applied sciences

MDPI

Article

Enhanced Chinese Domain Named Entity Recognition: An Approach with Lexicon Boundary and Frequency Weight Features

Yan Guo [1], Shixiang Feng [1], Fujiang Liu [2,*], Weihua Lin [3], Hongchen Liu [1], Xianbin Wang [4], Junshun Su [5] and Qiankai Gao [6]

[1] School of Computer Science, China University of Geosciences, Wuhan 430078, China; guoyan49@cug.edu.cn (Y.G.); fengsx@cug.edu.cn (S.F.); codernewhoc@cug.edu.cn (H.L.)
[2] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430078, China
[3] Hubei Key Laboratory of Regional Ecology and Environmental Change, School of Geography and Information Engineering, China University of Geosciences, Wuhan 430078, China; linweihua@cug.edu.cn
[4] Piesat Information Technology Co., Ltd., Beijing 100195, China; wangxianbin@piesat.cn
[5] Xi'ning Natural Resources Comprehensive Survey Center, China Geological Survey, Xi'ning 810016, China; sujunshun@163.com
[6] Kunming Natural Resources Comprehensive Survey Center, China Geological Survey, Kunming 650111, China; gaoqiankai@126.com
* Correspondence: liufujiang@cug.edu.cn

Abstract: Named entity recognition (NER) plays a crucial role in information extraction but faces challenges in the Chinese context. Especially in Chinese paleontology popular science, NER encounters difficulties, such as low recognition performance for long and nested entities, as well as the complexity of handling mixed Chinese–English texts. This study aims to enhance the performance of NER in this domain. We propose an approach based on the multi-head self-attention mechanism for integrating Chinese lexicon-level features; by integrating Chinese lexicon boundary and domain term frequency weight features, this method enhances the model's perception of entity boundaries, relative positions, and types. To address training prediction inconsistency, we introduce a novel data augmentation method, generating enhanced data based on the difference set between all and sample entity types. Experiments on four Chinese datasets, namely Resume, Youku, SubDuIE, and our PPOST, show that our approach outperforms baselines, achieving F1-score improvements of 0.03%, 0.16%, 1.27%, and 2.28%, respectively. This research confirms the effectiveness of integrating Chinese lexicon boundary and domain term frequency weight features in NER. Our work provides valuable insights for improving the applicability and performance of NER in other Chinese domain scenarios.

Keywords: Chinese named entity recognition; dual pointer network; lexicon enhancement; Chinese paleontology popular science

Citation: Guo, Y.; Feng, S.; Liu, F.; Lin, W.; Liu, H.; Wang, X.; Su, J.; Gao, Q. Enhanced Chinese Domain Named Entity Recognition: An Approach with Lexicon Boundary and Frequency Weight Features. *Appl. Sci.* **2024**, *14*, 354. https://doi.org/10.3390/app14010354

Academic Editor: Rafael Valencia-Garcia

Received: 19 November 2023
Revised: 7 December 2023
Accepted: 11 December 2023
Published: 30 December 2023

1. Introduction

Over the past few years, with the development of the Internet and information technology, various domains have accumulated a large amount of textual data, containing valuable information and knowledge [1]. Effectively utilizing this textual data can have a significant impact on various domains. Information extraction refers to automatically extracting key information from massive data, named entity recognition (NER), as a subtask, and has become a research hotspot in various domains [2].

In the domain of education, popular science (also known as popsci) is a scientific explanation targeted at the general audience [3]. Popular science education can cultivate the public's interest in science and promote the popularization and dissemination of science and technology in society, thereby fostering social development. Faced with the growing data in the domain of popular science education, how to quickly and accurately identify

keywords and reveal deep semantic relationships is an urgent issue for intelligent systems in popular science education. Many discoveries and research results in the domain of paleontology contribute to understanding the evolution of the Earth, as well as the origin and evolution of life and humanity [4], and one of the advantages of paleontology is popularization. In order to promote the latest research results to the public and advance the development of popular science education, the efficient extraction of key information from texts in the domain of paleontology popular science using deep learning-based information extraction technology has gained widespread attention from domain experts. One of the primary tasks in the domain of Chinese paleontology popular science NER is to identify different types of entities, such as a person, geological time, strata, and fossil, etc., from unstructured textual data. NER involves preprocessing, feature extraction, and classification to identify entities from unstructured textual data. By recognizing entities, valuable information can be extracted from textual data, laying the foundation for subsequent tasks [5], such as relationship extraction, sentiment analysis [6,7], and knowledge graph construction.

The NER methods can be summarized from rule-based methods to deep learning-based methods. In the early stage, rule-based methods utilized predefined and inductive assumptions and rules to identify entity names in the text and classify them [8,9]. However, with the increase in annotated data, rule-based methods need constant updates to discover more entities. Traditional machine learning methods, mainly based on supervised learning, involve training classification or sequence labeling models on annotated datasets to learn features of positive or negative instances, or type features. For example, based on a support vector machine (SVM) [10], hidden Markov model (HMM) [11,12], and conditional random field (CRF) [13], etc., those methods no longer require the manual construction of rules, but they still require feature selection. In recent years, with the development of deep learning, various neural networks have been used to address NER problems, such as convolutional neural networks (CNNs) [14], recurrent neural networks (RNNs), such as long short-term memory (LSTM) [15] and bidirectional LSTM (BiLSTM) [16], graph neural networks (GNNs) [17], and attention mechanisms [18], etc. Neural network methods typically use vectors to learn the syntax and context information of language. Pretrained models, represented by BERT (bidirectional encoder representation from transformers) [19], have a significant advantage in extracting contextual information features from text due to training on large-scale datasets. This notably improves various metrics for natural language processing (NLP) tasks. In the field of NER, significant achievements have been made by methods based on fine-tuning pretrained models, like BERT [19,20], BERT-CRF [21,22], and BERT-BiLSTM-CRF [23]. Pretrained models can learn contextual features of text on the basis of training on large-sample data, enabling fine-tuning in situations with limited annotated data and subsequently learning contextual features for downstream tasks.

The models mentioned above are based on sequence labeling to accomplish NER, which means performing a multi-classification task for each token. However, in the domain of Chinese paleontology popular science, as shown in Figure 1, the entity “古脊椎动物与古人类研究所” (Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences) consists of 12 Chinese characters; the span “寒武纪地层” (Cambrian Strata) includes the Geological Time entity, the Cambrian, and the Strata entity, Cambrian Strata; the span “古生物学家 Grant Zazula” (the Paleontologist named Grant Zazula) contains both Chinese and English characters. Due to the existence of the mentioned phenomena and problems, directly applying sequence labeling models will significantly degrade the model’s performance, and these problems directly hinder the further development of NER in this domain. To address this issue, the pointer labeling task is proposed. Pointer labeling is a framework for NER. It sets start and end pointers to record the starting and ending positions of each entity, and simultaneously annotates the entity type. This method excels in recognizing long entities and nested entities. However, using only the pointer labeling for NER results in the loss of Chinese lexical boundary information, causing the model to erroneously output any two boundaries as entities. Recently, the

integration of lexicon-level features into Chinese NER has attracted widespread attention. Integrating lexicon-level features involves incorporating lexicons as external information, helping to determine the span and type of entities. At the same time, in our domain, we found that the distribution of domain-specific lexicon also affects the performance of NER. Currently, the utilization of distributional features of a domain-specific lexicon is still in its early stages.

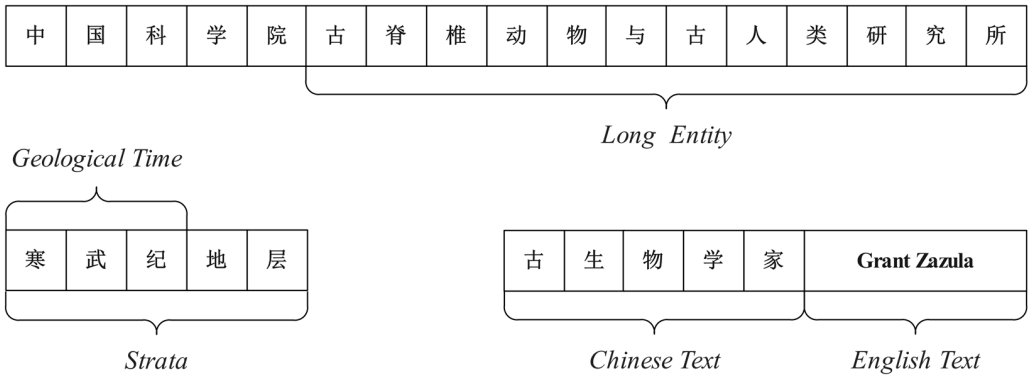


Figure 1. An example containing a long entity, nested entities, and a mixed text of Chinese and English.

To address the above-mentioned issues, in this paper, we propose an enhanced NER method based on lexicon boundary and frequency weight features, named the BERT-Lexicon-PointerNetwork (BERT-LPN) model. The model is proposed for extracting Chinese paleontology popular science entities and it consists of an encoder based on BERT, a lexicon feature fusion layer and a fully connected classifier layer from the bottom up: the encoder is a Chinese character-level model based on BERT, and it maps Chinese characters to a low-dimensional, highly dense real-vector space to extract latent semantic information from Chinese entity elements; the lexicon feature fusion layer takes lexicon boundary and weight information from Chinese word segmentation as input to capture the lexicon-level features of the input; finally the fully connected classifier layer takes BERT-encoded embeddings along with lexicon-level features as input to generate start position and end position labels for entities. The experimental results indicate that our proposed model outperforms previous models, and our approach achieves state-of-the-art (SOTA) performance on the constructed dataset.

The main contributions of this research are summarized as follows:

- Based on the characteristics of texts in the domain of Chinese paleontology popular science, we propose a NER model enhanced with Chinese lexicon boundary and frequency weight features. This model is designed for identifying and extracting entities from unstructured textual data.
- Based on the structural characteristics of our model, we propose a data augmentation method utilizing all entity types and sample entity types to alleviate the inconsistency between training and prediction tasks.

To comprehensively evaluate the performance of our model, we established a new dataset, namely the PPOST dataset, specifically designed for NER in the Chinese paleontology popular science domain. The dataset primarily consists of data from authoritative institutions in the domain of paleontology popular science in China. Experimental results on both public datasets and the PPOST dataset validate the effectiveness of our approach. In addition, we also analyze the possibility of applying our method to other domains, providing a reference for enhancing NER performance in other domains.

The remaining sections of this paper are organized as follows. Section 2 introduces NER work based on deep learning that is closely related to our work. Section 3 presents our proposed model and core algorithms. Section 4 provides a detailed description of the datasets used in our experiments and analyzes the experimental results. Finally, we conclude the article.

2. Related Work

In this section, we introduce works closely related to our proposed model, including NER models that integrate lexicon-level features, methods for addressing domain data scarcity, and NER models based on pointer labeling, as follows.

2.1. Lexicon-Based Chinese NER

Chinese, unlike languages, such as English, that use spaces as separators, encounters the issue of word segmentation, where Chinese takes words as the basic semantic units. To better adapt to Chinese NER, recent studies have found that introducing Chinese lexicon-level features can enhance the performance of NER models. For example, Zhang et al. [24] proposed a lattice LSTM model for Chinese NER. This model is an extension of character-based NER models, incorporating words as input and additional gates to control information flow. The model explicitly utilizes word and word order information. However, due to the RNN structure adopted by the lattice LSTM model, it cannot capture long-distance dependencies, and the introduction of lexical information is lossy. To address these limitations, Li et al. [25] proposed the FLAT model for Chinese NER. This model is based on a transformer [26] to solve the lexical loss problem in lattice LSTM and uses relative position encoding to adapt a transformer to NER, enabling efficient GPU parallel computation. These methods, combined with pre-trained language models, explore Chinese NER and achieve SOTA performance on several Chinese benchmark datasets. However, Guo et al. [27] and Liu et al. [28] believe that existing lexicon-based models only conform to lexicon features through shallow and randomly initialized encoding layers without integrating them into the underlying layers of pre-trained language models to explore deep lexicon knowledge. To address this, they proposed methods to deeply integrate external lexicon features into the pre-trained language model BERT, achieving more effective fusion of entity boundaries and lexical information.

All of these studies have explored the integration of Chinese lexicon-level features, but they did not address the practical issues faced by more detailed domains. Moreover, these methods rely on the quality of lexicons and have certain requirements for computational resources. In contrast, this paper focuses on extracting entities from the Chinese paleontology popular science domain, integrating Chinese lexicon-level features in a simple and effective manner while enhancing relative positional information. It provides a reference for addressing the practical issues faced by this domain.

2.2. Prompt and Pointer Network

Typically, there is not much annotated data for downstream tasks, and there is often a gap between pre-trained BERT models on large corpora and fine-tuning on downstream tasks. To address the gap, some researchers have proposed modeling downstream tasks in the form of pre-training tasks [29]. Taking BERT as an example, some tasks can be modeled as masked language modeling (MLM) or next sentence prediction (NSP), and this approach is particularly effective in scenarios with small datasets [30]. For instance, Cui et al. [31] and others proposed the TemplateNER method, which completes NER using prompt learning. This method employs a template-based approach to solve the problem of few-shot NER: a template is pre-defined, for example, “<candidate_span>is a <entity_type> entity”. With the help of a suitable prompt, this method reduces the difference between pre-training and fine-tuning, allowing the model to achieve good results with a small number of samples. It significantly outperforms traditional sequence labeling methods and distance-based few-shot NER methods in cross-domain and few-shot scenarios. Subsequently, methods based

on the prompt mechanism have been widely used in NER [32,33], such as modeling the task as machine reading comprehension and question answering (MRC-QA) format, fitting the NSP task for NER [34].

The NER model based on the dual pointer network aims to address both the challenges of recognizing long entities and nested entities. Lu et al. [35] proposed a unified extraction framework, where the NER module is based on a pre-trained model, employing a multi-layer dual pointer network. Input samples are constructed based on the template "[SPOT]<entity_type>[text]", predicting the start and end positions of possible entities in the samples, and NER is completed through the decoding algorithm. However, real-world datasets often exhibit a long-tail distribution problem, with a few types representing the majority of the samples. The multi-layer dual pointer network requires setting up a classifier for each entity type, potentially leading to some classifiers not being adequately trained, thus impacting the overall performance of NER model.

To mitigate this issue, some researchers proposed a model based on a single-layer dual pointer network and the prompt mechanism. Gong et al. [36] introduced a model based on a single-layer dual pointer network, constructing the input in the format "<entity_type>[SEP]<text>" to accomplish NER. However, Su et al. [37] argue that the conventional design of the dual-pointer network, when performing NER or MRC, typically employs two modules to separately identify the start and end positions of entities. This can lead to inconsistencies during training and prediction. To address this issue, they proposed global pointer model which uses a globally normalized approach for NER. In non-nested scenarios, it achieves results comparable to CRF, and in nested scenarios, it performs well. Moreover, it allows for fully parallelized training, significantly accelerating the training process. This method addresses the inconsistency between training and prediction tasks from a model perspective.

All these studies have explored NER methods based on pre-trained models and proposed unique insights and solutions for various issues. However, most of them optimized from the perspective of the model without considering the scarcity and differences in datasets across various domains in real-world applications. In contrast, this paper introduces a data augmentation method that enhances NER performance by increasing training data, alleviating the inconsistency between training and prediction tasks. This approach provides a reference for improving NER performance in domain-specific applications.

3. Methods

In this section, we provide a detailed description of the proposed model architecture, and the overall structure of our model is illustrated in Figure 2. Firstly, we introduce some preprocessing operations specific to this domain to acquire the necessary experimental foundation. Secondly, based on the WordPiece [38] algorithm, we use BERT to obtain character-level embeddings and acquire lexicon-level features through Chinese word segmentation; lexicon-level features fusion are accomplished using a multi-head self-attention mechanism. Finally, the embeddings based on BERT and the lexicon-level features serve as inputs to the dual pointer network, which is applied to complete the NER through decoding in the domain of Chinese paleontology popular science.

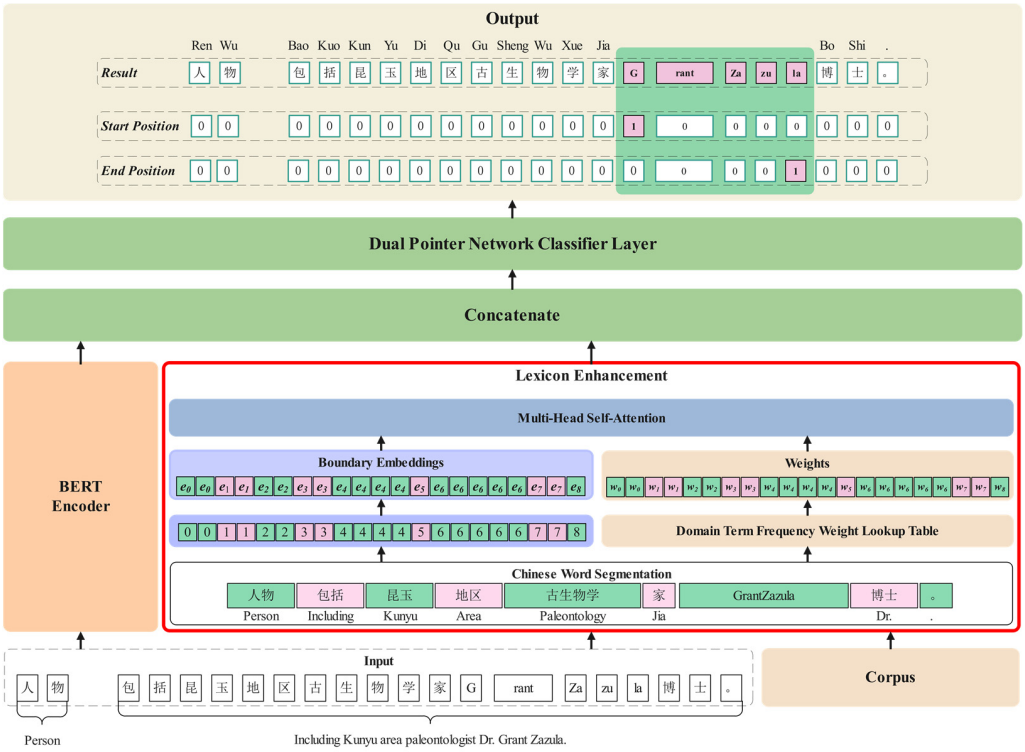


Figure 2. BERT-LPN: The input is processed through BERT to obtain character-level embedding vectors and the Chinese lexicon-level feature module to acquire lexicon-level embedding vectors (red box). After concatenation, a dual pointer network composed of two fully connected networks is used to accomplish the NER.

3.1. Preprocessing

To obtain domain term frequency weight features, we base our calculations on the collected corpus in the domain of Chinese paleontology popular science. TF-IDF [39] is a statistical method that consists of term frequency (TF) and inverse document frequency (IDF) and evaluates the importance of a term to a document set or one of the documents in a corpus, the TF-IDF scores of domain terms are computed as domain term frequency weights in our approach. Specifically, for the raw text corpus we collected, we employ the Chinese word segmentation tool pkuseg [40] for segmentation. After removing stop words and segmenting the Chinese text, the TF-IDF scores of each term are calculated as the weight feature for domain term frequency. Since there are N documents in total, term t has N TF-IDF scores, and we integrate these scores using the averaging strategy because this approach preserves more features, thereby better assessing the weight of a term in the corpus and obtaining the final feature for domain term frequency. The method for calculating domain term frequency weights is as follows:

$$W_{t,d} = tf_{t,d} \cdot log(\frac{N}{1 + df_t}), \tag{1}$$

$$Weight_t = \frac{\sum_{i=0}^{N} W_{t,d}^i}{N}, \tag{2}$$

where $tf_{t,d}$ represents the term frequency of term t in document d, log is the natural logarithm, df_t represents the number of documents out of N that contain term t, $W_{t,d}$ is the TF-IDF score of term t in document d, and $Weight_t$ represents the average weighted domain term frequency weight of term t. The final domain term frequency weights are shown in Table 1.

Table 1. Domain term frequency weight lookup table (DTFW-lookup Table).

Domain Term	Weight
恐龙 (Dinosaur)	$W_{恐龙} = 1.36 \times 10^{-3}$
侏罗纪 (Jurassic)	$W_{侏罗纪} = 3.67 \times 10^{-4}$
中国科学院 (Chinese Academy of Sciences)	$W_{中国科学院} = 7.38 \times 10^{-5}$
…	…
$Term_i$	$W_i = x$

Usually, the NER based on a dual pointer network and the MRC-QA mechanism faces the problem of inconsistency between training and prediction tasks. That is, during training, fitting is carried out with the correct entity type as a prompt, while in prediction tasks, it is often not possible to specify the correct prompt for prediction, meaning the prediction task also includes an entity classification task. To alleviate this inconsistency and enhance the model's generalization ability, we propose a data augmentation method based on the set difference between all entity types and sample entity types.

As shown in Figure 3, our dataset includes 13 types of entities, denoted as the set $AllTypes$ = {Strata, FossilClassification (FossilClass), Subject, Event, Position, Time, Feature, Person, Organization, GeologicalTime (GeoTime), Address, Fossil, BiologicalClassification (BioClass)}, while the current set of entity types included in the samples is denoted as set $Gold_{sample}$. Through set difference operations on the sets, the entity types not included in the current samples are generated. In this way, after calculating the difference set for all samples, the labels of the enhanced data are set to an empty set to avoid introducing noisy data, as follows :

$$Diff_{sample} = \left\{ type \in AllTypes : type \notin Gold_{sample} \right\}, \tag{3}$$

$$DA^{item}_{sample} = \left\{ Text_{sample}, EntityType = Diff^{item}_{sample}, \ Label = \varnothing \right\}, \tag{4}$$

where $Diff^{item}_{sample}$ is an element in the collection $Diff_{sample}$, and DA^{item}_{sample} is a piece of data generated when the entity type is $Diff^{item}_{sample}$.

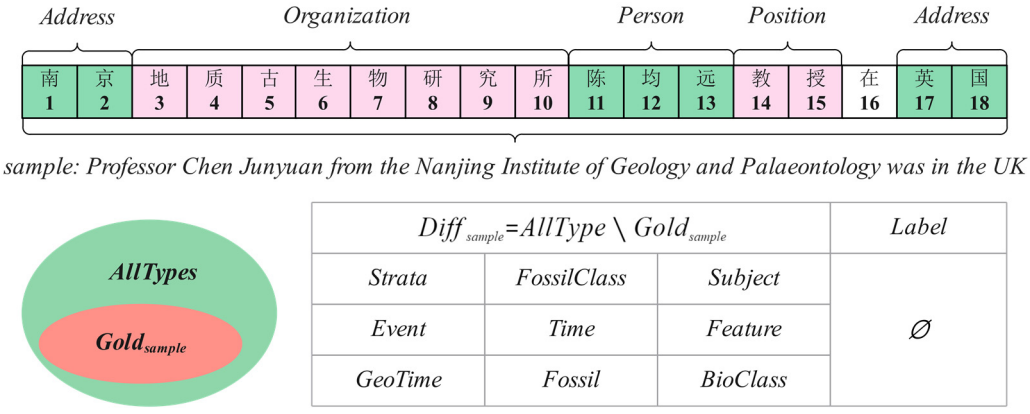


Figure 3. Data augmentation: Calculate the set difference between the set $AllTypes$ and the set $Gold_{sample}$ to obtain augmented data.

Then, based on the given predefined data augmentation ratio, a certain number of augmented data are generated through a random sampling algorithm and merged into the training dataset for model training. By introducing augmented data, the model's perception of entity types not present in the sample data is strengthened, effectively alleviating the inconsistency between the training and prediction tasks.

3.2. BERT Encoder

To effectively capture textual information from text, the encoder is constructed based on a pretrained language model, BERT. Given an input (i.e., a sentence or a word) $S = \left[s_1, s_2, \ldots, s_{|S|}\right]$, the encoder tokenizes it using a predefined vocabulary and encodes s_i into a vector representation $\overline{s_i}$. For detailed steps, please refer to the literature [19]. The encoder embeds the text S into vector representations $\Gamma^S = \mathbb{R}^{len \times dim}$, where len represents the maximum acceptable input length for the model, dim represents the dimension of BERT token embedding vectors, and in the case of BERT-base dim is set to 768. Generally, Γ^S_j can be considered as the embedding of the j^{th} token, and it is worth noting that due to the usage of the WordPiece tokenization algorithm, variations may occur in our methods. Unlike other Chinese NER models, we use the WordPiece tokenization algorithm to represent Chinese characters as independent tokens and English as their subword representations, and the method can simplify the model input and reduce the complexity of model. Additionally, based on BERT, we fine-tune the model in conjunction with the MRC-QA task. The detailed steps are as follows:

$$input = " < entity\_type >, < text > ", \tag{5}$$

$$tokenized = WordPiece(input) = \{t_1, t_1, \ldots, t_n\}, \tag{6}$$

$$\chi_i = e(t_i). \tag{7}$$

In the above, using the provided *text* sequence and its corresponding entity type *entity_type*, we construct the *input* and obtain *tokenized* representations for the input using the WordPiece algorithm. The variable χ_i represents the embedding vector by BERT for the given *tokenized* representation.

3.3. Lexicon Encoder

To better harness Chinese lexicon-level information, we have introduced lexicon boundary and frequency weight features. Given the input, it is segmented into individual lexicons or characters after applying a segmentation tool, as follows:

$$SegId_{input} = \{(0, a), (a+1, b), \ldots, (x, y), \ldots, 0 \ldots, 0\}, \tag{8}$$

where (x, y) represents the segment composed of the characters from the x^{th} to the y^{th} position in the input sequence, in other words, in the ordered set of $SegId_{input}$, each element represents a Chinese lexicon or character. Let us define a function $f(i, j)$—element i repeats j times results in an ordered set, as follow:

$$f(i, j) = \{i \; repeats \; j \; times\}, \tag{9}$$

giving $(x, y)_{id} \in SegId_{input}$, where id is the index of $(x, y)_{id}$ in the ordered set $SegId_{input}$, and we substitute $f(i, j)$ to obtain $f(id, y - x)$, where $y - x$ is the length of the current lexicon or character segment. In summary, we obtain the lexicon boundary features of the input sequence through Algorithm 1, and its time complexity is $O(n)$, where n is the length of the input sequence. At the same time, if n does not reach the maximum allowed length of the model, padding with zeros is applied after the last element of *boundary* until it reaches its maximum. On the other hand, as shown in Figure 4, when there are three entities which belong to the same type (BioClass) in a given input (lexicons 0, 2, and 4), without considering relative positional information, the model may arbitrarily combine their boundaries in the output. For example, the phrase "恐龙位于爬行动物和鸟类" (Dinosaurs are between reptiles and birds, with characters between the start of lexicon 0 and the end of lexicon 4) could be incorrectly combined into a single BioClass entity. With the inclusion of boundary information, the model becomes sensitive to the length and span of entities.

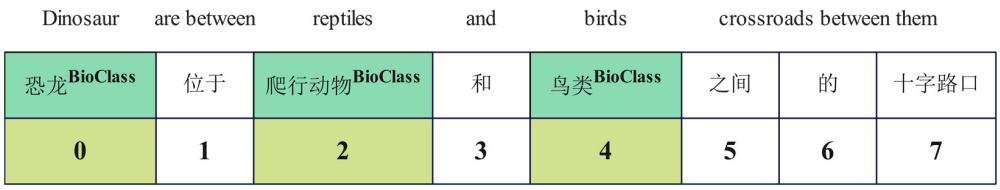


Figure 4. Boundary features incorporate simultaneously lexical boundary information and relative positional information: the method introduces relative positional information from another perspective, making it sensitive to the length and span of entities, reducing the likelihood of the model predicting the combination of any two entity boundaries as the target.

Algorithm 1 Lexicon Boundary Generation

Input: *Ordered Set of* $SegId_{input}$
Output: *Ordered Set of Lexicon Boundary*
1: $boundary = \{\}$
2: **for** each $(x, y)_{id} \in SegId_{input}$ **do**
3: $item = \{\}$
4: **for** each $index \in [0, y - x]$ **do**
5: $item = item \cup \{i\}$
6: **end for**
7: $boundary = boundary \cup item$
8: **end for**
9: **return** *boundary*

Given the $boundary_i$, we look up the boundary index embedding e_i from the specialized embedding matrix. This matrix contains fixed-size embeddings for each boundary index 0, 1, 2, …, and these embeddings are learned through backpropagation.

Similarly, $(x, y)_{weight} \in SegId_{input}$, where *weight* represents the term frequency weight of the xth to yth subsequence of the input, and this weight value is obtained through the DTFW-lookup Table in Section 3.1. Substitute $f(i, j)$ to obtain $f(weight, y - x)$, where $y - x$ is the length of the current segment. This step will result in the lexicon or character frequency weight representation w_i for the input. Specifically, when the lexicon or character do not exist in the DTFW-lookup Table, we use zero as the default value, as follows:

$$w_i = \begin{cases} DTFW - Lookup\ Table_{term}, & if\ term\ \in\ DTFW - Lookup\ Table \\ 0, & else. \end{cases} \tag{10}$$

Ultimately, for the given $boundary_i$, the lexicon-level feature representation at position i is as follows (whereas $\circ$ denotes concatenation):

$$\theta_i = e_i \circ w_i. \tag{11}$$

Simultaneously, in order to capture features of different dimensions for boundaries and weights, we employ a multi-head self-attention mechanism for feature fusion. The workflow of the above algorithm is illustrated in Figure 5.

In this mechanism, the multi-head self-attention mechanism calculates attention weights and outputs through the query (Q), key (K), and value (V) inputs. Specifically, for the inputs Q, K, and V, the calculation formula for the output vectors is as follows:

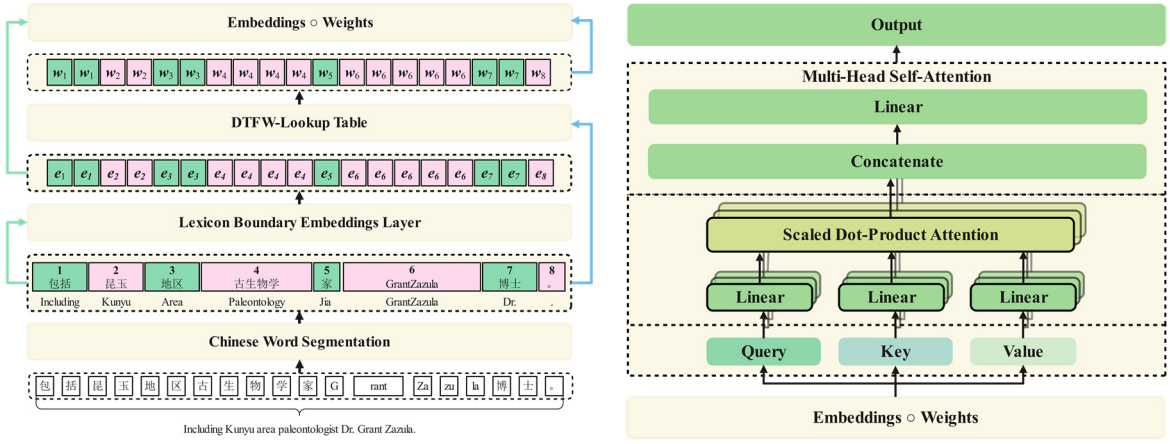
$$MultiHead(Q, K, V) = head_1 \circ head_2 \circ \ldots \circ head_h \times W^o, \tag{12}$$

where $W^O \in \mathbb{R}^{hd_v \times d_{model}}$, and $head_i$ is calculated as follows:

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right), \tag{13}$$

$$where\ Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{14}$$

where $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ (h represents the number of heads, and d_{model} is the dimension of the input vector. Generally, $d_k = d_v = d_{model}/h$).



(**a**) (**b**)

Figure 5. (**a**) Method for enhancing Chinese lexicon-level features: after undergoing Chinese word segmentation processing, we obtain the lexicon-level representation. We encode and embed its boundaries while simultaneously looking up the DTFW-lookup Table to acquire domain term frequency weights, and concatenate embeddings and weights; (**b**) after concatenating them, we use the multi-head self-attention mechanism for feature fusion, with the final output serving as the Chinese lexicon-level features vector.

3.4. Classifier

To address the challenges posed by long entities and nested entities, we apply a dual pointer fully connected network classifier. The classifier consists of two pointers: one indicating the start position of an entity and the other indicating the end position. In addition, a sigmoid function is applied to map the fully connected network output into the [0, 1] range. Finally, based on a specified *threshold*, the output is discretized into {0, 1} for representation, where "1" indicates that the position is the start position or end position of the entity, and "0" indicates the opposite. The specific classification process is illustrated in Figure 6. Given the embedding vector representation of the input, *context*, is determined as follows:

$$context_i = \chi_i \circ \theta_i. \tag{15}$$

We utilize the *context* vector as the input for the classifier and decode based on predefined *threshold*, as follows:

$$output = classifier(context), \tag{16}$$

$$position_i^s, position_i^e = \begin{cases} 1, & if\ position_i > threshold \\ 0, & else, \end{cases} \tag{17}$$

where $position_i^s$ denotes the start position classifier's output at position i, and $position_i^e$ denotes the end position classifier's output at position i.

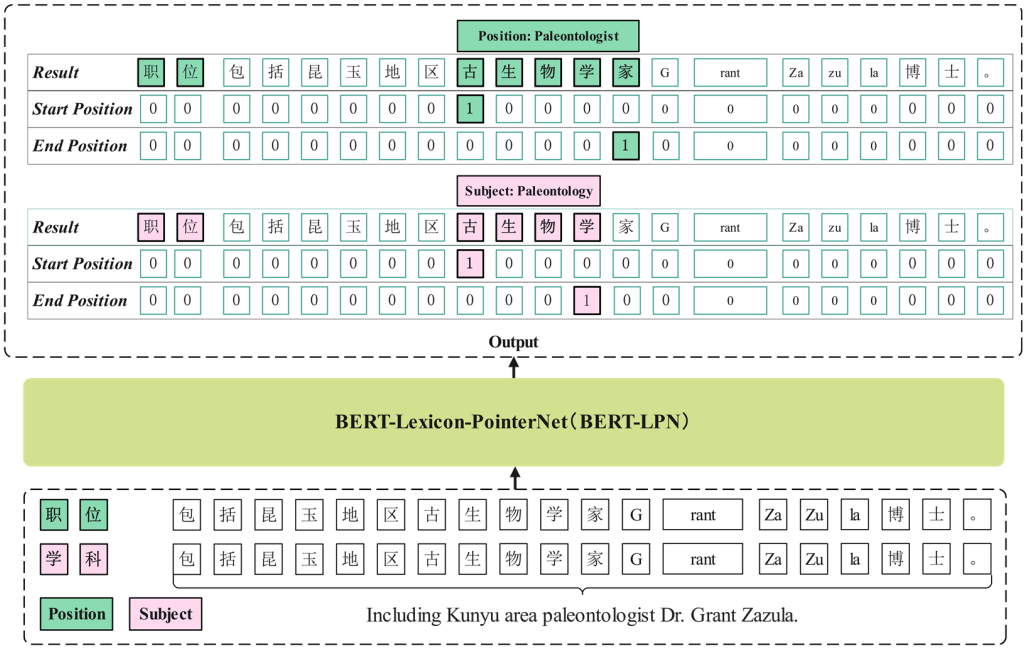


Figure 6. For nested-type entities, our model processes multiple pairs of "<entity_type>,<text>" inputs, and the classifier ultimately outputs the start and end position labels for the entities.

3.5. Loss Function

The loss function is applied to gauge the performance of the model, specifically the disparity between the model's predictions and the actual targets. This discrepancy is typically referred to as the loss value, and the objective of the loss function is to minimize this value. When the model's predictions perfectly align with the targets, the loss value is zero, indicating optimal model performance. We apply the binary cross-entropy loss (BCELoss) function to compute the disparity between the model's outputs and the actual labels.

Additionally, the AdamW [41] optimizer, which is a variant of the Adam [42] optimizer, introduces a weight decay term to control the regularization of model parameters. Weight decay helps prevent overfitting by encouraging the model to use simpler parameter settings, thereby enhancing generalization capabilities. It is important to note that we do not explicitly introduce a regularization term in the loss function. In our loss function, we calculate the difference between the model's outputs and the true labels using binary cross-entropy, and we apply the AdamW optimizer with weight decay to enhance the model's generalization without explicitly introducing a regularization term in the loss function, as follows:

$$\mathcal{L}oss = \frac{1}{2}(\mathcal{L}^s + \mathcal{L}^e), \tag{18}$$

where $\mathcal{L}^s$ denotes the start position classifier's loss and $\mathcal{L}^e$ denotes the end position classifier's loss. A sentence consists of N tokens, $\mathcal{L}$ is calculated as follows:

$$\mathcal{L}(Y, P) = -\frac{1}{N}\sum_{i=1}^{N}[\mathcal{Z}(y_i, p_i) + \mathcal{Z}(1 - y_i, 1 - p_i)], \tag{19}$$

$$\text{where } \mathcal{Z}(y, p) = y * log(p), \tag{20}$$

where log is the natural logarithm.

4. Experiments

4.1. Preparation

4.1.1. Dataset

We evaluated our proposed method on the following datasets: Resume [24], Youku [43], DuIE [44], and our own PPOST dataset.

The Resume dataset is a Chinese NER dataset focused on resumes. Youku provides an open NER dataset in the entertainment domain. DuIE is a large-scale manually annotated dataset, and all sentences in this dataset are extracted from Baidu Baike and Baidu News Search. The text in this dataset covers various domains found in real-world applications, including news, entertainment, and user-generated content. Due to the large size of the dataset, we randomly sampled a portion of the data, referred to as SubDuIE, for evaluation. Table 2 shows the details of these datasets.

Table 2. Statistics on the four dataset for the experiments.

Dataset	Resume	Youku	SubDuIE	PPOST [1]
Train	3821	8001	19,521	13,124
Eval	463	1000	1950	1396
Test	477	1001	1948	1395
Types	8	9	13	13
Avg [2]	33	18	66	83

[1] The PPOST dataset includes 9% of data generated using the data augmentation method; [2] "Avg" represents the average number of characters in the samples of this dataset.

The PPOST dataset consists of manually labeled entity data in the domain of Chinese paleontology popular science. The label distribution of the PPOST dataset is illustrated in the Figure 7. It can be seen that even after data augmentation processing, the phenomenon of a long-tail distribution still exists in practical production scenarios.

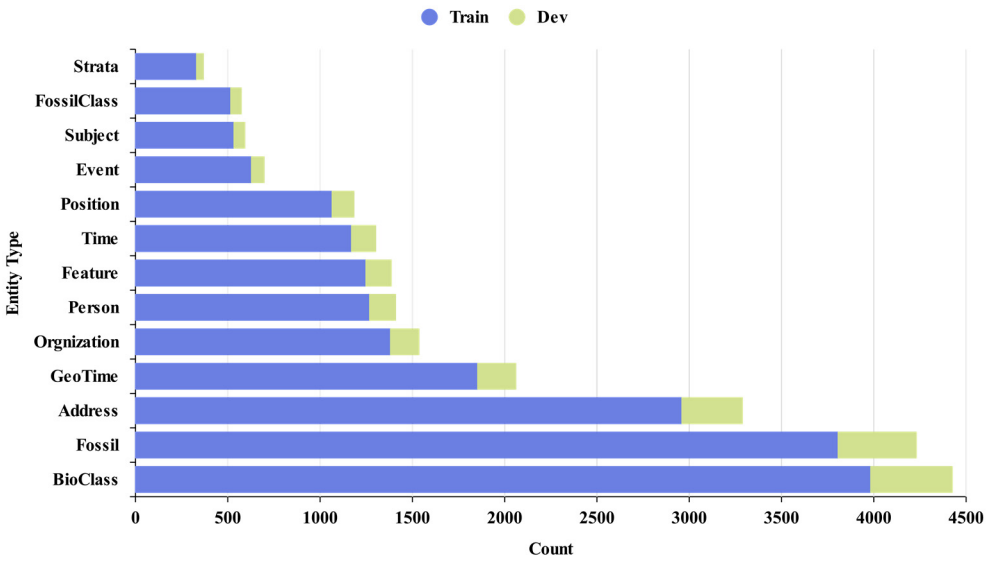


Figure 7. The label distribution of the PPOST dataset.

4.1.2. Evaluation Metrics

To evaluate the performance on Chinese NER, following most of the baselines, we use precision (P), recall (R) and F1-score (F1) as the metrics, they are computed on the entity-level number of true positives (TP), false positives (FP) and false negatives (FN):

$$Precision = \frac{TP}{TP + FP} * 100\%, \quad (21)$$

$$Recall = \frac{TP}{TP + FN} * 100\%, \tag{22}$$

$$F_1 = 2 * \frac{Precision \times Recall}{Precision + Recall} * 100\%, \tag{23}$$

where F_1 is obtained by directly averaging the F1 scores of all types, that is, Macro-F1 [45].

4.1.3. Baseline Methods

Aiming to measure and analyze the proposed method, we compare its performance on the four datasets mentioned above with mainstream models in recent years:

- Methods without lexicon features: BiLSTM-CRF [16], BERT [19], BERT-CRF [23], BERT-BiLSTM-CRF [46], BERT-GlobalPointer [37] and BERT-Pointer [36]. These models are based on deep learning, such as BiLSTM or BERT, using sequence labeling or pointer labeling frameworks. In particular, these models do not introduce Chinese lexicon-level features.
- Methods with lexicon features: lattice LSTM [24] and lexicon-enhanced BERT (LEBERT) [28]. Lattice LSTM can be seen as an extension of the character-based NER model, which adds words as input and additional gates to control information flow; LEBERT directly integrates external lexicon knowledge into the BERT layer through the lexicon adapter layer.

4.1.4. Training

In all experiments, we used the same BERT pre-trained model. Due to differences between datasets, we adjusted batch size and max_seq_len separately for each dataset, and all experiments were conducted on a single NVIDIA (Santa Clara, CA, USA) GeForce RTX 3070Ti GPU. For specific details and settings of other hyperparameters, please refer to Tables A1–A3 in Appendix A.

4.2. Results and Analysis

4.2.1. Comparative with Other Models

To validate the effectiveness of the proposed method in the Chinese NER, we applied an early stopping mechanism and calculated the average results over five runs of experiments. Tables 3 and 4 present the experimental results for the models mentioned in Section 4.1.3.

Table 3. Precision, recall, and F1-score statistics on the Resume and Youku datasets.

Model	Resume			Youku		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
BiLSTM-CRF	93.70	93.30	93.50 [47]	80.31	79.22	79.76
BERT	94.20	95.80	95.00 [19]	85.06	76.75	80.69
BERT-CRF	-	-	**96.87** [48]	83.00	81.70	82.40 [43]
BERT-BiLSTM-CRF	-	-	95.59 [49]	89.56	80.48	84.78
Lattice LSTM	94.81	94.11	94.46 [24]	84.43	81.28	82.82
LEBERT	-	-	96.08 [28]	91.93	82.10	86.74
BERT-GlobalPointer	96.94	93.71	95.30	92.76	82.02	87.06
BERT-Pointer	97.15	94.27	95.69 [27]	95.06	83.47	**88.89**
BERT-LPN	98.12	95.70	**96.90**	95.74	83.23	**89.05**

Significant bold values are the current SOTA in performance achieved, and the bold in the last row of these tables denotes the new achievements we obtained in our model.

Experimental results on two shorter-text datasets, Resume and Youku, and two longer-text datasets, SubDuIE and PPOST, show that our model improves F1-score by 0.03%, 0.16%, 1.27%, and 2.28%, respectively, compared to previous models. This demonstrates that our feature fusion strategy effectively utilizes Chinese lexicon-level information, enhances Chinese NER performance. However, the advantages of our method are not fully reflected in the Resume and Youku datasets, where the sample texts are shorter, and entities

are simpler. Moreover, the DTFW-lookup Table for the original corpus of these datasets was not calculated. The more significant improvement on longer-text datasets may be attributed to the use of the WordPiece algorithm, simplifying the model input and integrating Chinese lexicon-level features, enhancing the model's generalization performance and feature capture capability. Specifically, on the PPOST dataset, compared to sequence labeling models, the performance of pointer labeling models has significantly improved. This experimental result indicates that pointer labeling models can effectively improve the recognition performance of long and nested entities in the domain of Chinese paleontology popular science.

Table 4. Precision, recall, and F1-score statistics on the SubDuIE and PPOST datasets.

Model	SubDuIE			PPOST		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
BiLSTM-CRF	72.37	58.71	64.83	61.14	47.37	53.38
BERT	79.48	55.49	65.35	66.79	46.88	55.09
BERT-CRF	78.60	57.66	66.52	68.82	46.40	55.43
BERT-BiLSTM-CRF	76.65	56.71	65.19	67.78	46.22	54.96
Lattice LSTM	79.00	62.68	69.90	71.07	50.38	58.96
LEBERT	83.31	61.58	70.81	79.48	53.13	63.69
BERT-GlobalPointer	83.20	61.84	70.95	80.40	67.98	73.67
BERT-Pointer	83.67	61.81	**71.10**	81.50	71.60	**76.23**
BERT-LPN	84.47	63.30	**72.37**	83.50	74.07	**78.51**

4.2.2. Comparative with Various Entity Types

To comprehensively evaluate our model's recognition performance on various entities in our domain, we collected the performance of our model in identifying different entities on the PPOST dataset, as shown in Figure 8. The experimental results indicate that the recognition accuracy of various entities is mainly positively correlated with the number of labels in the dataset. This may be due to the scarcity of samples for certain types, leading to the model's inability to learn richer features of these type entities. Specifically, since we define the Feature type mostly for the attributes of things, such as the volume of fossils or the appearance of ancient organisms, and with diverse characteristics, the recognition performance for this part of entities is relatively lower. Therefore, it is more appropriate to consider conducting attribute extraction tasks for these entities.

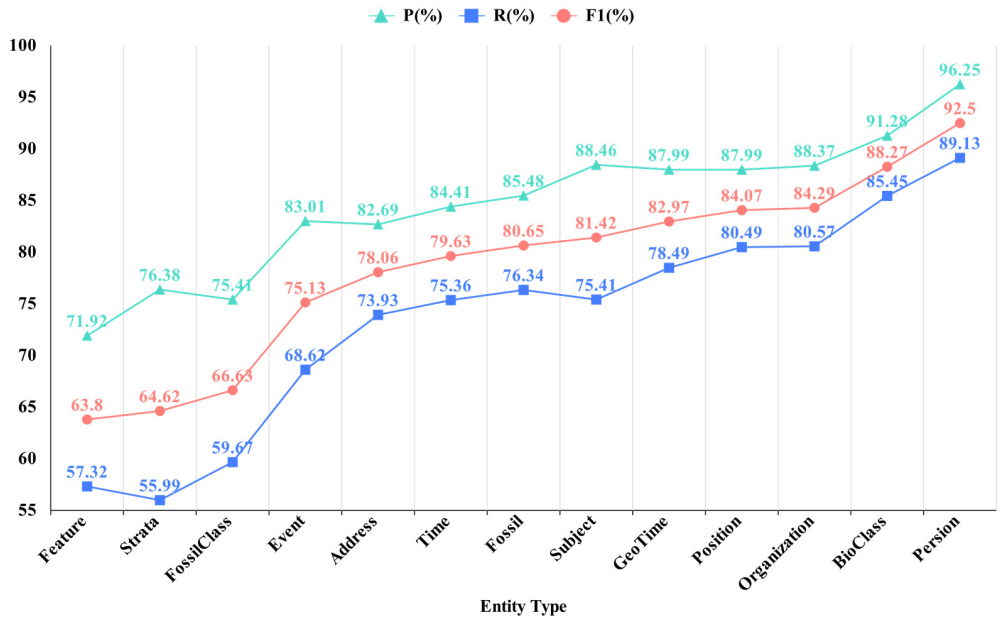


Figure 8. Precision, recall, and F1-score of 13 types statistics on the PPOST.

To evaluate the recognition performance of our model for long entities and nested entities, we sampled 324 instances of long entities with an entity length exceeding 8 characters and an average length of 10 characters and 213 instances of nested entities from the PPOST dataset. The relevant experimental results are shown in Table 5. The results indicate that compared to models based on CRF, models based on pointer labeling can effectively enhance the recognition performance of long and nested entities. Moreover, our BERT-LPN model achieves the best performance among the models mentioned.

Table 5. Precision, recall, and F1-score statistics of long entities and nested entities on the PPOST.

Model	Long Entities			Nested Entities		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
BiLSTM-CRF	43.36	39.76	41.48	47.68	40.01	43.50
BERT	46.90	41.47	44.02	45.19	37.31	40.87
BERT-CRF	48.43	41.29	44.58	42.17	35.31	38.44
BERT-BiLSTM-CRF	49.00	40.39	44.28	43.16	37.99	40.41
Lattice LSTM	51.49	45.62	48.38	44.11	41.01	42.50
LEBERT	52.04	45.84	48.74	45.80	42.61	44.15
BERT-GlobalPointer	57.42	53.73	55.51	65.80	56.76	60.95
BERT-Pointer	59.00	54.71	**56.77**	67.67	58.30	**62.64**
BERT-LPN	60.39	55.19	**57.67**	68.32	59.26	**63.47**

4.2.3. Comparative with Complexity of Various Models

Compared to neural networks, CRF models themselves have a relatively high time performance overhead [50], with an overall time complexity of $O(m * n^2)$, where m represents the length of the observed sequence, and n denotes the number of hidden states. To assess the computational requirements of our model, we only recorded the trainable parameters and training speed of the pointer labeling models, as shown in Table 6. The experimental results indicate that the additional time performance overhead of our method is negligible. This is because, through preprocessing, we obtained the DTFW-lookup Table, eliminating the need for repetitive computation of lexicon frequency weight during training, requiring only table lookup operations. Additionally, compared to the BERT-GlobalPointer model, we used only two fully connected layers for positional marking, reducing the parameters of fully connected layers and improving the training speed by 15%. In contrast to the BERT-Pointer model, we need to calculate the boundaries of lexicon and look up the DTFW-Lookup Table to obtain lexicon frequency weights. Moreover, the boundary embedding layer is trainable, resulting in an additional 1.7% performance overhead.

Table 6. Parameters and training speed statistics of the GlobalPointer, LPN, and Pointer models based on BERT on the PPOST dataset.

Model	Parameters	Speed
BERT-GlobalPointer	29.00 M	1× *
BERT-LPN	28.97 M	1.15× *
BERT-Pointer	28.94 M	1.17× *

* 1.15× and 1.17× represent 15% and 17% faster than 1×, respectively.

4.2.4. Comparative with Various Strategies

In order to investigate the contribution of each strategy to our approach, we conducted a series of ablation experiments on the SubDuIE and PPOST datasets, and the results are shown in Figure 9.

The ablation studies were designed as follows:

- Strategy 1: Strategy of adding only domain term frequency weight features without additional methods for feature fusion;

- Strategy 2: Strategy of adding only Chinese lexicon boundary features without additional methods for feature fusion;
- Strategy 3: Strategy of simultaneously adding Chinese lexicon boundary and domain term frequency weight features without additional methods for feature fusion;
- Strategy 4: Strategy of simultaneously adding Chinese lexicon boundary and domain term frequency weight features, using the BiLSTM for feature fusion.

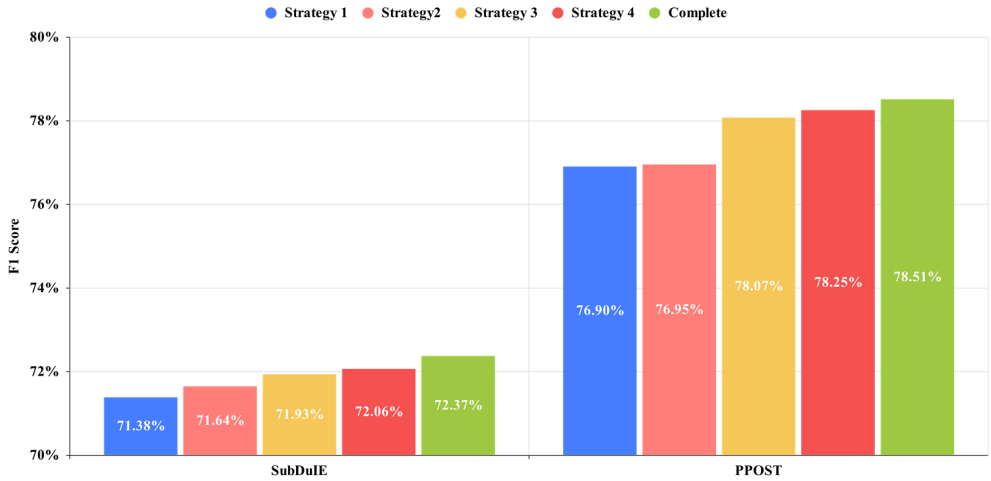


Figure 9. Performance comparison of different strategies on the SubDuIE and PPOST datasets in terms of F1-score. The Complete represents our proposed BERT-LPN model, simultaneously adding Chinese lexicon boundary and domain term frequency weight features and using the multi-head self-attention for feature fusion.

We remove any strategy that will lead to a decrease in entity recognition accuracy. Specifically, compared to the original pointer network, Strategy 1, Strategy 2, and Strategy 3 have all shown improvements on our experimental SubDuIE and PPOST datasets. Their F1-scores increased by 0.28%, 0.54%, 0.67% and 0.67%, 0.72%, 1.84%, respectively. This indicates that lexicon boundary features and domain term frequency weight features can effectively integrate Chinese lexicon-level features, improving the performance of the NER model. Compared to Strategy 3, Strategy 4 and the Complete model showed F1-score improvements of 0.13% and 0.44% and 0.18% and 0.44%, respectively, on the above two datasets. This suggests that the feature fusion of Chinese lexicon-level feature vectors can more effectively integrate contextual information features, enhancing the performance of the NER model. Compared to Strategy 4, the Complete model showed F1-score improvements of 0.31% and 0.26% on the above two datasets, respectively, indicating that the multi-head self-attention mechanism better handles dependencies between information.

4.2.5. Comparative with Various Data Augmentation Ratios

To verify the effectiveness of the data augmentation method, we conducted an additional set of experiments on the BERT-LPN model. The experiments were performed on the PPOST dataset, and the results are shown in Figure 10. The experimental results indicate that, on the PPOST dataset, our proposed data augmentation method has a certain performance improvement within the enhancement ratio range of (0, 0.1], where the ratio value is the ratio of the generated data to the original total data. This improvement may be attributed to the model further learning the ability to identify entities that do not exist in the input within this enhancement ratio range. However, when the ratio is more than 0.1, the model's performance sharply declines, and the performance is even worse than that of the original dataset. This may be due to an excess of augmented data in the training set, causing the model to preferentially predict the absence of entities in the input.

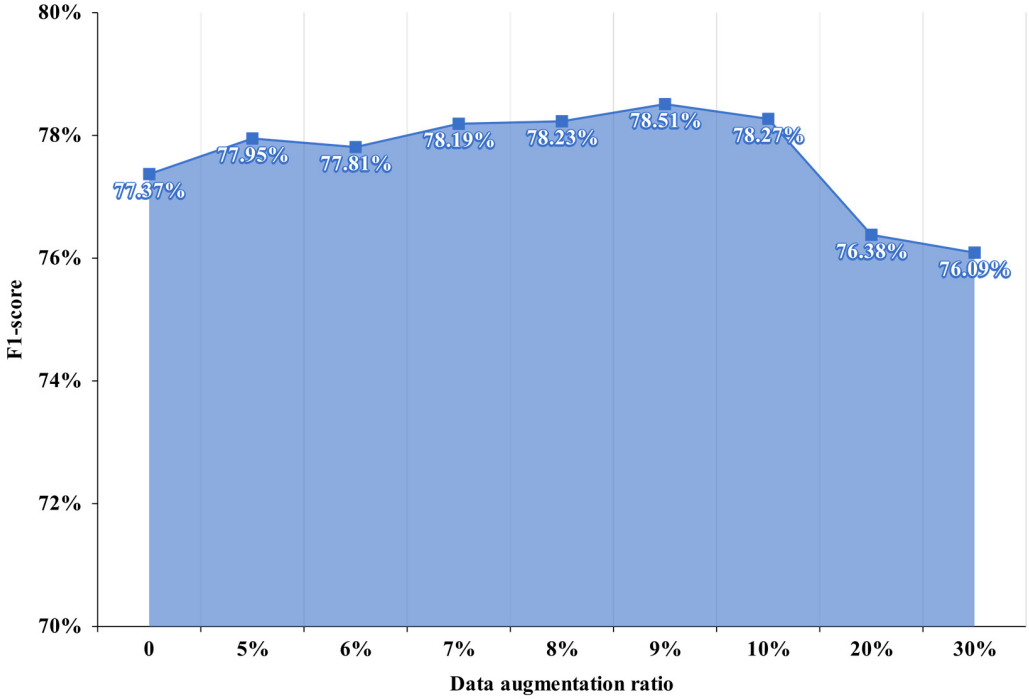


Figure 10. F1-score of the BERT-LPN at different data augmentation ratios on the PPOST dataset.

4.3. Extraction Results and Error Analysis

In this study, a set of experiments were conducted to validate the extracted NER results and analyze errors. Some examples of experimental results are shown in Table 7. From the experimental data with IDs 1 and 2, it can be observed that our model can effectively recognize nested entities, such as "寒武纪地层" (Cambrian strata), and long entities, like "古脊椎动物与古人类研究所" (Institute of Vertebrate Paleontology and Paleoanthropology). From the experimental data with IDs 3 and 4, it is evident that when the input data contains multiple entities of the same type (Address: "湖南" (Hunan), "江西" (Jiangxi), "浙江" (Zhejiang), "安徽" (Anhui)), our model, with the introduction of relative positional information, can effectively distinguish the boundaries of these four entities, reducing the likelihood of erroneously combining arbitrary start and end positions to form incorrect entities.

Additionally, we summarized and analyzed the model's recognition errors, identifying the following main issues:

1. We cannot accurately identify some continuous entity characters separated by the symbol "—". For example, experimental data with ID 5 indicates that in the entity "中—上奥陶统" (Middle—Upper Ordovician), "中—" (Middle—) is not considered part of the entity. This may be due to a limited number of samples for such entities or the negative impact of the "—" symbol on the recognition of contextual features;
2. We cannot accurately identify entities with uncommon terms as boundaries. For instance, experimental data with ID 6 shows that in the entity "獬豸盘角鹿" (Discokeryx xiezhi), the model fails to recognize "獬豸" (xiezhi) as part of the entity. This could be because "獬豸" (xiezhi) is an uncommon term, and its domain term frequency weight is relatively low, preventing the model from fully learning the features of uncommon lexicons.

Table 7. Illustrative examples of recognition results.

ID	Sentence	Ground Truth	Prediction
1	“寒武纪地层记录” (Record of Cambrian strata)	寒武纪 (Cambrian), 寒武纪地层 (Strata).	寒武纪 (Cambrian) √, 寒武纪地层 (Strata) √.
2	“中国科学院古脊椎动物与古人类研究所” (Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences)	中国科学院 (Chinese Academy of Sciences), 古脊椎动物与古人类研究所 (Institute of Vertebrate Paleontology and Paleoanthropology).	中国科学院 (Chinese Academy of Sciences) √, 古脊椎动物与古人类研究所 (Institute of Vertebrate Paleontology and Paleoanthropology)√.
3	“主要分布于我国湖南, 江西、浙江、安徽等地” (Mainly distributed in Hunan, Jiangxi, Zhejiang, Anhui and other places in China)	湖南 (Hunan), 江西 (Jiangxi), 浙江 (Zhejiang), 安徽 (Anhui).	湖南 (Hunan) √, 江西 (Jiangxi) √, 浙江 (Zhejiang) √, 安徽 (Anhui) √.
4	“恐龙位于爬行动物和鸟类之间的十字路口” (Dinosaurs were at the crossroads between reptiles and birds)	恐龙 (Dinosaurs), 爬行动物 (Reptiles), 鸟类 (Birds).	恐龙 (Dinosaurs) √, 爬行动物 (Reptiles) √, 鸟类 (Birds) √.
5	“我国华南地区中—上奥陶统黑色页岩” (Middle-Upper Ordovician black shale in South China)	华南地区 (South China), 中—上奥陶统 (Middle-Upper Ordovician).	华南地区 (South China) √, 上奥陶统 (Upper Ordovician)×.
6	“獬豸盘角鹿属于长颈鹿科” (Discokeryx xiezhi belongs to the family Giraffeidae)	獬豸盘角鹿 (Discokeryx xiezhi), 长颈鹿科 (Giraffeidae).	盘角鹿 (Discokeryx) ×, 长颈鹿科 (Giraffeidae) √.

“×” represents entity type or boundaries recognition errors. “√” represents complete and accurate recognition of entity type and boundaries.

5. Conclusions and Future Work

In the domain of Chinese paleontology popular science, NER is a fundamental step for extracting information and knowledge from a massive amount of popular science articles or books. In this study, we researched and compared mainstream Chinese deep learning-based NER methods to address the challenges in our domain. We constructed a corpus for the Chinese paleontology popular science domain and its annotated NER dataset, PPOST. We proposed a domain NER model, BERT-LPN, based on the BERT pre-trained language model. The model was applied to four different datasets, and its performance was evaluated in terms of precision, recall, and F1-score. The experimental results show that the proposed BERT-LPN method significantly outperforms baseline models and other deep learning models in both general domain and Chinese paleontology popular science texts.

The contributions of this research can be viewed from two perspectives. Methodologically, this study introduces a deep learning model named BERT-LPN. Experimental results demonstrate that, compared to other deep learning models, BERT-LPN performs better in extracting entities, providing multiple advantages for Chinese NER, including the following:

1. Domain-specific information: In the task of Chinese NER within a specific domain, lexicon-level features can encompass domain-specific lexicons, thereby enhancing the model’s performance within that domain;
2. Improved semantic understanding: Lexicon-level features can assist the model in better comprehending the semantic information within the text, capturing the meanings of lexicons and their contextual relationships. This helps the model in more accurately distinguishing between different entity types in longer texts.

From an application perspective, this study developed a Chinese paleontology popular science NER dataset, promoting the development of NER, popular science knowledge graphs, and popular science education in the Chinese paleontology domain. Moreover, in other NER tasks in different domains, researchers typically have easy access to original corpora, and our method relies only on the target domain's corpus, greatly increasing the possibility of generalizing our model to other domains. Future research will focus on the following three aspects:

- Corpus construction is an ongoing process, and we will further optimize the Chinese paleontology popular science corpus and PPOST dataset. Specifically, for the Feature entity type, it will be split into more specific attributes to further improve NER performance;
- Studying the performance of our model in other Chinese domains to promote the development of Chinese NER;
- Studying how to handle ambiguous entity boundaries and uncommon terms situations based on our approach.

Author Contributions: Conceptualization, Y.G. and F.L.; methodology, S.F. and H.L.; software, Y.G. and S.F.; validation, X.W., J.S. and Q.G.; formal analysis, Y.G. and S.F.; investigation, S.F. and H.L.; resources, W.L.; data curation, S.F. and H.L.; writing—original draft preparation, Y.G. and S.F.; writing—review and editing, Y.G. and F.L.; visualization, S.F.; supervision, W.L. and X.W.; project administration, W.L. and X.W.; funding acquisition, W.L. and X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by International Research Center of Big Data for Sustainable Development Goals, grant number CBAS2022GSP05, State Key Laboratory of Remote Sensing Science, grant number 6142A01210404, Hubei Key Laboratory of Intelligent Geo-Information Processing, grant number KLIGIP-2022-B03, and Metallogenic patterns and mineralization predictions for the Daping gold deposit in Yuanyang County, Yunnan Province, grant number 2022026821.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This manuscript utilizes publicly available datasets, which can be accessed at the following links: Resume dataset (https://tianchi.aliyun.com/dataset/144345 (accessed on 17 July 2023)), Youku dataset (https://github.com/allanj/ner_incomplete_annotation (accessed on 10 August 2023)), and DuIE dataset (https://www.luge.ai/#/luge/dataDetail?id=5 (accessed on 15 September 2023)). Restrictions apply to the availability of the PPOST data, and the data are available from the authors.

Acknowledgments: We gratefully acknowledge the support of the School of Computer Science, China University of Geosciences, Wuhan, and the School of Geography and Information Engineering, China University of Geosciences, Wuhan. We also thank the Piesat Information Technology Co., Ltd., and the China Geological Survey.

Conflicts of Interest: Author Xianbin Wang was employed by the company Piesat Information Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A

There are some details about the experimental process using systems, such as BERT, embedding layer, and multi-head self-attention, that require fine-tuning of the hyperparameters, and some details about this process are given in the following Tables A1 and A2. Some hardware and software environments for the experiments are given in the following Table A3.

Table A1. Some common hyperparameter settings about the experiments.

Hyperparameter	Value
Data Augmentation Ratio	0.09
Epoch	20
BERT Model	bert-base-chinese
BERT LR	1×10^{-5}
BERT Dropout Rate	0.35
Linear LR	3×10^{-4}
Optimizer	AdamW
Boundary Embedding Dim	16
Multi-head self-attention head	8
Threshold	0.6

Table A2. Some hyperparameter settings about the experiments on the four datasets.

Dataset	max_seq_len	Batch Size
Resume	64	64
Youku	64	64
SubDuIE	256	32
PPOST	256	32

Table A3. Some hardware and software environments about the experiments.

Environment	Value
Processor	12th Gen Intel(R) Core(TM) i5-12600KF
RAM	32.0 GB
GPU	NVIDIA GeForce RTX 3070Ti GPU 8GB
Python Version	3.8.16
PyTorch Version	2.0.0

References

1. Tao, D.; Yang, P.; Feng, H. Utilization of text mining as a big data analysis tool for food science and nutrition. *Compr. Rev. Food Sci. Food Saf.* **2020**, *19*, 875–894. [CrossRef] [PubMed]
2. Singh, S. Natural language processing for information extraction. *arXiv* **2018**, arXiv:1807.02383.
3. Contributors, W. Popular Science—Wikipedia, the Free Encyclopedia. 2023. Available online: https://en.wikipedia.org/wiki/Popular_science (accessed on 1 July 2023).
4. Zhai, X. Research on Tourism Promotion of Shandong Zhucheng Dinosaur National Paleontologic Geopark. In Proceedings of the 2015 International Conference on Education, Management and Computing Technology, Tianjin, China, 13–14 June 2015; Atlantis Press: Amsterdam, The Netherlands, 2015.
5. Mansouri, A.; Affendey, L.S.; Mamat, A. Named entity recognition approaches. *Int. J. Comput. Sci. Netw. Secur.* **2008**, *8*, 339–344.
6. Ye, J.; Zhou, J.; Tian, J.; Wang, R.; Zhou, J.; Gui, T.; Zhang, Q.; Huang, X. Sentiment-aware multimodal pre-training for multimodal sentiment analysis. *Knowl. Based Syst.* **2022**, *258*, 110021. [CrossRef]
7. Chennafi, M.E.; Bedlaoui, H.; Dahou, A.; Al-Qaness, M.A.A. Arabic Aspect-Based Sentiment Classification Using Seq2Seq Dialect Normalization and Transformers. *Knowledge* **2022**, *2*, 388–401. [CrossRef]
8. Saha, S.K.; Chatterji, S.; Dandapat, S.; Sarkar, S.; Mitra, P. A hybrid approach for named entity recognition in indian languages. In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian languages, Hyderabad, India, 12 January 2008.
9. Tanabe, L.; Xie, N.; Thom, L.H.; Matten, W.; Wilbur, W.J. GENETAG: A tagged corpus for gene/protein named entity recognition. *BMC Bioinform.* **2005**, *6*, S3. [CrossRef] [PubMed]
10. Ju, Z.; Wang, J.; Zhu, F. Named entity recognition from biomedical text using SVM. In Proceedings of the 2011 5th International Conference on Bioinformatics and Biomedical Engineering, Wuhan, China, 10–12 May 2011.
11. Morwal, S.; Jahan, N.; Chopra, D. Named entity recognition using hidden Markov model (HMM). *Int. J. Nat. Lang. Comput. (IJNLC)* **2012**, *1*. [CrossRef]
12. Zhou, G.; Su, J. Named entity recognition using an HMM-based chunk tagger. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002.
13. Konkol, M.; Konopík, M. CRF-Based Czech Named Entity Recognizer and Consolidation of Czech NER Research. In *International Conference on Text, Speech and Dialogue*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 153–160.

14. Khan, A.; Sohail, A.; Zahoora, U.; Qureshi, A.S. A survey of the recent architectures of deep convolutional neural networks. *Artif. Intell. Rev.* **2020**, *53*, 5455–5516. [CrossRef]
15. Staudemeyer, R.C.; Morris, E.R. Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. *arXiv* **2019**, arXiv:1909.09586.
16. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
17. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Philip, S.Y. A comprehensive survey on graph neural networks. *IEEE Trans-Actions Neural Netw. Learn. Syst.* **2020**, *32*, 4–24. [CrossRef] [PubMed]
18. Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; Huang, X. Pre-trained models for natural language processing: A survey. *Sci. China Technol. Sci.* **2020**, *63*, 1872–1897. [CrossRef]
19. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
20. Darji, H.; Mitrović, J.; Granitzer, M. German BERT Model for Legal Named Entity Recognition. *arXiv* **2023**, arXiv:2303.05388.
21. Souza, F.; Nogueira, R.; Lotufo, R. Portuguese named entity recognition using BERT-CRF. *arXiv* **2019**, arXiv:1909.10649.
22. Song, Z.; Xu, W.; Liu, Z.; Chen, L.; Su, H. A BERT-Based Named Entity Recognition Method of Warm Disease in Traditional Chinese Medicine. In Proceedings of the 2023 IEEE 18th Conference on Industrial Electronics and Applications (ICIEA), Ningbo, China, 18–22 August 2023; pp. 1226–1231.
23. Dai, Z.; Wang, X.; Ni, P.; Li, Y.; Li, G.; Bai, X. Named Entity Recognition Using BERT BiLSTM CRF for Chinese Electronic Health Records. In Proceedings of the 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Suzhou, China, 19–21 October 2019; pp. 1–5.
24. Zhang, Y.; Yang, J. Chinese NER Using Lattice LSTM. *arXiv* **2018**, arXiv:1805.02023.
25. Li, X.; Yan, H.; Qiu, X.; Huang, X.-J. FLAT: Chinese NER Using Flat-Lattice Transformer. *arXiv* **2020**, arXiv:2004.11795.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; Curran Associates Inc.: Long Beach, CA, USA, 2017.
27. Guo, Q.; Guo, Y. Lexicon enhanced Chinese named entity recognition with pointer network. *Neural Comput. Appl.* **2022**, *34*, 14535–14555. [CrossRef]
28. Liu, W.; Fu, X.; Zhang, Y.; Xiao, W. Lexicon Enhanced Chinese Sequence Labeling Using BERT Adapter. *arXiv* **2021**, arXiv:2105.07148.
29. Sun, Y.; Zheng, Y.; Hao, C.; Qiu, H. NSP-BERT: A Prompt-based Few-Shot Learner Through an Original Pre-training Task--Next Sentence Prediction. *arXiv* **2021**, arXiv:2109.03564.
30. Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* **2023**, *55*, 1–35. [CrossRef]
31. Cui, L.; Wu, Y.; Liu, J.; Yang, S.; Zhang, Y. Template-Based Named Entity Recognition Using BART. *arXiv* **2021**, arXiv:2106.01760.
32. Huang, Y.; He, K.; Wang, Y.; Zhang, X.; Gong, T.; Mao, R.; Li, C. Copner: Contrastive learning with prompt guiding for few-shot named entity recognition. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022.
33. Ding, N.; Chen, Y.; Han, X.; Xu, G.; Wang, X.; Xie, P.; Zheng, H.; Liu, Z.; Li, J.; Kim, H.-G. Prompt-learning for Fine-grained Entity Typing. *arXiv* **2021**, arXiv:2108.10604.
34. Li, X.; Feng, J.; Meng, Y.; Han, Q.; Wu, F.; Li, J. A Unified MRC Framework for Named Entity Recognition. *arXiv* **2019**, arXiv:1910.11476.
35. Lu, Y.; Liu, Q.; Dai, D.; Xiao, X.; Lin, H.; Han, X.; Sun, L.; Wu, H. Unified Structure Generation for Universal Information Extraction. *arXiv* **2022**, arXiv:2203.12277.
36. Gong, O. Chinese Information Extraction Using Pointer Network, in GitHub Repository. 2022. Available online: https://github.com/taishan1994/PointerNet_Chinese_Information_Extraction (accessed on 10 December 2023).
37. Su, J.; Murtadha, A.; Pan, S.; Hou, J.; Sun, J.; Huang, W.; Wen, B.; Liu, Y. Global Pointer: Novel Efficient Span-based Approach for Named Entity Recognition. *arXiv* **2022**, arXiv:2208.03054.
38. Song, X.; Salcianu, A.; Song, Y.; Dopson, D.; Zhou, D. Fast WordPiece Tokenization. *arXiv* **2020**, arXiv:2012.15524.
39. Rajaraman, A.; Ullman, J.D. *Mining of Massive Datasets*; Cambridge University Press: Cambridge, UK, 2011.
40. Luo, R.; Xu, J.; Zhang, Y.; Zhang, Z.; Ren, X.; Sun, X. Pkuseg: A toolkit for multi-domain chinese word segmentation. *arXiv* **2019**, arXiv:1906.11455.
41. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
42. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
43. Jie, Z.; Xie, P.; Lu, W.; Ding, R.; Li, L. Better modeling of incomplete annotations for named entity recognition. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Long and Short Papers. Volume 1.
44. Li, S.; He, W.; Shi, Y.; Jiang, W.; Liang, H.; Jiang, Y.; Zhang, Y.; Lyu, Y.; Zhu, Y. Duie: A large-scale chinese dataset for in-formation extraction. In Proceedings of the Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, 9–14 October 2019. Proceedings, Part II 8.

45. Lewis, D.D.; Schapire, R.E.; Callan, J.P.; Papka, R. Training algorithms for linear text classifiers. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, 18–22 August 1996.
46. Tang, X.; Huang, Y.; Xia, M.; Long, C. A Multi-Task BERT-BiLSTM-AM-CRF Strategy for Chinese Named Entity Recognition. *Neural Process. Lett.* **2023**, *55*, 1209–1229. [CrossRef]
47. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural Architectures for Named Entity Recognition. *arXiv* **2016**, arXiv:1603.01360.
48. Wang, X.; Jiang, Y.; Bach, N.; Wang, T.; Huang, Z.; Huang, F.; Tu, K. Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning. *arXiv* **2021**, arXiv:2105.03654.
49. Zhao, J.; Cui, M.; Gao, X.; Yan, S.; Ni, Q. Chinese Named Entity Recognition Based on BERT and Lexicon Enhancement. In Proceedings of the RICAI 2022: 2022 4th International Conference on Robotics, Intelligent Control and Artificial Intelligence, Dongguan China, 16–18 December 2022.
50. Strubell, E.; Verga, P.; Belanger, D.; McCallum, A. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions. *arXiv* **2017**, arXiv:1702.02098.