Table 2. Experimental parameters

Content	Parameter name	Value
Input length	Maximum length of sequence	128
	Batch size	64/32/4
Pre-Trained word vector	Lexicon size	8,824,330
	Word Vector Dimension	200
BERT-based	Network layers	12
	Hidden layer dimension	768
	Number of Attention heads	12
	Parameter size	110M
Train	Epoch	20
	Learning rate	1e-5
	Dropout	0.1
	Optimizer	AdamW
	Random Seed	2021

The purpose of the Note4 dataset is to identify four named entity types, namely People (PER), Places (LOC), Animals (ANI) and Plants (PLT), with data from Baidu Encyclopedia and Wikipedia, using the BIO annotation method and balancing the high-frequency and low-frequency entities by category sampling. The MSRA dataset, which requires the identification of three different categories of entities—Person (PER), Location (LOC), and Organization (ORG)—is a frequently used dataset for Chinese named entity recognition tasks. Two thousand Weibo posts that have been filtered from the Chinese social media site Sina Weibo make up the Weibo dataset. These postings include four different categories of entities: Person (PER), Location (LOC), Organization (ORG), and Geopolitical Entity (GPE). Resume dataset is derived from resume summaries of over a thousand senior managers on financial websites, mainly including eight types of entities: Person (PER), Country (CONT), Education (EDU), Location (LOC), Title (TITLE), Race (RACE), Organization (ORG), and Profession (PRO).

3.2. **Experimental Environment and Parameters.** Our experiment's operating system is Linux, which is paired with an Intel Core i7 processor and a GeForce RTX 3090 GPU. As indicated in Table 2. The framework used is Pytorch. The model is built on the Chinese version of Google's open-source BERT-Base pre-trained model, which includes 12 Transformer layers. For external word vectors, Tencent AI Lab's open-source and thoroughly pre-trained Chinese word vectors were selected. These word vectors were trained using a directed skip-gram algorithm on massive amounts of news and web texts, with a dimension of 200 and a size of 16GB, covering more than eight million Chinese words. Regarding to the model settings, the character-word deep interaction module designed in this paper is applied between the first and second Transformer encoder layers of the BERT model, and both BERT parameters and external word vectors are optimized during training. In terms of experimental parameters, the maximum length of the Chinese text sequence is set to 128, the batch size for training on the Note4 Chinese business dataset is set to 64, the MSRA dataset is set to 32, and the other two datasets are set to 4.

3.3. **Evaluation.** The precision, recall, and F1 score of the model on the dataset are the main assessment criteria for this challenge. Compared to a simple average, the F1 score provides a more thorough assessment of a model's performance. It is calculated as the harmonic mean of precision and recall. The formulas for these three indicators are as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% \tag{10}$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% \tag{11}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \tag{12}$$

Specifically, the counting rules for sample results are based on characters; a character is considered a true positive (TP) if its predicted entity position and category are correct, and a false positive (FP) if not. When a character is accurately predicted to be 0 and does not belong to an entity category, it is considered a true negative (TN); when it is mistakenly classified as an entity, it is considered a false negative (FN). F1 is split into two calculating approaches for multi-classification tasks: Micro-F1 and Macro-F1. Macro-F1 computes the average over all entity kinds after calculating precision and recall for each entity type independently. As seen by the following formulas, Micro-F1 on the other hand considers all entity kinds in order to determine the total precision, recall, and F1 score:

$$\text{Precision}_{\text{Micro}} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \tag{13}$$

$$\text{Recall}_{\text{Micro}} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \tag{14}$$

$$\text{Micro-}F_1 = \frac{2 \times \text{Precision}_{\text{Micro}} \times \text{Recall}_{\text{Micro}}}{\text{Precision}_{\text{Micro}} + \text{Recall}_{\text{Micro}}} \tag{15}$$

Micro-F1 better reflects the overall model effectiveness when there is an imbalance in entity categories. Thus, the assessment metric in the named entity recognition experiments reported in this research is the Micro-F1 score.

3.4. **Validation of model validity.** The performance of the lexical knowledge-enhanced model described in this research was assessed by contrasting it with named entity model tests conducted on the Note4, MSRA, Weibo, and Resume datasets. Table 3 displays the experimental outcomes, with the values representing the Micro-F1 score. The pre-trained model described in this research greatly increases performance on Chinese named entity recognition tasks because to its strong semantic encoding capabilities, according to the experimental data, when compared to the existing named entity recognition models. The efficiency of the suggested model is confirmed by the model's performance on the four data sets, which outperforms the top three deep models.

As shown in Table 3, named entity recognition models that incorporate Chinese word granularity information achieve better performance than other single character level models. Using the BERT+Word model, for instance, its performance on the Note4, MSRA, and Weibo datasets, as measured by Micro-F1, is 0.42%, 0.17%, and 0.77% higher than the BERT model, respectively. Additionally, it is seen that the BERT+Word model and the ERNIE model, which is developed from BERT, outperform all other models. This indicates that, for Chinese named entity recognition tasks, it makes sense to combine the BERT pre-training model with word-level data specific to Chinese entities. Furthermore, the model in this paper integrates word information within the model, compared to the BERT+Word model that interacts of character and word at the Embedding level. This leads to performance improvements by 0.61%, 0.58%, 2.03%, and 0.61% in Micro-F1 values, respectively, on the Note4, MSRA, Weibo, and Resume datasets. This improves

Table 3. Model implementation comparison

Model	Note4	MSRA	Weibo	Resume
BiLSTM+CRF	86.16	91.87	56.75	94.41
Lattice LSTM[20]	86.28	93.18	59.92	94.46
LR-CNN	85.65	93.23	60.15	93.95
BERT	91.01	94.95	67.55	95.86
BERT+Word	91.43	95.12	68.32	95.46
ERNIE[24]	92.01	95.08	67.96	94.82
ZEN[26]	89.79	95.29	66.73	95.41
FLAT[23]	80.56	95.46	68.07	95.78
Ma et al.[22]	81.34	95.35	69.11	95.54
Lexicon Knowledge Enhancement	92.04	95.70	70.35	96.07

Table 4. Experiment of interaction scheme

Model	Span F1/%		Type Acc/%	
	Note4	MSRA	Note4	MSRA
BERT	91.68	96.07	93.16	97.29
BERT+Word	93.38	96.33	94.24	97.45
Lexicon Knowledge Enhancement	94.16	96.58	94.84	97.52

the overall model's comprehension of Chinese semantics by explicitly representing word granularity information at the model level.

3.5. **Comparison of Model-level interaction and Embedding-level interaction.** In contrast to models that carry out interaction between character and word vectors at the external Embedding layer, the method in this paper is a form of lexicon enhancement at the model level. It achieves deep interaction between characters and words by designing a unique interaction module that directly injects external word vectors into the model. BERT+Word is a benchmark method that performs character-word interaction outside the model, mainly by concatenating the output part of the BERT model with word vectors, and then feeding them into LSTM and CRF for further integration and inference.

In Table 4, on all four datasets, the model in this research performs better than the BERT+Word model. To further verify how model-level character-word interaction improves performance in entity recognition tasks, this section sets up two new evaluation metrics Span F1 and Type Accuracy (Type Acc) for comparison. The experimental results are shown in Table 4. Span F1 represents the correctness of the entity span in NER, while Type Acc calculates the consistency in calculate form and accuracy, indicating the proportion of entities with both span and type correctly predicted among all predicted entities. Both BERT+Word and the model in this paper have higher Span F1 and Type Acc values on Note4 and MSRA datasets than the original BERT model at the single character level, further proving that the model's capacity to identify and categorize entity boundaries is enhanced by the inclusion of Chinese word granularity information. Compared with the BERT+Word model, the model in this paper improves the Span F1 and Type Acc values on the Note4 and MSRA datasets by 0.78%, 0.25%, 0.6%, and 0.07%, respectively. The experimental results once again validate that compared to models that perform character-word interaction at the Embedding layer outside the model,

Table 5. Exploration of interaction location

Interaction type	Location	F1/%
Single interaction	1	92.04
	3	91.73
	6	91.34
	9	91.11
	12	90.67
Multiple interactions	1,3	89.54
	1,3,6	88.53
	1,3,6,9	88.28
Interactions between each layer	All	86.23

injecting word-level knowledge into the model to achieve deep interaction between word-level and character-level features allows the model to more fully learn Chinese semantic information.

3.6. **Exploration of the Position for Character-Word Interaction Module.** This section applies the character-word interaction module between different Transformer layers of the model and conducts experiments using the Note4 dataset. Table 5 displays the outcomes of the experiment. The experiments set up various schemes, including performing a single instance of character-word interaction between some two layers within the BERT model, multiple instances of character-word interactions between different layers, and character-word interactions between each Transformer layer.

It can be observed from Table 5 that enhancing the model with lexicon knowledge at the lower layers of BERT, incorporating word granularity information to facilitate interaction with other layers of BERT, allows for a more comprehensive learning of vector semantic representations. Conversely, integrating character-word fusion interactions after higher Transformer layers in BERT results in a shallower level of interaction between semantic vectors at earlier layers. Moreover, performing character-word interactions after all Transformer layers in BERT leads to severe overfitting, substantially degrading model performance, with an F1 score of only 86.23% on Note4 dataset. Therefore, adding a character-word interaction module after only the first Transformer layer of BERT yields the best results, with an F1 value of 92.04%.

3.7. **Validation of the Pre-Trained Model Size's Effect.** In the previous sets of experiments, a fixed 12 Transformer layers was used in the model. However, large-scale pre-trained models consume massive resources and have slow inference speeds, making the choice of model size particularly important. This section of experiments attempts to use fewer Transformer layers, specifically 12, 9, 6, and 3 layers, to verify the impact of different numbers of model layers. Additionally, considering the real-time requirements in practical applications, the experiments compare the inference time of the models on the test set, with the outcomes displayed in Table 6.

Table 6. Model size experiment

Layers	Accuracy/%	Recall/%	F1/%	GPU inference time (ms/bar)
12	91.75	92.37	92.04	8.81
9	90.58	91.11	90.85	4.19
6	89.45	89.67	89.55	2.21
3	87.26	85.94	86.59	1.62

As can be seen from Table 6, lightweight BERT models, with fewer parameters and simpler model structures, have faster inference speeds. Compared to the original BERT model, they are quicker to deploy and operate and need less resources in real applications, but the model's performance also suffers as a result.

4. **Conclusion.** In this paper, a Chinese named entity recognition model based on lexical knowledge enhancement is proposed with the goal of improving the performance of Chinese named entity recognition (NER) model. The model injects character granularity and word granularity information into the underlying Transformer layer of the BERT model through the attention mechanism, and realizes the deep interaction and fusion between character and word inside the model, providing a new solution to the character-word fusion problem in Chinese named entity recognition. Compared with existing methods, the model in this paper helps to realize deeper knowledge fusion inside BERT, significantly enhances the model's ability to understand Chinese semantics, and achieves significant performance improvement in Chinese named entity recognition tasks. The Micro-F1 values on Note4, MSRA, Weibo and Resume datasets reach 92.04%, 95.70%, 70.35% and 96.07%, respectively.

However, the introduction of word granularity information has led to a significant increase in the number of model parameters, which in turn raises the demand for computing power. Therefore, how to perform model lightweight compression while ensuring that the accuracy loss remains within an acceptable range will be a key direction for future research.

Acknowledgment. This research was partially supported by the Natural science research project of Anhui Xinhua University (No.2024zr001), the Anhui Province Teaching Demonstration Course (No. 2020SJJXSFK1294).

REFERENCES

[1] G. A. Goldstein, et al., "Shared computational principles for language processing in humans and deep language models," *Nature Neuroscience*, vol. 25, no. 3, pp. 369–380, 2022.

[2] K. Humphreys, et al., "Description of the LaSIE-II system as used for MUC-7," in *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, vol. 4, 1998.

[3] D. E. Appelt, et al., "SRI International FASTUS System MUC-6 Test Results and Analysis," in *Proceedings of the 6th Conference on Message Understanding, MUC 1995, Columbia, Maryland, USA, November 6-8, 1995*, Association for Computational Linguistics, 1995.

[4] R. G. Morgan, et al., "University of Durham: Description of the LOLITA system as used in MUC-6," in *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

[5] Y. Y. Liu, et al., "Chinese named entity recognition method based on machine reading comprehension," *Pattern Recognition and Artificial Intelligence*, vol. 33, no. 7, pp. 653–659, 2020.

[6] H. Isozaki and H. Kazawa, "Efficient Support Vector Classifiers for Named Entity Recognition," in *International Conference on Computational Linguistics*, 2002, pp. 1–7.

[7] G. Zhou and J. Su, "Named entity recognition using an HMM-based chunk tagger," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, July 6-12, 2002, Philadelphia, PA, USA, 2002.

[8] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of ICML*, vol. 1, no. 2, p. 3, 2002.

[9] A. McCallum and W. Li, "Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons," *Association for Computational Linguistics*, vol. 4, pp. 188–191, 2003.

[10] M. A. Meng-Cheng, A. Hamdulla, and T. Tohti, "Chinese Named Entity Recognition Based on Conditional Random Fields Multi-Feature Fusion," *Modern Computer*, 2019.

[11] Y. U. Hong-Kui, et al., "Chinese named entity identification using cascaded hidden Markov model," *Journal on Communications*, vol. 27, no. 2, pp. 87–94, 2006.

[12] P. Pu, et al., "Transformer Optimization and Application in Named Entity Recognition of Apple Diseases and Pests," *Transactions of the Chinese Society for Agricultural Machinery*, vol. 54, no. 6, pp. 264–271, 2023.
[13] P. Zhao, et al., "Recognition of the agricultural named entities with multifeature fusion based on ALBERT," *IEEE Access*, vol. 10, pp. 98936–98943, 2022.
[14] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv*, 2016.
[15] R. Collobert, et al., "Natural Language Processing (almost) from Scratch," *Journal of Machine Learning Research*, vol. 12, no. 1, pp. 2493–2537, 2011.
[16] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.
[17] M. D. Zeiler, et al., "On rectified linear units for speech processing," in *IEEE International Conference on Acoustics*, IEEE, 2013.
[18] J. Devlin, et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019)*, vol. 6. Long and short papers, Minneapolis, Minnesota, USA, 2019.
[19] A. Vaswani, et al., "Attention Is All You Need," *arXiv*, 2017.
[20] Y. Zhang and J. Yang, "Chinese NER Using Lattice LSTM," *arXiv*, 2018.
[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[22] M. Peng, et al., "Simplify the Usage of Lexicon in Chinese NER," *arXiv*, 2019.
[23] X. Li, et al., "FLAT: Chinese NER Using Flat-Lattice Transformer," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.
[24] Y. Sun, et al., "ERNIE: Enhanced Representation through Knowledge Integration," *arXiv*, 2019.
[25] C. Jia, et al., "Entity Enhanced BERT Pre-training for Chinese NER," in *Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2020.
[26] S. Diao, et al., "ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations," in *Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2020.
[27] Y. Xiao, et al., "Chinese NER Using Multi-View Transformer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
[28] L. Zhang, et al., "Enhanced Chinese named entity recognition with multi-granularity BERT adapter and efficient global pointer," *Complex & Intelligent Systems*, vol. 10, no. 3, pp. 4473–4491, 2024.
[29] X. Liu, et al., "Sequential lexicon enhanced bidirectional encoder representations from transformers: Chinese named entity recognition using sequential lexicon enhanced BERT," *PeerJ Computer Science*, no. 10, pp. e2344, 2024.