Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.063008

ARTICLE

Chinese Named Entity Recognition Method for Musk Deer Domain Based on Cross-Attention Enhanced Lexicon Features

Yumei Hao[1,2], **Haiyan Wang**[1,2,*] **and Dong Zhang**[3]

[1]School of Information Science and Technology, Beijing Forestry University, Beijing, 100083, China
[2]Engineering Research Center for Forestry-Oriented Intelligent Information Processing, National Forestry and Grassland Administration, Beijing, 100083, China
[3]School of Ecology and Nature Conservation, Beijing Forestry University, Beijing, 100083, China

*Corresponding Author: Haiyan Wang. Email: wanghaiyan2008@bjfu.edu.cn

Received: 01 January 2025; Accepted: 14 February 2025; Published: 16 April 2025

ABSTRACT: Named entity recognition (NER) in musk deer domain is the extraction of specific types of entities from unstructured texts, constituting a fundamental component of the knowledge graph, Q&A system, and text summarization system of musk deer domain. Due to limited annotated data, diverse entity types, and the ambiguity of Chinese word boundaries in musk deer domain NER, we present a novel NER model, CAELF-GP, which is based on cross-attention mechanism enhanced lexical features (CAELF). Specifically, we employ BERT as a character encoder and advocate the integration of external lexical information at the character representation layer. In the feature fusion module, instead of indiscriminately merging external dictionary information, we innovatively adopted a feature fusion method based on a cross-attention mechanism, which guides the model to focus on important lexical information by calculating the correlation between each character and its corresponding word sets. This module enhances the model's semantic representation ability and entity boundary recognition capability. Ultimately, we introduce the decoding module of GlobalPointer (GP) for entity type recognition, capable of identifying both nested and non-nested entities. Since there is currently no publicly available dataset for the musk deer domain, we built a named entity recognition dataset for this domain by collecting relevant literature and working under the guidance of domain experts. The dataset facilitates the training and validation of the model and provides data foundation for subsequent related research. The model undergoes experimentation on two public datasets and the dataset of musk deer domain. The results show that it is superior to the baseline models, offering a promising technical avenue for the intelligent recognition of named entities in the musk deer domain.

KEYWORDS: Named entity recognition; musk deer; cross-attention; lexicon enhancement

1 Introduction

Musk deer, belonging to the family Muscidae within the order Artiodactyla, exhibits notable scientific, ecological, and medical significance. Chinese literature dedicated to musk deer encompasses a substantial reservoir of valuable information and extensive knowledge. The extraction of domain-specific information from these textual sources serves as crucial data support for various research endeavors, including the conservation and development of musk deer resources. Consequently, exploring effective methodologies for data mining and information extraction from these unstructured texts emerges as a significant and worthwhile pursuit. With the expanding digitized scale of data associated with the musk deer domain, manual extraction of pertinent information becomes increasingly difficult. Addressing this challenge, the deep learning-based NER methods present a viable solution, enabling the automatic extraction of relevant knowledge entities from vast amounts of unstructured texts. This approach establishes a foundational framework for subsequent endeavors such as the construction of the musk deer domain knowledge graph, the musk deer domain Q&A system, and the musk deer domain text summarization system.

NER refers to the process of locating and classifying entities from unstructured text based on predefined entity categories. For example, in the medical field, entity categories are classified into diseases, symptoms, body parts, and so on. NER task has demonstrated commendable performance in general domains, such as the news domain. However, in specific domains, the scarcity of annotated data arises due to the requisite expertise of annotators in the respective professional domain, coupled with the labor-intensive nature of the task. Furthermore, the presence of proprietary vocabularies in these specific domains poses a challenge to the semantic expressiveness of NER models. In contrast to general domain NER tasks, those specific to domains exhibit greater application value and encounter more formidable challenges, prompting researchers to delve into domain-specific NER studies. Wang et al. [1] proposed a NER model that integrates multiple features, including characters, radicals, word boundaries, and parts of speech, by analyzing the characteristics of texts related to forest diseases. Yu et al. [2] proposed a mineral NER model based on the pre-trained language model BERT for the problem of NER in mineral literature. Additionally, Zhang et al. [3] devised a NER method for challenges in the Chinese financial domain, addressing issues like long entity names, ambiguous boundaries, and diverse forms of expression. Compared to English, the nature of inconspicuous Chinese word boundaries makes it difficult to recognize entity boundaries, which is another difficulty in Chinese domain-specific NER tasks. Current works mainly adopts a lexicon-based approach to solve the problem of inconspicuous Chinese word boundaries. Ma et al. [4] proposed a method for simple fusion of vocabulary information on the model embedding layer. Specifically, for each character, all the vocabularies with the character as the start character, the middle character, the end character, and the separate character are obtained in turn, and the sets of these vocabularies are encoded, and finally the vocabulary vectors and the character vectors are concatenated. This approach, while leveraging lexical information, demonstrates some performance improvement. However, it fails to account for the varying importance of different lexical sets in character feature representation. Building upon this, Zhong et al. [5] enhanced the method by introducing a lexical set-level attention mechanism, aiming to acknowledge the significance of different lexical sets. Nevertheless, this improved method overlooks the correlation between lexical sets and their corresponding characters when computing the importance of lexical sets.

In this study, musk deer domain texts exhibit a greater diversity of entity categories, including diseases, genes, tissue sites, etc., and demonstrate entity nesting. In contrast to the general domain, these texts span various research fields, thereby diversifying their textual characteristics. Moreover, there is an absence of a publicly available NER dataset tailored to musk deer domain. Addressing the challenges encountered in the NER task in the musk deer domain, we introduce a model aimed at augmenting lexical features through the utilization of the cross-attention mechanism. Inspired by the work of Chen et al. [6], which employed the attention mechanism for extracting multi-scale features in image classification, we integrate lexical information at the character representation layer and incorporate the cross-attention mechanism during the fusion process. Different from the approach proposed by Ma et al. [4], which directly concatenates lexical information at the character representation layer, our model dynamically assigns weights to the lexical features of each character. This is achieved by calculating the relevance magnitude of each character in the input sequence with the corresponding four lexical sets. The objective is to precisely guide the model towards focusing on more crucial lexical information, thereby enhancing its ability to represent semantic information and identify Chinese word boundaries. The primary contributions of this paper are outlined as follows:

(1) We propose the CAELF-GP model based on a cross-attention enhanced lexical features. The BERT pre-trained language model is used as the base encoder. Particularly, we introduce a lexical feature fusion layer based on the cross-attention mechanism after the character representation layer, which guides the model to focus on important lexical features, thereby improving the model's semantic representation ability and entity boundary recognition capability. The decoding module of GlobalPointer [7] is employed to predict entity types, addressing the recognition of both nested and non-nested entities simultaneously.

(2) For the lack of relevant datasets in the field of musk deer, we construct the corresponding field dataset and lexicon by collecting relevant literature, cleaning and manually labelling the data.

(3) Experimental evaluations on two public datasets and the literature dataset of the musk deer domain are conducted with the CAELF-GP model and the baseline models. Results indicate that our model outperforms the baselines on the NER task in musk deer domain.

The remainder of this paper is structured as follows: Section 2 provides an overview of related work. Section 3 describes the architecture of the proposed model. Section 4 details the dataset, experimental setup, and analyzes the experimental results. Finally, Section 5 presents the conclusions and directions for future work.

2 Related Works

NER plays a crucial role in information extraction tasks. Early approaches to NER involved rule-based and dictionary-based methods, along with traditional machine learning approaches. The advent of deep learning has led to outstanding performance in NER tasks with deep learning approaches. Researchers have explored various frameworks to tackle NER. Huang et al. [8], Chiu et al. [9], and Luo et al. [10] consider NER as a sequence labeling task. Li et al. [11] redefine NER as a machine reading comprehension problem, effectively addressing the recognition of nested and non-nested entities. However, this approach necessitates repeated text encoding, resulting in increased computational complexity. Su et al. [7] introduced the GlobalPointer model, which is a span-based approach. It incorporates a multiplicative attention mechanism and leverages relative position information to predict entities. The model globally assesses the start and end positions of candidate entities, enabling the discrimination-free identification of both nested and non-nested entities. Chen et al. [12] formulated NER as a sequence-to-sequence generation, yielding favorable outcomes in cross-domain tasks. In this paper, for the problem of entity nesting in texts of musk deer domain, we select the GlobalPointer model as the backbone model, and further explore the solution for the problems such as difficult identification of entity boundaries and the presence of lots of proprietary vocabularies.

Research indicates that, due to errors in Chinese word segmentation, character-based NER systems are usually superior to word-based methods [13]. However, character-based NER systems have limited semantic information and fail to leverage lexical information, which is crucial for entity boundary recognition. Numerous studies concentrate on the effective integration of lexical information into character-based NER systems. These studies can be categorized into two primary domains: dynamic architecture-based methodologies and adaptive embedding-based methodologies.

2.1 Dynamic Architecture-Based Methodologies

Zhang et al. [14] proposed the Lattice LSTM model, the earliest model to apply lexical enhancement methods in Chinese NER tasks. This model introduces a word cell structure based on the LSTM, incorporating all lexical information for characters ending in that character. While this model achieved effective performance improvement in NER by integrating lexical information, it suffers from drawbacks such as low computational performance, information loss, and poor portability. These issues stem from its use of RNN structure, inability to parallelize calculations, lack of utilization of lexical information with each character as an intermediate character, and limited adaptability to LSTM networks. Gui et al. [15] suggested using CNN to encode character features, allowing for parallelized computation, and employed attention mechanisms to integrate lexical information. To address the issue of lexical information loss, Sui et al. [16] constructed a collaborative graph network between characters and words, utilizing a Graph Attention Network (GAN) for feature extraction. Gui et al. [17] proposed a neural network model based on a dictionary. As RNN and CNN structures are unable to capture long-distance dependencies in sentences, Li et al. [18] introduced the Flat-Lattice Transformer model. This model excels at capturing long-distance dependencies through its fully connected self-attention mechanism. It integrates lattice structures by assigning head and tail position encodings to each character and word. Liu et al. [19] considered that current approaches only fuse lexical information through a shallow random initialisation layer, and proposed to integrate external lexical knowledge directly into the BERT layer through a lexical adapter.

2.2 *Adaptive Embedding-Based Methodologies*

Besides the aforementioned method, which modifies the model framework to incorporate lexical information, certain studies suggest approaches that exclusively integrate lexical information at the embedding layer, showcasing enhanced transferability. For instance, Liu et al. [20] suggested encoding fixed-length representations of vocabulary information ending with each character to serve as the word vector for that character. The character's word vector is then concatenated with its character vector to form the final vector representation of the character. However, this method still lacks vocabulary information for each character acting as an intermediate character. Ma et al. [4] proposed a simple method to integrate lexical information at the model's embedding layer by sequentially obtaining all vocabulary sets for each character as the starting, middle, ending, and standalone character. These sets are then encoded and concatenated with the character encoding. This method maximizes the utilization of vocabulary information, prevents information loss, and experimental results affirm its efficacy in enhancing the performance of Chinese NER. Motivated by the aforementioned methodology, we employ the lexical augmentation strategy grounded in the foundational model of GlobalPointer. Given that the study conducted by Ma et al. [4] exclusively delved into the migratory implementation of their approach onto NER models based on the sequence-labeling architectures, we endeavor to extend this inquiry by examining its applicability to span-based NER models. Furthermore, an innovative contribution of this paper lies in the introduction of a cross-attention mechanism, aimed at guiding the model to utilize lexical information more effectively. This augmentation serves to enhance the model's semantic representation proficiency and its ability to discern word boundaries.

3 Methodology

The CAELF-GP model with enhanced lexical features based on cross-attention employs the span-based architecture, offering the advantage of recognizing both nested and non-nested entities. The model comprises distinct layers: the character representation layer, feature fusion layer, span representation layer, and label prediction layer, as illustrated in Fig. 1.

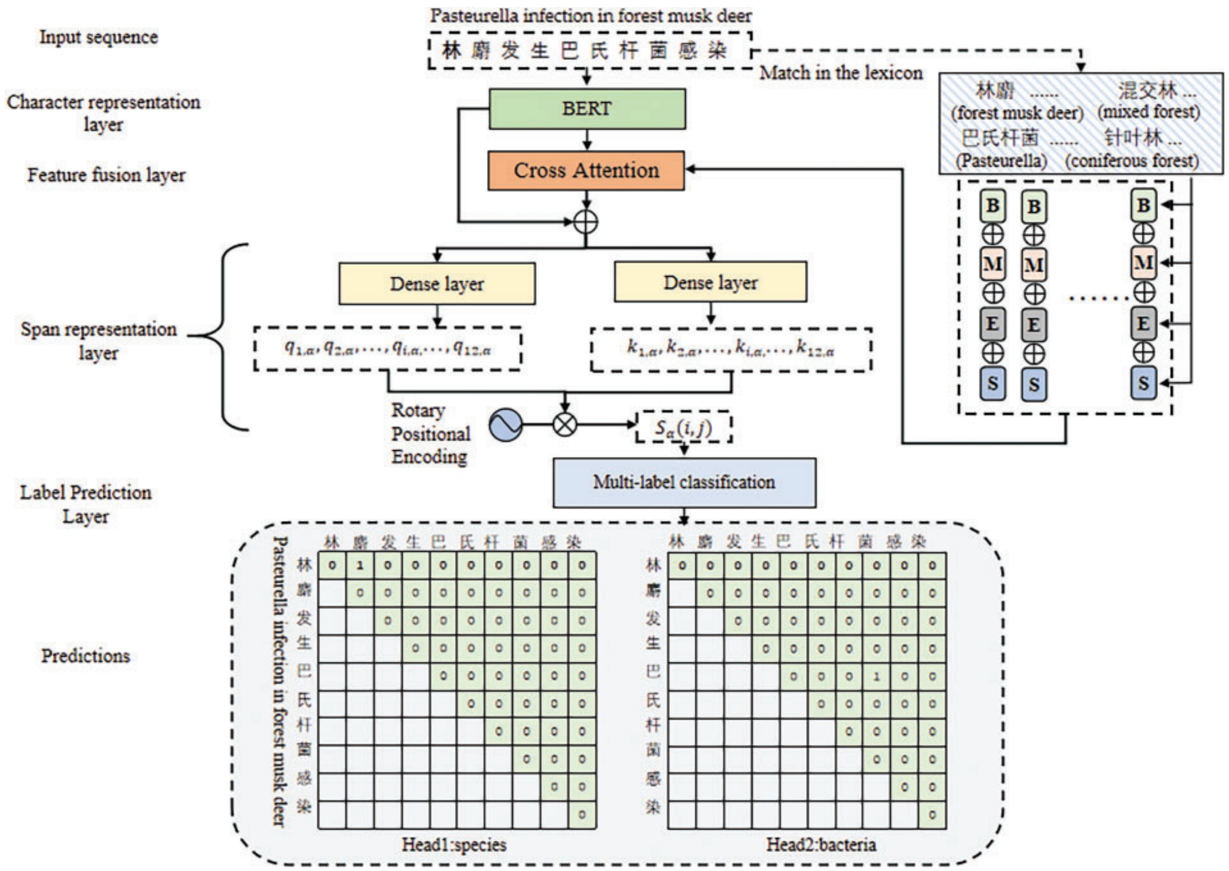

Input sequence
Pasteurella infection in forest musk deer
林麝发生巴氏杆菌感染
Match in the lexicon
Character representation layer
BERT
Feature fusion layer
Cross Attention
林麝 ...... (forest musk deer)
混交林 ... (mixed forest)
巴氏杆菌 ...... (Pasteurella)
针叶林 ... (coniferous forest)
B
M
E
S
Span representation layer
Dense layer
Dense layer
$q_{1,\alpha}, q_{2,\alpha}, \ldots, q_{i,\alpha}, \ldots, q_{12,\alpha}$
$k_{1,\alpha}, k_{2,\alpha}, \ldots, k_{i,\alpha}, \ldots, k_{12,\alpha}$
Rotary Positional Encoding
$S_\alpha(i,j)$
Label Prediction Layer
Multi-label classification
Predictions
Pasteurella infection in forest musk deer
林 麝 发 生 巴 氏 杆 菌 感 染
Head1:species
Head2:bacteria

Figure 1: The overall architecture of the CAELF-GP model

The character representation layer utilizes the BERT model to encode the input text sequence, yielding vector representations for each character. The feature fusion layer enhances lexical features by incorporating external information, addressing challenges in semantic representation and entity boundary recognition. Notably, we introduce an innovative cross-attention mechanism within the feature fusion method, directing the model's focus to crucial lexical information based on the correlation between input sequence characters and the sets of matching words. The span representation layer employs a multiplicative attention mechanism on vector representations of characters, integrating positional coding to score all candidate entities within each entity class. The label prediction layer adopts the decoding module from the GlobalPointer model, classifying candidate entities with multiple labels based on their scores. This decoding structure accommodates the identification of both nested and non-nested entities. The training process of the model is shown in Algorithm 1 below:

Algorithm 1: Pseudocode for the CAELF-GP model

```
Input: initial parameters θ, training data X, external lexicon L
Output: Optimized model parameters $\hat{\theta}$
Repeat
    Convert sentences to character vectors: $h_i \leftarrow PLM(c_i)$
    Obtain the four word vectors corresponding to each character according to Eqs. (2)–(7):
    $v_i^S(B), v_i^S(M), v_i^S(E), v_i^S(S)$
```

(Continued)

Algorithm 1 (continued)

```
According to Eqs. (8)–(12), obtain the character vector representation $x_i$ of the fused vocabulary features:
    $z_i \leftarrow CrossAttention(h_i, v_{i,j}^{s})$,
    $x_i \leftarrow Concat\,(h_i, z_i)$
According to Eqs. (13)–(15), calculate the score of span belonging to type $\alpha$ based on the head and tail characters of span $S_\alpha(i,j)$: $S_\alpha(i,j) \leftarrow {q_{i,\alpha}}^{T} R_{j-i} k_{j,\alpha}$
Calculate the loss: $\mathrm{loss} \leftarrow \log(1 + \sum_{(i,j)\in P_\alpha} e^{-s_\alpha(i,j)}) + \log(1 + \sum_{(i,j)\in Q_\alpha} e^{s_\alpha(i,j)})$
Update model parameters: $\hat{\theta} \leftarrow argmin_\theta(loss)$
Until n iterations are completed
```

3.1 Character Representation Layer

The CAELF-GP model uses BERT as the encoder. BERT's input consists of three parts: word embeddings, segment embeddings, and position embeddings. Word embeddings are typically at the level of words, subwords, or characters. The purpose of segment embeddings is to differentiate between pairs of sentences. Position embeddings are used to introduce the positional information of characters into the model, addressing the issue of Transformer models failing to capture the absolute positions of tokens in a sequence.

Specifically, for a sentence S composed of n characters ($S = \{c_1, c_2, \ldots c_n\}$), we input it into the pre-trained language model BERT to obtain the corresponding representations for each character:

$$h_1, h_2, \ldots h_n = PLM(c_1, c_2, \ldots c_n). \tag{1}$$

3.2 Feature Fusion Layer

3.2.1 Lexical Feature

For Chinese words, the starting character, middle character, ending character, and characters that independently form a word often contain rich semantic information and word boundary information. Moreover, as observed from the experimental results in Section 4.2, the model incorporating the "BMES" word sets achieved the best performance. Therefore, in the CAELF-GP model, the lexical features corresponding to each character c_i are derived from words where the character serves as the starting character, middle character, or ending character, as well as words formed solely by the character itself. These words form four corresponding word sets, B (c_i), M (c_i), E (c_i), and S (c_i), which can be represented using the following Eqs. (2)–(5). Here, $w_{i,k}$ represents a word in the input sequence S that starts with c_i, ends with c_k, and exists in the external dictionary L.

$$B(c_i) = \{w_{i,k}, \forall w_{i,k} \in L, i < k \le n\}, \tag{2}$$
$$M(c_i) = \{w_{j,k}, \forall w_{j,k} \in L, 1 \le j < i < k \le n\}, \tag{3}$$
$$E(c_i) = \{w_{j,i}, \forall w_{j,i} \in L, 1 \le j \le i\}, \tag{4}$$
$$S(c_i) = \{t_i, \exists t_i \in L\}. \tag{5}$$

After obtaining the "BMES" word sets corresponding to each character, these sets are compressed into vectors of fixed dimensions. During computation, the word frequency of each word in the corpus (in our experiments, the corpus refers to the data composed of the training and validation sets) is used as its weight. Additionally, weight normalization is applied to all words in the four sets for overall comparison. Taking the example of the word set B corresponding to the character c_i, its set vector $v_i^S(B)$ can be obtained using the following Eqs. (6) and (7). Here, $z(w)$ represents the word frequency of the word w in the corpus, and e^w $(\cdot)$ denotes the operation of converting the word into a word vector.

$$Z = \sum\nolimits_{w \in B \cup M \cup E \cup S} z(w), \tag{6}$$

$$v_i^S(B) = \frac{4}{Z} \sum_{w \in B} z(w)\, e^w(w). \tag{7}$$

3.2.2 Feature Fusion Based on Cross-Attention

Each character within the input sequence exhibits varying degrees of association with the information of its respective four lexical sets. To guide the model towards prioritizing essential lexical information, the incorporation of a cross-attention mechanism is proposed. This mechanism assigns distinct weights to lexical features, facilitating the derivation of a character vector representation that encompasses relevant lexical attributes.

To make it easier to understand the role of the cross-attention mechanism, we will illustrate it using its application in a question-answering task as an example. In a question-answering task, both the question and the relevant paragraph are input into the model, and the model needs to generate an answer based on the provided input. The role of the cross-attention mechanism is to allow the model, when generating the answer, to treat the question as a query and dynamically attend to different parts of the paragraph, thereby locating the most relevant content in the paragraph and generating the answer. In addition to its application in question-answering tasks, the cross-attention mechanism has also achieved outstanding results in many other tasks, including image processing [21,22] and multimodal learning [23,24]. Distinguished as an extension of the self-attention mechanism, the primary distinction lies in the source of input: the self-attention mechanism draws from the same sequence, whereas the cross-attention mechanism draws from disparate sequences. In this work, the input to the cross-attention mechanism primarily originates from the original input sequence and the corresponding external word sets, as shown in Fig. 2. The operational procedure involves generating a Query (Q) vector for each character vector h_i of the input sequence, utilizing the trainable parameter matrix W_i^Q. Additionally, for the four lexical set vectors corresponding to the character, Key (K) and Value (V) vectors are formulated using trainable parameter matrices W_i^K and W_i^V, respectively. Here, $W_i^Q \in \mathbb{R}^{d_k \times d_e}$, W_i^K, $W_i^V \in \mathbb{R}^{d_k \times d_{ge}}$, with d_k denoting the dimension of the K vector, d_e denoting the dimension of the character vectors output from BERT, and d_{ge} denoting the dimension of the word vector.

$$Q_i = W_i^Q \cdot h_i, \tag{8}$$

$$K_{i,j} = W_i^K \cdot v_{i,j}^s, \tag{9}$$

$$V_{i,j} = W_i^V \cdot v_{i,j}^s. \tag{10}$$

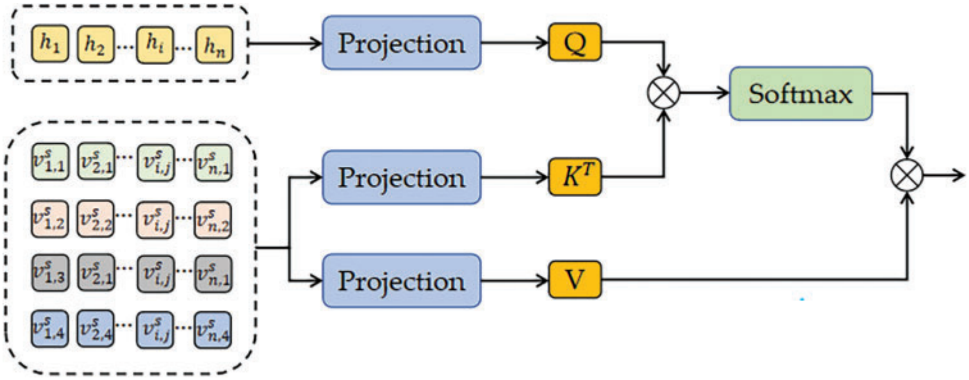


Figure 2: Cross-Attention architecture

By computing the scaled dot product between the Q vector of the current character and the K vectors associated with the respective word sets, attention scores for the current character across the four word sets are derived. Subsequently, attention weights are obtained through the application of the softmax function. Utilizing the acquired weights, the V vectors undergo weighting and summation to yield the character vector representation incorporating lexical features, denoted as z_i. Since there are four word sets defined in this work, specifically identified as "BMES", the value of n in the Eq. (11) is 4.

$$z_i = \sum_{j=1}^{n} softmax\left(\frac{Q_i \cdot K_{i,j}}{\sqrt{d_k}}\right) V_{i,j}. \tag{11}$$

Ultimately, in order to retain more information, the vector representation of each character by BERT and the character vector representation incorporating lexical features are concatenated as the final vector representation of the character:

$$x_i \leftarrow [h_i; z_i]. \tag{12}$$

3.3 Span Representation Layer

After obtaining the vector representations for all characters in the sentence, a span can be represented by its first and last characters. Upon acquiring the sentence representation X, it undergoes processing through fully-connected layers to derive character head and tail vector representations denoted as $q_{i,\alpha}$ and $k_{i,\alpha}$. These representations are instrumental in the identification of entities belonging to the specified type α.

$$q_{i,\alpha} = W_{q,\alpha} x_i + b_{q,\alpha}, \tag{13}$$

$$k_{i,\alpha} = W_{k,\alpha} x_i + b_{k,\alpha}. \tag{14}$$

The score $S_\alpha(i, j)$ for predicting span $s[i{:}\, j]$ as an entity of type α can be obtained by multiplying the vector representations of the first and last characters of span $s[i{:}\, j]$. The calculation of $S_\alpha(i, j)$ applies Rotary Positional Encoding (RoPE) [25], which explicitly injects relative positional information to better utilize word boundary information. Here, R is a transformation matrix, and it satisfies ${R_i}^T R_j = R_{j-i}$.

$$\begin{aligned} S_\alpha(i, j) &= (R_i q_{i,\alpha})^T (R_j k_{j,\alpha}), \\ &= {q_{i,\alpha}}^T {R_i}^T R_j k_{j,\alpha} \\ &= {q_{i,\alpha}}^T R_{j-i} k_{j,\alpha} \end{aligned} \tag{15}$$

3.4 Label Prediction Layer

In the task of identifying entities of type α from sentences with a sequence length of n, the selection of k entities is required from a pool of n (n + 1)/2 candidate entities. Thus, the task transforms into a multi-label classification task. We employ a loss function designed for multi-label classification tasks, and its calculation formula is as follows. During the decoding phase, all segments $s[i\colon j]$ satisfying $S_\alpha\,(i, j) > 0$ are considered as outputs of entities of type α.

$$\text{loss} = \log\left(1 + \sum\nolimits_{(i,j)\in P_\alpha} e^{-s_\alpha(i,j)}\right) + \log\left(1 + \sum\nolimits_{(i,j)\in Q_\alpha} e^{s_\alpha(i,j)}\right). \tag{16}$$

P_α denotes the set of all spans belonging to entities of type α, and Q_α denotes the set of all spans belonging to non-entities or entities not of type α.

4 Experiments

4.1 Dataset and Labeling Strategy

To assess the superiority of the CAELF-GP model, we conducted experiments using two public datasets and a self-constructed literature abstract NER dataset specific to the musk deer field (MDNER). Table 1 presents information for each dataset. Specifically, two public datasets originate from distinct domains: Weibo [4], gathered from the social media website Weibo, and CLUENER [26], obtained from the news field. As the test set of the CLUENER dataset lacks annotations and cannot be employed to evaluate the model's effectiveness, we combine the training and validation sets of CLUENER. Subsequently, we redistribute them in an 8:1:1 ratio, designating them as the training set, validation set, and test set.

Table 1: Statistics of datasets

Datasets	Train	Dev	Test	Sentence length	Number of entity types
CLUENER	9.6 k	1.2 k	1.2 k	37.38	10
Weibo	1.4 k	0.27 k	0.27 k	54.57	8
MDNER	1.7 k	0.22 k	0.22 k	53.09	11

MDNER datasets primarily derive from pertinent research literature accessible through the China National Knowledge Infrastructure (CNKI), encompassing journal articles, master's and doctoral theses. Due to the limited amount of Chinese literature on musk deer, and the reference value of closely related species for musk deer studies, we also gather relevant literature on these species (primarily red deer and sika deer) to augment the dataset. We search the literature on CNKI (https://www.cnki.net/ (accessed on 1 January 2025)) using the subject words "musk deer", "red deer" and "sika deer", and manually filter out irrelevant publications. We use publicly available literature published on the internet as our data source, which does not include any private data. Additionally, during the data collection process, we carefully review the data to ensure that it does not contain any sensitive information. Finally, a total of 1401 abstracts of documents are collected, with a total of 2165 sentences. In order to conduct NER task, these sentences are labeled with entities using Label-Studio (https://labelstud.io/ (accessed on 1 January 2025)). Given that our model is based on span as the basic unit for entity recognition, the annotation of the raw data needs to be marked with the name, category and location of entities present within the sentence. Based on the primary research themes in musk deer literature and the opinions of domain experts, we categorized entities into 11 types, encompassing nature reserves, regions, tissue parts, diseases, genes, cells, bodily elements, food, bacteria, and research methods. During the data annotation process, the presence of specialized vocabularies increased the difficulty of understanding the text. In order to accurately annotate the entities, we carry out the data annotation work under the guidance of domain experts.

4.2 Experiment Setup

The bert-base-Chinese pre-trained language model, released by Google, serves as the foundational encoder. For lexicons, experiments on two public datasets use the lexicon provided by [14], which are pre-trained on Chinese Giga-Word using word2vec [27]. This lexicon is characterized by word vector of dimension 50, encompassing 5.7 k single-character words, 291.5 k two-character words, 278.1 k three-character words, and 129.1 k other words. The experiments conducted on the MDNER dataset involved the utilization of word vectors derived from training word2vec models on a corpus of pertinent literature. In the training process, we use the Adam optimizer, and adopt the dropout technique to avoid overfitting. To determine the hyperparameters of the model, we employ a random search method. By randomly selecting parameter combinations and conducting multiple preliminary experiments, the optimal performance of the model along with the corresponding parameter configurations are recorded. Table 2 shows the optimal hyperparameters of the proposed model and the range of the random search for the parameters. The CAELF-GP model and the comparison model in this paper undergo training and testing on an NVIDIA RTX A4000 GPU.

Table 2: Optimal hyperparameter values and their search ranges

Hyperparameters	Optimal value	Random search range
Batch size	16 (CLUENER), 5 (others)	[5, 16, 32]
Learning rate	2×10^{-5}	[2×10^{-5}, 5×10^{-5}, 5×10^{-4}]
Dropout	0.3	[0.15, 0.2, 0.3]
Word vector dimension	100 (MDNER)	[50, 100]

In the evaluation of NER models, performance is primarily assessed by comparing the entities identified by the model to the ground-truth entities, which are manually annotated. In this study, we use widely accepted evaluation metrics: Precision (P), Recall (R), and F1 score (F1). Precision refers to the proportion of correctly predicted entities among all entities identified by the model, and reflects the accuracy of the prediction. Recall represents the proportion of correctly predicted entities among all ground-truth entities, and shows the sensitivity of the model. F1 is the harmonic mean of precision and recall, and is a comprehensive evaluation of model performance. The values of the above three indicators recorded in the experimental results below are the average values of the results obtained from multiple experiments. The formulas for these metrics are as follows:

$$Precision = \frac{\sum_{i=1}^{c} TP_i}{\sum_{i=1}^{c} TP_i + \sum_{i=1}^{c} FP_i}, \tag{17}$$

$$Recall = \frac{\sum_{i=1}^{c} TP_i}{\sum_{i=1}^{c} TP_i + \sum_{i=1}^{c} FN_i}, \tag{18}$$

$$F1 = 2 * \frac{Precison * Recall}{Precision + Recall}. \tag{19}$$

In the above formulas, c represents the total number of entity categories. For entity type i, TP_i denotes the number of entities of this type correctly predicted by the model, that is, the number of entities whose predicted label and true label both correspond to type i. FP_i refers to the number of entities predicted as type i by the model, but whose true label is not of this type. FN_i represents the number of entities whose true label is type i, but were not predicted as such by the model.

To select a more suitable word sets combination as lexical features to be integrated into the model, we considered various possible combinations and conducted preliminary experiments. Fig. 3 shows the F1 scores obtained by models integrating different word sets when tested on the MDNER dataset. It can be seen that the model incorporating the "BMES" word set achieved the best performance, with an F1 score of 86.74%. Therefore, in subsequent experiments, the CAELF-GP model adopts the strategy of integrating "BMES" word sets features.

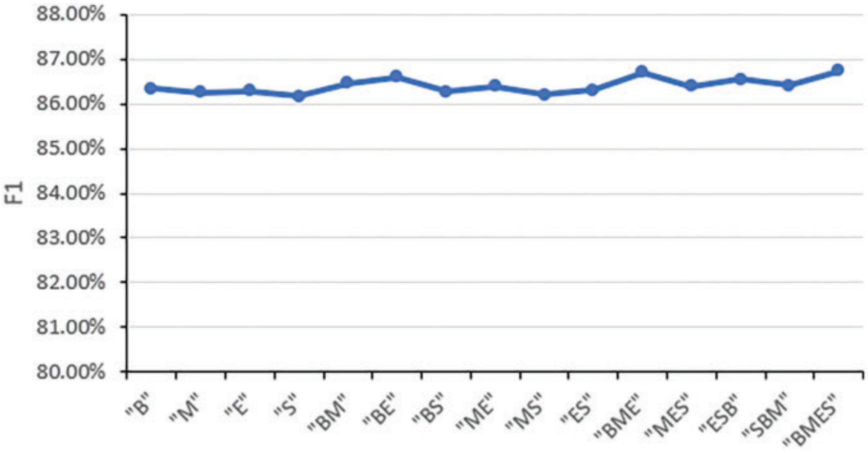


Figure 3: The performance of models integrating features from different word sets on the MDNER dataset

4.3 Results and Discussion

4.3.1 Comparison with Other Models

In order to validate the superiority of the CAELF-GP model based on cross-attention augmented lexical features, we conducted comparative experiments between the CAELF-GP model and the following models on the Weibo dataset, the CLUENER dataset, and the MDNER dataset. Among them, models (a)–(e) are baseline models for addressing the NER task, while (f) and (g) are models proposed in recent studies that have demonstrated excellent performance and also adopt lexical enhancement strategies. Tables 3–5 show the performance of each model obtained from our experiments on the above datasets. The best results are shown in bold in the tables.

Table 3: Results of different models on Weibo

Models	*P*/%	*R*/%	*F*1/%
Lattice LSTM	62.97	48.07	54.52
BERT-CRF	67.99	67.65	67.82
BERT-BiLSTM-CRF	67.10	64.44	65.74
SoftLexicon	70.6	67.87	69.21
GlobalPointer	**72.40**	67.44	69.83
MFGFF-BiLSTM-EGP	72.07	70.06	**71.05**
BABERT	71.22	**70.35**	70.78

(Continued)

Table 3 (continued)

Models	*P*/%	*R*/%	*F*1/%
CAELF-GP	71.15	70.30	70.72

Table 4: Results of different models on CLUENER

Models	*P*/%	*R*/%	*F*1/%
Lattice LSTM	73.00	73.15	73.07
BERT-CRF	76.76	80.76	78.71
BERT-BiLSTM-CRF	77.34	79.65	78.48
SoftLexicon	78.54	80.75	79.63
GlobalPointer	77.85	80.69	79.24
MFGFF-BiLSTM-EGP	78.65	**80.81**	79.72
BABERT	78.03	80.49	79.24
CAELF-GP	**79.39**	80.63	**80.01**

Table 5: Results of different models on MDNER

Models	*P*/%	*R*/%	*F*1/%
Lattice LSTM	84.28	81.85	83.05
BERT-CRF	83.21	84.24	83.72
BERT-BiLSTM-CRF	83.11	84.85	83.97
SoftLexicon	85.26	85.91	85.58
GlobalPointer	85.77	86.61	86.19
MFGFF-BiLSTM-EGP	86.59	**87.46**	**87.02**
BABERT	85.91	86.27	86.09
CAELF-GP	**86.68**	86.80	86.74

(a) Lattice LSTM: a model that adapts the original LSTM structure to incorporate lexicon information.

(b) BERT-CRF: a baseline model for NER tasks based on pre-trained BERT models and CRF.

(c) BERT-BiLSTM-CRF: a model that add the BiLSTM structure as a sequence encoding layer based on model (b) to extract additional contextual features in sentences.

(d) SoftLexicon: a method that integrates information from all matching words in the dictionary for each character into the character-based NER model by encoding vocabulary information at the character representation layer.

(e) GlobalPointer: a span-based NER framework that employs a multiplicative attention mechanism and relative position information to predict entities by globally considering the start and end positions of candidate entities.

(f) MFGFF-BiLSTM-EGP: a model combines the multiple fine-grained feature fusion (MFGFF) module with the BiLSTM neural network and uses the efficient GlobalPointer (EGP) to predict entity positions [28]. The MFGFF module consists of a RoBERTa-based character encoder, a word encoder based on character-word matching, and a sentence encoder based on SBERT.

(g) BABERT: an architecture that utilizes unsupervised statistical boundary information and directly encodes this boundary information into the BERT pre-trained language model [29].

The experimental results, as shown in Tables 3–5, demonstrate that the CAELF-GP model consistently outperforms various baseline models across three datasets in terms of F1 value and achieves comparable performance to recently proposed models, namely MFGFF-BiLSTM-EGP and BABERT. Specifically, on the MDNER dataset, the CAELF-GP model attains notable performance with precision of 86.68%, recall of 86.80%, and F1 value of 86.74%. It achieved an improvement of 0.55%–3.69% on F1 value compared to other baseline models. Notably, SoftLexicon and BERT-BiLSTM-CRF, both adopting a sequence-labeling architectures with similar structures, differ in that the former utilizes lexical information. The results underscore the superiority of SoftLexicon over the BERT-BiLSTM-CRF model, affirming the findings of Ma et al. [4]. The CAELF-GP model exhibits F1 improvements of 0.89%, 0.77%, and 0.55%, respectively, relative to the baseline GlobalPointer model. This underscores the efficacy of incorporating lexical information at the embedding layer, a strategy transferable to span-based NER models, enhancing their overall performance. Notably, the model's most substantial performance enhancement occurs on the Weibo dataset. This can be attributed to the dataset's limited size and noise, wherein the CAELF-GP model, through the integration of external lexical information and the strategic guidance provided by the cross-attention mechanism, achieves heightened accuracy in named entity recognition.

4.3.2 Ablation Experiment

To assess the efficacy of integrating lexical information and the cross-attention-based feature fusion approach to enhance model performance, we conducted ablation experiments on the aforementioned three datasets. As shown in Table 6, the exclusion of either the sole cross-attention mechanism or the complete removal of the lexical information fusion module (reverting the model to its foundational GlobalPointer form) led to a discernible deterioration in model performance. Specifically, the removal of the cross-attention mechanism resulted in a reduction of 0.40%, 0.74%, and 0.66% in F1 value on the Weibo, CLUENER, and MDNER datasets, respectively. The outcomes affirm that the incorporation of the cross-attention mechanism facilitates a more effective utilization of lexical information, thereby enhancing the overall performance of the model.

Table 6: Ablation results ($F1$/%)

Models	Weibo	CLUENER	MDNER
CAELF-GP	70.72	80.01	86.74
CAELF-GP (w/o cross-attention)	70.32	79.27	86.08
GlobalPointer	69.97	79.08	85.85

4.3.3 Entity Recognition Results on Musk Deer Dataset

To further assess the CAELF-GP model on NER task in the musk deer field, we record the recognition results of the model on different categories of entities, as shown in Table 7.

As shown in Table 7, it can be seen that the model performs well in recognizing entities of species, regions, cells and bodily elements with F1 values of 95.05%, 86.45%, 85.71% and 85.71%, respectively, but the F1 value for recognizing food is only 69.39%. The poor performance of the model in recognizing food-type entities may be attributed to two main reasons. On the one hand, it could be due to the limited number of food-type entities in the training set, which prevents the model from fully extracting relevant feature information. On the other hand, it may stem from the complexity of expressions for some food-type entities. For example, terms like "low-protein diet" were not fully recognized as a single entity by the model but were instead identified as separate parts, such as "low-protein" and "protein." In future work, we can expand our dataset by incorporating corpus that include food-type entities. Specifically, these entities in the additional corpus will have diverse expressions. Furthermore, we can explore more sophisticated entity boundary detection mechanisms, such as effectively incorporating statistical word boundary information [29], to address the challenges of recognizing food entities.

Table 7: Entity recognition results

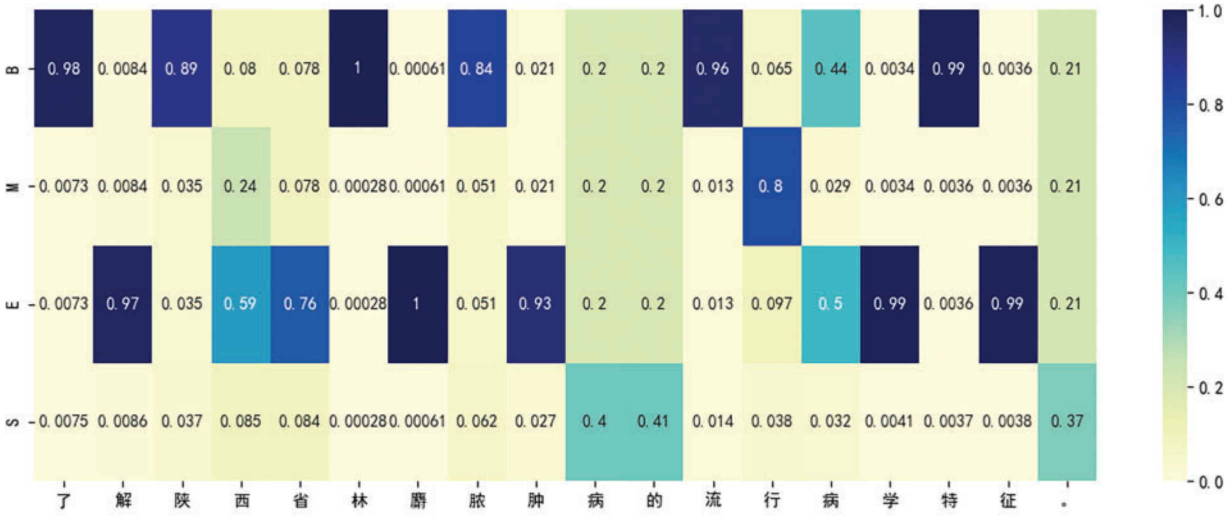
Entity	*P*/%	*R*/%	*F*1/%	Number of entities[2]
Species	95.05	95.05	95.05	2356
Area	84.81	88.16	86.45	638
Nature reserve	75.76	86.21	80.65	195
Tissue site	75.29	82.05	78.53	726
Disease	76.47	68.42	72.22	243
Gene	91.11	77.36	83.67	304
Cell	85.71	85.71	85.71	58
Bodily elements	81.82	90.00	85.71	732
Food	65.38	73.91	69.39	185
Bacteria	77.78	70.00	73.68	111
Research methods	88.00	82.50	85.16	1285
Average	81.56	81.76	81.47	621.18

Note: [2]The number of entities in the table is for the training set.

4.3.4 Case Study

To elucidate the performance enhancement of the CAELF-GP model, an empirical analysis is conducted by randomly selecting a sentence from the test set within the musk deer domain. The sentence undergoes processing through the trained model, and the resultant weight values corresponding to the four word sets ("BMES") of each character in the sentence are visually presented in the form of a heat map, as depicted in Fig. 4. Taking the sentence "To investigate the epidemiological characteristics of forest musk deer abscess disease in Shaanxi Province". As an illustrative example, it manifests three entities: "Shaanxi Province", "forest musk deer", and "abscess disease".

Upon scrutiny of Fig. 4, a discernible pattern emerges, revealing that, for the characters "Shaan" and "province", the model allocates more weight to the "B" and "E" word sets, respectively. This implies that "Shaan" is more likely to be the beginning of an entity, and "province" is more likely to be the end of an entity, thereby aiding the model in recognizing the entity "Shaanxi Province". The weight distribution for the entities "forest musk deer" and "abscess disease" exhibits a similar pattern. This indicates that the proposed lexical feature enhancement model based on cross-attention can improve the recognition capability of entity boundaries and semantic representation by focusing on more crucial lexical information, leading to an improvement in named entity recognition.



(To investigate the epidemiological characteristics of musk deer abscess in Shaanxi province.)

Figure 4: An example of distribution of the CAELF-GP model's attention on the "BMES" word sets

5 Conclusions and Future Work

In this work, we propose a Chinese NER model, CAELF-GP, with enhanced lexical features based on a cross-attention mechanism that is applicable to the musk deer domain. Given the diverse range of entity types within this domain and the inherent challenges in recognizing Chinese entity boundaries, we employ a vocabulary enhancement strategy. Our work confirms the efficacy of integrating vocabulary information to enhance the performance of the span-based NER model. Diverging from the approach of Ma et al. [4], who simply fused lexical information at the model embedding layer, we introduced a novel lexical feature fusion module based on the cross-attention mechanism, This mechanism guides the model to more effectively utilize lexical information by considering the varying degrees of association between each character and its corresponding word sets. This augmentation significantly enhances the model's semantic representation and its ability to recognize entity boundaries. Through rigorous experiments on two publicly available datasets and a musk deer domain dataset, we empirically demonstrate the superior performance of the CAELF-GP model compared to baseline models. The outcomes of this study provide effective technical support for downstream tasks such as the construction of knowledge graphs in the field of musk deer and the development of auxiliary diagnosis systems for musk deer diseases. This, in turn, promotes the conservation and research efforts related to musk deer.

Our model has shown good performance in the NER task in the musk deer domain, but there are still some limitations and areas worth further exploration. Due to the limited scope of available corpus and the continuous emergence of new domain-specific vocabulary, it is difficult for the lexicon to cover all domain terms. Moreover, in the model, the character representation integrates two types of information: dynamic word vectors generated by a pre-trained language model and static word vectors from an external lexicon. However, static word vectors cannot accurately represent polysemous words, so this fusion approach may introduce some bias when representing words with multiple meanings.

The performance of our model relies on the utilization of a high-quality external lexicon. Subsequently, we aim to train an even higher-quality musk deer domain lexicon based on a more extensive range of the related corpus. By replacing the musk deer domain-specific lexicon used in the model with lexicons from other domains, our model can be applied to NER tasks in those domains. Specifically, when the model is applied to larger datasets or more diverse domains, the increased training data or expanded domain lexicon results in relatively longer training and inference times. This is primarily due to the introduced lexical feature fusion module, which adds word-character matching and feature fusion processes, thereby increasing computational resource consumption. In future work, we will further explore to improve the computational efficiency of the lexical feature fusion module. Additionally, it is worthy to delve into a deeper analysis of the impact of fusing lexical information at a deeper layer of the model, specifically, the span representation layer, on the enhancement of model effectiveness.

Acknowledgement: The authors would like to express gratitude to the anonymous reviewers and editors for their kind comments and suggestions.

Funding Statement: This research was funded by 5·5 Engineering Research & Innovation Team Project of Beijing Forestry University (No. BLRC2023C02).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Yumei Hao, Haiyan Wang, Dong Zhang; data collection: Yumei Hao, Dong Zhang; analysis and interpretation of results: Yumei Hao, Haiyan Wang; draft manuscript preparation: Yumei Hao, Haiyan Wang, Dong Zhang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Haiyan Wang, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Wang Q, Su X. Research on named entity recognition methods in Chinese forest disease texts. Appl Sci. 2022;12(8):3885. doi:10.3390/app12083885.
2. Yu Y, Wang Y, Mu J, Li W, Jiao S, Wang Z, et al. Chinese mineral named entity recognition based on BERT model. Expert Syst Appl. 2022;206:117727. doi:10.1016/j.eswa.2022.117727.
3. Zhang H, Wang X, Liu J, Zhang L, Ji L. Chinese named entity recognition method for the finance domain based on enhanced features and pretrained language models. Inf Sci. 2023;625(17):385–400. doi:10.1016/j.ins.2022.12.049.
4. Ma R, Peng M, Zhang Q, Wei Z, Huang X. Simplify the usage of lexicon in Chinese NER. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10. p. 5951–60. doi:10.18653/v1/2020.acl-main.528.
5. Zhong S, Chen X, Zhao M, Zhang Y. A Chinese named entity recognition method incorporating a lexical set-level attention mechanism. J Jilin Univ. 2022;52(5):1098–105 (In Chinese). doi:10.13229/j.cnki.jdxbgxb20200984.
6. Chen CR, Fan Q, Panda R. CrossViT: cross-attention multi-scale vision transformer for image classification. In: Proceedings of the 18th IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 347–56. doi:10.1109/ICCV48922.2021.00041.
7. Su J, Murtadha A, Pan S, Hou J, Sun J, Huang W, et al. Global pointer: novel efficient span-based approach for named entity recognition. arVix:2208.03054. 2022.
8. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv:1508.01991. 2015.
9. Chiu JPC, Nichols E. Named entity recognition with bidirectional LSTM-CNNs. arXiv:1511.08308. 2015.
10. Luo Y, Xiao F, Zhao H. Hierarchical contextualized representation for named entity recognition. In: Proceedings of the AAAI 34th Conference on Artificial Intelligence; 2020 Feb 7–12; New York, NY, USA. p. 8441–8.
11. Li X, Feng J, Meng Y, Han Q, Wu F, Li J. A unified MRC framework for named entity recognition. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10. p. 5849–59.

12. Chen X, Li L, Qiao S, Zhang N, Tan C, Jiang Y, et al. One model for all domains: collaborative domain-prefix tuning for cross-domain NER. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence; 2023 Aug 19–25; Macao, China. p. 5030–8.
13. Li X, Meng Y, Sun X, Han Q, Yuan A, Li J. Is word segmentation necessary for deep learning of Chinese representations? arXiv:1905.05526. 2019.
14. Zhang Y, Yang J. Chinese NER using lattice LSTM. arXiv:1805.02023. 2018.
15. Gui T, Ma R, Zhang Q, Zhao L, Jiang YG, Huang X. CNN-based Chinese NER with lexicon rethinking. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence; 2019 Aug 10–16; Macao, China. p. 4982–8.
16. Sui D, Chen Y, Liu K, Zhao J, Liu S. Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network. In: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019 Nov 3–7; Hong Kong, China. p. 3828–38.
17. Gui T, Zou Y, Zhang Q, Peng M, Fu J, Wei Z, et al. A lexicon-based graph neural network for Chinese NER. In: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019 Nov 3–7; Hong Kong, China. p. 1040–50.
18. Li X, Yan H, Qiu X, Huang X. FLAT: Chinese NER using flat-lattice transformer. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10. p. 6836–42.
19. Liu W, Fu X, Zhang Y, Xiao W. Lexicon enhanced Chinese sequence labeling using BERT adapter. In: The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; 2021 Aug 1–6. p. 5847–58.
20. Liu W, Xu T, Xu Q, Song J, Zu Y. An encoding strategy based word-character. In: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019 Jun 2–7; Minneapolis, MN, USA; p. 2379–89.
21. Ates GC, Mohan P, Celik E. Dual cross-attention for medical image segmentation. Eng Appl Artif Intell. 2023;126(4):107139. doi:10.1016/j.engappai.2023.107139.
22. Zhou Y, Huo C, Zhu J, Huo L, Pan C. DCAT: dual cross-attention-based transformer for change detection. Remote Sens. 2023;15(9):2395. doi:10.3390/rs15092395.
23. Xue H, Ma J, Guo X. A hierarchical multi-modal cross-attention model for face anti-spoofing. J Vis Commun Image Represent. 2023;97(8):103969. doi:10.1016/j.jvcir.2023.103969.
24. Wei X, Zhang T, Li Y, Zhang Y, Wu F. Multi-modality cross attention network for image and sentence matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 10938–947.
25. Su J, Ahmed M, Lu Y, Pan S, Bo W, Liu Y. RoFormer: enhanced transformer with rotary position embedding. Neurocomputing. 2024;568:127063. doi:10.1016/j.neucom.2023.127063.
26. Xu L, Tong Y, Dong Q, Liao Y, Yu C, Tian Y, et al. CLUENER2020: fine-grained named entity recognition dataset and benchmark for Chinese. arXiv:2001.04351. 2020.
27. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: 27th Annual Conference on Neural Information Processing Systems; 2013 Dec 5–10; Lake Tahoe, NV, USA. p. 3111–9.
28. Wang X, Peng C, Li Q, Yu Q, Lin L, Li P, et al. A Chinese nested named entity recognition model for chicken disease based on multiple fine-grained feature fusion and efficient global pointer. Appl Sci. 2024;14(18):8495. doi:10.3390/app14188495.
29. Jiang P, Long D, Zhang Y, Xie P, Zhang M, Zhang M. Unsupervised boundary-aware language model pretraining for Chinese sequence labeling. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; 2022 Dec 7–11; Abu Dhabi, United Arab Emirates. p. 526–37. doi:10.18653/v1/2022.emnlp-main.34.