energies

MDPI

Article

A Named Entity Recognition Model for Chinese Electricity Violation Descriptions Based on Word-Character Fusion and Multi-Head Attention Mechanisms

Lingwen Meng [1], Yulin Wang [2], Yuanjun Huang [3], Dingli Ma [3], Xinshan Zhu [2] and Shumei Zhang [2,*]

[1] Electric Power Research Institute of Guizhou Power Grid Co., Ltd., Guiyang 550002, China; menglw000@163.com
[2] School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China; yulin_2000@tju.edu.cn (Y.W.); xszhu@tju.edu.cn (X.Z.)
[3] Southern Power Grid Digital Grid Group Co., Ltd., Guizhou Branch, Guiyang 550002, China; 18198350949@139.com (Y.H.); 13765842161@139.com (D.M.)
* Correspondence: shumeizhang@tju.edu.cn

Abstract: Due to the complexity and technicality of named entity recognition (NER) in the power grid field, existing methods are ineffective at identifying specialized terms in power grid operation record texts. Therefore, this paper proposes a Chinese power violation description entity recognition model based on word-character fusion and multi-head attention mechanisms. The model first utilizes a collected power grid domain corpus to train a Word2Vec model, which produces static word vector representations. These static word vectors are then integrated with the dynamic character vector features of the input text generated by the BERT model, thereby mitigating the impact of segmentation errors on the NER model and enhancing the model's ability to identify entity boundaries. The combined vectors are subsequently input into a BiGRU model for learning contextual features. The output from the BiGRU layer is then passed to an attention mechanism layer to obtain enhanced semantic features, which highlight key semantics and improve the model's contextual understanding ability. Finally, the CRF layer decodes the output to generate the globally optimal label sequence with the highest probability. Experimental results on the constructed power grid field operation violation description dataset demonstrate that the proposed NER model outperforms the traditional BERT-BiLSTM-CRF model, with an average improvement of 1.58% in precision, recall, and F1-score. This demonstrates the effectiveness of the model design and further enhances the accuracy of entity recognition in the power grid domain.

Keywords: named entity recognition (NER); attention mechanism; electric grid on-site operation violation description; word-character fusion

check for updates

Academic Editor: Juan-José González de la Rosa

Received: 9 December 2024
Revised: 13 January 2025
Accepted: 16 January 2025
Published: 17 January 2025

Citation: Meng, L.; Wang, Y.; Huang, Y.; Ma, D.; Zhu, X.; Zhang, S. A Named Entity Recognition Model for Chinese Electricity Violation Descriptions Based on Word-Character Fusion and Multi-Head Attention Mechanisms. *Energies* **2025**, *18*, 401. https://doi.org/10.3390/en18020401

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The power industry is a core pillar of national energy infrastructure, closely tied to the lifeblood of national development and crucial to future revitalization. With the advancement of computing and artificial intelligence, countries around the world are promoting the intelligent and digital development of the power industry, leading to the generation of vast amounts of data that are available for analysis and mining. Power Named Entity Recognition (NER) refers to the task of extracting and categorizing entities with specific meanings, such as names of people, places, equipment, and other relevant entities, from

unstructured power-related data [1]. Similar to other domains, NER in the power sector faces the aforementioned challenges. Moreover, relying solely on manual extraction based on expert knowledge to identify valuable information from unstructured data is not only labor-intensive and time-consuming but also prone to inaccuracies due to subjective factors and the lack of experience. Consequently, the accuracy of the extracted information cannot be guaranteed. Therefore, research into NER for power industry data is of critical importance.

Natural Language Processing (NLP) is a subfield of computer science and artificial intelligence aimed at enabling computers to understand, interpret, and generate natural language. It integrates techniques from linguistics, computer science, and machine learning to handle various language-related tasks [2]. Named Entity Recognition (NER) is a key technique in NLP, and used to identify entities in text that have specific meanings, typically including names of people, places, organizations, dates, times, currencies, and other such information. The goal of NER is to extract these entities from text and classify them, thus aiding in the understanding of the structure and content of the text [3]. NER plays an important role as a fundamental task in the field of information extraction, with significant applications in information retrieval [4], machine translation [5], intelligent question answering [6], and text classification [7].

Compared to English Named Entity Recognition (NER), Chinese NER research is more challenging. English words naturally have spaces and word boundaries, which facilitate entity recognition. In contrast, Chinese requires boundary determination for entities, and the language features phenomena such as homophones, polysemy, and historical and contemporary semantic shifts [8]. For example, in the sentence “南京市长江大桥” (“Nanjing Yangtze River Bridge”), depending on the context, it can be parsed as either “南京|市长|江大桥” (“Nanjing|Mayor|Yangtze Bridge”) or “南京市|长江|大桥” (“Nanjing City|Yangtze|Bridge”). Such ambiguities make Chinese NER more difficult. Existing NER techniques can be broadly classified into three categories: rule- and dictionary-based methods, statistical machine learning-based methods, and deep learning-based methods. Deep learning methods have become the mainstream and widely applied approach in Named Entity Recognition (NER) due to their remarkable performance. Particularly in the context of domain-specific NER, there are several challenges, such as the presence of a large number of specialized terms, the necessity for researchers to possess extensive domain knowledge, and the lack of high-quality annotated data. These factors make NER in specialized domains require more refined models and methods, as well as greater support in terms of domain knowledge and data, to enhance the accuracy and robustness of the recognition process.

Based on the issues and challenges of Named Entity Recognition (NER) in the power domain outlined above, this paper proposes a Chinese electrical power violation description entity recognition model based on dictionaries and attention mechanisms: the WBERT-BiGRU-MHA-CRF model. This model builds upon the BERT-BiGRU-CRF framework, further incorporating word and character-level information from the power violation description domain and introducing a multi-head attention mechanism. The multi-head attention mechanism captures different contextual information in parallel, enhancing the model’s expressive power and its ability to handle long-distance dependencies, thus improving both accuracy and robustness. First, the BERT model is fine-tuned based on the collected power grid violation description data. This fine-tuning enhances the model’s understanding of domain-specific terminology and context, thereby improving its performance in NER tasks. Next, a specialized dictionary is built from the same data, and a Word2Vec model is used to generate static word embeddings. These embeddings are fused with the dynamic character embeddings produced by BERT and input into the BiGRU model, allowing the model

to capture richer semantic information to help identify the boundaries of power-related entities. Afterwards, a multi-head attention mechanism module is added after the BiGRU model to enhance the ability to capture contextual information, improving the precision of feature representations and boosting the model's overall performance. Finally, the output of the attention mechanism is fed into a CRF model to obtain the globally optimal entity sequence recognized by the model.

2. Related Work

The development of Named Entity Recognition (NER) has evolved from early rule-based and dictionary-based methods to statistical machine learning approaches, and later to deep learning-based methods. Early Chinese NER research primarily focused on automatically identifying entities such as names of people, places, and organizations in Chinese texts through the formulation of specific rules [9]. For example, Shaalan et al. [10] optimized Arabic NER by combining contextual features with a place name dictionary and applying a filtering mechanism to handle incorrectly formatted named entities. With the advancement of statistical machine learning, some researchers began incorporating these methods into Chinese NER tasks. The process generally involves data collection and annotation, followed by feature extraction from the labeled text data. Hidden Markov Models (HMM) [11] and Conditional Random Fields (CRF) [12] have been widely applied in NER research. For example, Wu et al. [13] proposed a hybrid model that combines Conditional Random Fields (CRFs) and Dynamic Gazetteers (DGs) for the task of Chinese Named Entity Recognition (NER), where the DGs can dynamically add new entity information during the process of named entity recognition.

With the rapid development of deep learning technologies, traditional neural network-based methods have shown significant advantages in Named Entity Recognition (NER) tasks. Collobert et al. [14] first introduced the CNN-CRF model for NER. Huang et al. [15] incorporated BiLSTM into NER and proposed the classic BiLSTM-CRF model, which has become one of the well-known baseline models. In this model, the input to the BiLSTM network is the word embedding feature, while the CRF layer enforces constraints on the entity labels at the semantic level, decoding the corresponding labels. Vaswani et al. [16] proposed the Transformer model, which captures global features all at once, revolutionizing NER research with its capability to handle long-range dependencies effectively. In 2018, Google introduced BERT [17], a pre-trained model using a bidirectional Transformer architecture, outperforming the single-layer Transformer model. Souza et al. [18] applied the BERT-CRF model to NER tasks, leveraging BERT's transfer learning ability and the CRF model's structured prediction capabilities, achieving promising results. Dai et al. [19], in their research on NER for electronic health records, compared Word2Vec and BERT embeddings on top of the BiLSTM-CRF model, ultimately proving the effectiveness of the BERT-BiLSTM-CRF model. For Chinese NER, Cui et al. [20] improved the BERT model using full-word masking techniques and pre-trained it on Chinese Wikipedia data. Experimental results showed that the BERT-wwm model, which employs full-word masking, significantly enhanced performance. In recent years, scholars have been increasingly interested in exploring the positive impact of integrating popular techniques, such as transfer learning and attention mechanisms, into neural network structures for improving NER tasks. These innovations have further advanced the accuracy and robustness of NER models, especially in specialized fields like the power industry, where domain-specific entities and complex contexts require more sophisticated handling.

However, Chinese power data texts often exhibit complex linguistic structures and contain a large number of specialized terms and technical jargon. Directly using general-purpose dictionaries for Chinese word segmentation in power data texts often fails to

correctly identify the entity boundaries of these specialized terms. As a result, many researchers have attempted to incorporate external information through dictionaries and combined it with deep neural networks to enhance Named Entity Recognition (NER) for unstructured power data texts. Zhang et al. [21] enhanced the LSTM model by designing a lexical processing unit, connecting the character units at the beginning and end of words. This improved the model's ability to recognize entity boundaries, avoiding the impact of segmentation errors and effectively improving the performance of NER. This approach became one of the baseline models for Chinese entity recognition that integrates lexical information. Cong et al. [22], addressing the challenges in NER for power data, designed a power data NER model that incorporated lexical enhancement and global pointers. This model fused lexical information into each character embedding representation, which was then input into a pre-trained RoBERTa model to further extract features, followed by entity recognition using a global pointer mechanism. In addition to integrating domain-specific vocabulary, new techniques such as attention mechanisms have also provided researchers with valuable directions for improving NER in the power sector. Feng et al. [23] used an end-to-end deep learning model, the BERT-BiLSTM-CRF model, to achieve effective extraction of named entities from regulations, reports, and other related materials in the electric power domain. Du et al. [24] proposed a knowledge graph construction technique for power grid fault handling based on an attention mechanism integrated with a BiLSTM-CRF model. This knowledge graph enables efficient fault information filtering, supports intelligent decision-making, and enhances the efficiency of power grid fault repair operations. He et al. [25] proposed an asymptotic multi-type feature fusion model for recognizing Chinese named entities in unstructured maintenance records of electric power primary equipment. The effectiveness of the proposed approach was validated using electric power maintenance datasets as well as public datasets, including the MSRA dataset and the Chinese People's Daily corpus. Chen et al. [26] proposed the PBERTBiLC method for entity recognition in unstructured power text information. Based on the BERT-BiLSTM-CRF model, this method integrates self-attention mechanisms to fuse contextual semantic information from power texts, achieving high accuracy in entity recognition. These efforts highlight the continuous advancements in improving NER for power data by integrating specialized knowledge, attention mechanisms, and transformer-based models, which have collectively contributed to more accurate and efficient entity recognition in the power domain.

3. Model Framework

The model proposed in this paper for Named Entity Recognition (NER) in Chinese power grid violation descriptions, based on word-character fusion and multi-head attention mechanisms, is shown in Figure 1. It consists of four layers: the word-character fusion feature embedding layer, the feature embedding encoding layer, the multi-head attention mechanism layer, and the label decoding layer. The model architecture is depicted in Figure 1.

The word2vec embeddings for electrical power vocabulary, pre-trained on a corpus of Chinese electrical power violation descriptions, are kept frozen during the training of the named entity recognition model. In contrast, all other layers are continuously optimized as part of the training process.

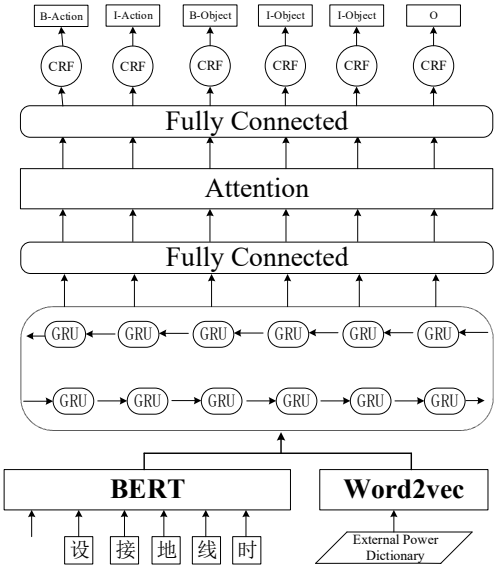


Figure 1. Flowchart of Named Entity Recognition for electric grid violation descriptions.

3.1. Word-Character Fusion Feature Embedding Layer

To address the issue that general Named Entity Recognition (NER) algorithms cannot accurately recognize domain-specific terms or specialized vocabulary in the power sector, this paper employs a dictionary matching method to extract lexical information from the text. This approach provides the model with additional semantic and entity boundary information, thereby improving the effectiveness of the NER task.

Step 1. Character Embedding Based on the BERT Model

The input text is first passed through the BERT pre-trained model to generate dynamic character vector representations of the text sequence. The structure of the BERT model is shown in Figure 2.

The input vectors for BERT consist of three main components: token embeddings, segment embeddings, and position embeddings. The '[CLS]' token is a special marker used for classification, typically placed at the beginning of the sequence, while the '[SEP]' token is used to separate sentences, usually placed at the end of the sentence. An example of the input ("装设接地线时") is shown in Figure 2.

Step 2. Lexical Embedding Based on the Word2Vec Model and Power Dictionary

In addition to obtaining the dynamic character vector representation of the input text sequence through the BERT model, a power grid field operation violation description corpus and a power dictionary were also constructed. The power dictionary can serve as a supplement to the general vocabulary of the BERT model.

Word2vec [27] is a word embedding model proposed by Google in 2013. The Word2Vec model includes two methods for training word embeddings: the Continuous Bag of Words (CBOW) model and the Skip-Gram model. In the CBOW model, the goal is to predict the center word based on several surrounding words. In contrast, the Skip-Gram model works in the opposite direction; it uses the center word as the starting point to predict and generate the surrounding context words. The structure of the Word2Vec model is shown in Figure 3, which includes the input layer, projection layer, and output layer.

In terms of training accuracy, the Skip-Gram model provides a more accurate representation of the semantic information in word vectors, yielding better results compared to the CBOW model. Moreover, the Skip-Gram model is better suited for handling small corpora and low-frequency words. Therefore, this paper employs the Skip-Gram model for training word vectors on the power grid field operation violation description text.

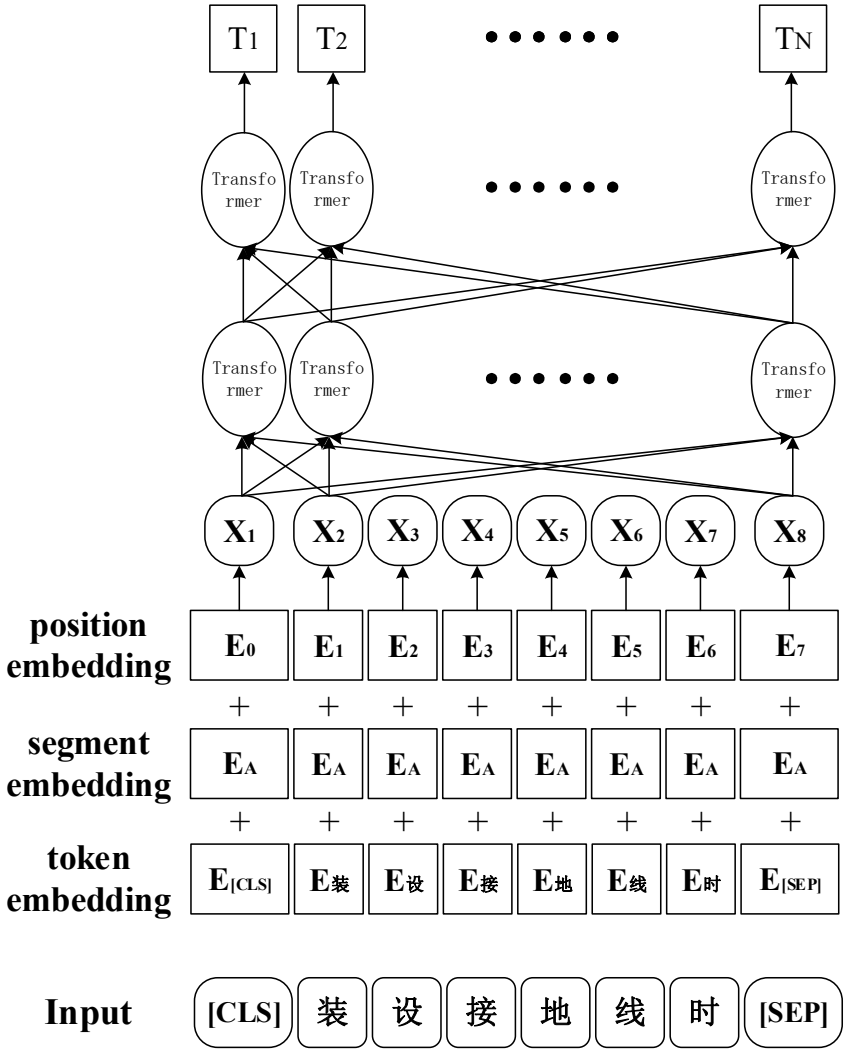


Figure 2. Architecture and examples of input vectors in the BERT model.

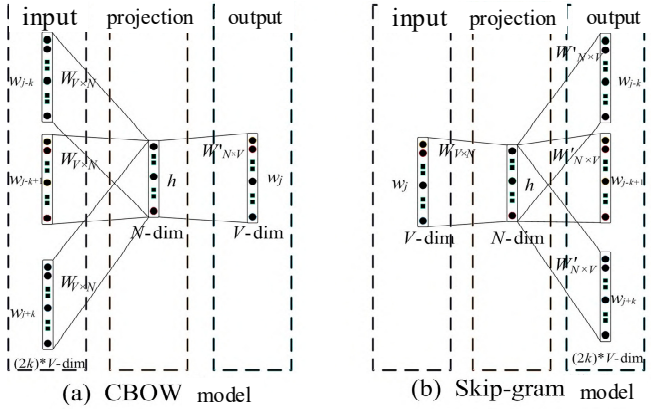


Figure 3. Architecture diagram of Word2Vec model.

The objective of the Skip-Gram model is to maximize the log-likelihood function, which is optimized using the stochastic gradient descent (SGD) algorithm. After training, the resulting matrix W' corresponds to low-dimensional dense word vectors of size $N \times V$, where V is the vocabulary size. Therefore, the essence of the Skip-Gram model is to apply softmax normalization to the cosine similarity between the input word's input vector

and the target word's output vector. The optimization objective function of the model is as tfollows:

$$L = \sum_{\omega_j \in V} \lg P(\omega_{j-k}, \omega_{j-k+1}, \cdots, \omega_{j+k} \Big| \omega_j) \tag{1}$$

where V represents the vocabulary of the corpus in which ω_j is located.

In the Skip-Gram model, the input layer only takes a given word ω_j, using this word as the center word. With a window size k, the probability of the $2k$ most likely context words is given by the following:

$$P(\omega_{j-k}, \omega_{j-k+1}, \cdots, \omega_{j+k} | \omega_j) = \frac{\exp(u_j^T h)}{\sum_i \exp(u_i^T h)} \tag{2}$$

where h is the column vector of the input layer matrix, representing the input vector of the center word; u_j is the row vector of the output layer matrix, representing the output vector of ω_j; and u_j^T is the weight used in the weighted summation. Since the probabilities of all words need to be calculated, $j = 1, 2, \cdots, V$, corresponding to a series of weights $u_1^T, u_2^T, \cdots, u_V^T$.

3.2. Feature Embedding Encoding Layer

The feature embedding encoding layer utilizes the BiGRU model to extract contextual features. Both GRU and LSTM are improvements upon RNNs, and they control the flow of information through different gating mechanisms, thereby mitigating the issues of vanishing and exploding gradients. The LSTM model consists of three gates: the forget gate, the input gate, and the output gate. In contrast, the GRU model has only two gates, as shown in Figure 4. The GRU integrates the input and forget gates of LSTM into a single update gate (denoted as z_t in Figure 4), and its other gate is the reset gate (denoted as r_t in Figure 4). As such, the GRU model has a simpler network structure, fewer parameters, and is easier to train, significantly improving training efficiency. Moreover, the performance of entity recognition models based on GRU is comparable to that of LSTM models.

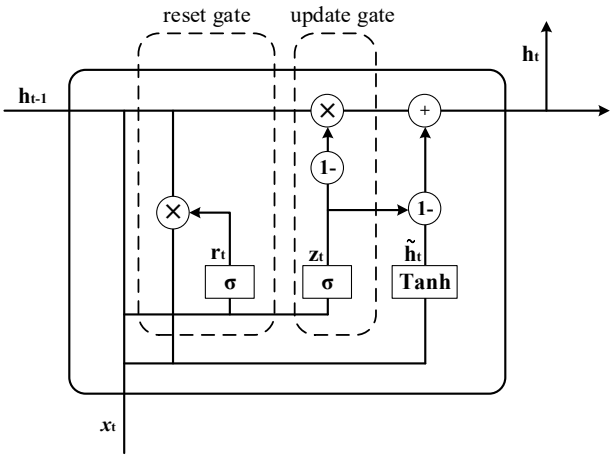


Figure 4. Internal structure diagram of GRU model.

The main equations of the GRU model are the following:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{3}$$

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{4}$$

$$\widetilde{h}_t = \tanh(W_{\widetilde{h}} \cdot [r_t * h_{t-1}, x_t]) \tag{5}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \widetilde{h} \tag{6}$$

where x_t represents the input embedding vector at time step t in the text sequence, h_t denotes the hidden state output at time step t, $\widetilde{h}$ is the candidate hidden state at time step t, and h_{t-1} is the hidden state output from the previous time step. $W_r, W_z, W_{\widetilde{h}}$ are the weight matrices. σ is the Sigmoid activation function, which maps the outputs of the reset and update gates to the range [0,1], and tanh is the hyperbolic tangent activation function, used to calculate the candidate hidden state $\widetilde{h}_t$. The notation [] indicates concatenation of two vectors, $*$ represents element-wise matrix multiplication, and $+$ represents matrix addition.

The single-layer GRU model outputs the state in a forward direction, which may not fully capture the contextual information. Therefore, this paper adopts a bidirectional GRU network model. BiGRU (Bidirectional Gate Recurrent Unit) utilizes both forward and backward GRU layers to extract contextual semantic information. The outputs are then weighted and summed, followed by a linear layer to map the d-dimensional vector to an m-dimensional vector. This process results in the final output label vector list $H = \{h_1, h_2, \cdots\cdots, h_n\}, H \in R^{n \times m}$, where n is the length of the text sequence and m is the number of entity type labels.

The relevant equations for the BiGRU modeling process are as follows:

$$\overrightarrow{h} = GRU(x_t, \overrightarrow{h}_{t-1}) \tag{7}$$

$$\overleftarrow{h}_t = GRU(x_t, \overleftarrow{h}_{t-1}) \tag{8}$$

$$h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t] \tag{9}$$

The hidden state output h_t at time step t of the BiGRU model is determined by the input at time step t, the forward hidden state output $\overrightarrow{h}_{t-1}$ at time step $t-1$, and the backward hidden state output $\overleftarrow{h}_{t-1}$ at time step $t-1$.

3.3. Multiple Self-Attention Mechanism Layer

The Self-Attention mechanism (SA) is a variant of the attention mechanism, which extracts information by computing the correlations between different elements in a sequence [28]. The BiGRU model, however, is unable to fully capture long-distance semantic features embedded in the text. By introducing the attention mechanism, more attention (larger feature weight) is allocated to the features related to the entities in the electric grid operation violation descriptions, while irrelevant features receive less attention (smaller feature weight). The attention mechanism helps effectively capture contextual semantic features associated with the current information, improving the model's ability to extract local features. Attention weights are determined by the word vectors themselves, rather than the distance between words, which helps solve the long-distance dependency problem in the BiGRU model. The schematic diagram of the Self-Attention mechanism is shown in Figure 5. For the input embedding representations, the model calculates three matrices: the Query matrix Q, the Key matrix K, and the Value matrix V. These matrices are obtained through linear transformations of the input vectors, using learned weight matrices. The formula is expressed as follows:

$$Q = H^L W^q, K = H^L W^k, V = H^L W^v \tag{10}$$

where H^L represents the multi-head attention receiving the output embedding representation from BiGRU, and W^q, W^k, W^v represent the corresponding transformation weight matrices.

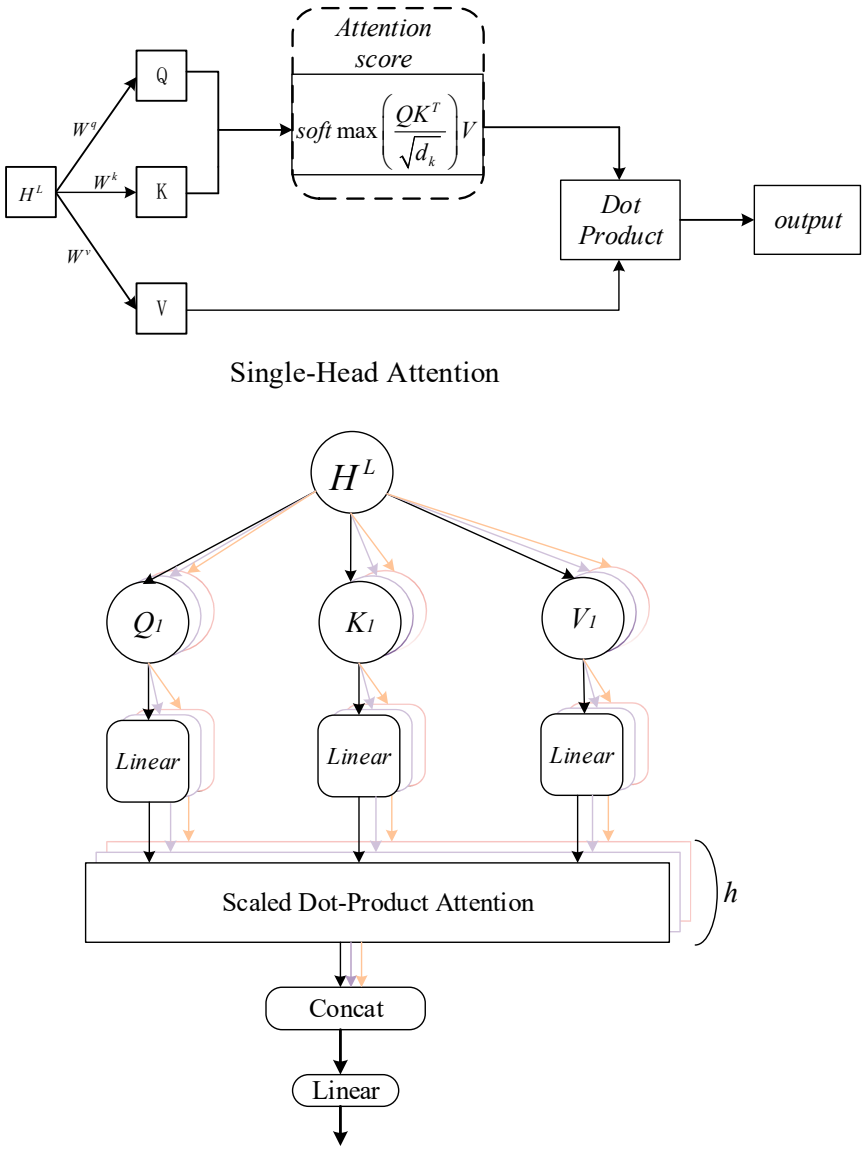


Figure 5. Diagrams of single-head and multi-head attention mechanisms.

Then, a dot product operation is performed to calculate the similarity between Q and K, and the softmax function is applied for normalization to obtain the weight coefficients. Finally, the matrix V is multiplied by the attention weights to obtain the attention output. The specific formula is the following:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{11}$$

where d_k represents the dimension of the key vectors, and the scaling factor $\sqrt{d_k}$ is used to normalize the dot product. This scaling helps prevent the dot product values from becoming too large or too small, keeping them within a reasonable numerical range.

The Multi-head Self-Attention (MHA) mechanism consists of multiple self-attention modules, which can simultaneously capture information from different positions and representation subspaces. This enables more effective extraction of context-dependent semantic features, helping to address the long-distance dependency issues in the BiGRU model that may arise due to gradient vanishing or explosion problems [29]. In this paper, the multi-head self-attention mechanism is chosen to extract the global semantic features from

the grid operation violation description text. The schematic diagram of the mechanism is shown in Figure 5. In the multi-head attention mechanism, a single attention is expanded into h independent heads, with each head corresponding to a self-attention module that calculates the weights of its query (Q), key (K), and value (V). Then, the outputs from all heads are concatenated together and passed through a linear transformation to obtain the final multi-head self-attention representation. The calculation formula is as follows:

$$Q_i = H^L W_i^q, K_i = H^L W_i^k, V_i = H^L W_i^v, \quad i = 1, 2, \ldots, h \tag{12}$$

$$head_i = Attention(Q_i, K_i, V_i) \tag{13}$$

$$\alpha = MultiHead(Q, K, V) = concat(head_1, head_2, \ldots, head_h)W^o \tag{14}$$

where H^L is the input, W_i^q, W_i^k, W_i^v is the weight matrix corresponding to the i-th head, and W^o is the output weight matrix. The operation concat denotes matrix concatenation. After computing the attention output for each head, these outputs are concatenated along the last dimension to form a large vector. The concatenated vector is then passed through a linear transformation to obtain the final multi-head attention output.

3.4. Label Decode Layer

The addition of the multi-head self-attention mechanism module after the BiGRU model can handle long-distance dependency issues caused by gradient vanishing or gradient explosion, yielding the most relevant output labels. However, it cannot resolve the dependency relationships between labels, such as the fact that the "B-Action" entity label should not directly follow the "I-File" entity label, or that the label of the first word in a sentence should start with "B-" or "O" rather than "I-". Due to these issues, the output may not be a reasonable prediction result for the model. The core function of CRF (Conditional Random Field) [30] is to model the dependencies between labels through a transition score matrix. It takes into account the implicit rules of the label sequence, thereby outputting a globally optimal label sequence.

Let $S, S \in R^{n \times m}$ be the output score matrix of the BiGRU model integrated with Multi-Head Attention (MHA), where s_{ij} denotes the score for the j-th label associated with the i-th character x_i in the text sequence. The score for the predicted label sequence $Y = \{y_1, y_2 \cdot \cdots \cdot y_n\}$ corresponding to the input text sequence $X = \{x_1, x_2 \cdot \cdots \cdot x_n\}$ is computed using the following functional equation:

$$score(X, Y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} S_{i, y_i} \tag{15}$$

where A is the transfer score matrix, $A \in R^{(m+2)\times(m+2)}$, $A_{yi,yi+1}$ represents the transfer score from label y_i to label y_{i+1}, and $S_{i,yi}$ represents the probability that the i-th character in the input text sequence is predicted to be label y_i.

Then, the probability of the output label sequence Y is computed using the normalized activation function, softmax, as follows:

$$P(Y|X) = \frac{e^{score(X,Y)}}{\sum_{\widetilde{Y} \in Y_X} e^{score(X,\widetilde{Y})}} \tag{16}$$

where Y_X is the set of all possible label sequences and $\widetilde{Y}$ is the true label sequence.

Finally, the Viterbi algorithm is applied to find the globally optimal label sequence Y^* for the text sequence X. Y^* is the label sequence that maximizes the output probability. The specific formula is as follows:

$$Y^* = \underset{\widetilde{Y} \in Y_X}{\operatorname{argmax}}\, score(X, \widetilde{Y}) \tag{17}$$

4. Experiments and Analysis of Results

4.1. Experimental Environment

The experimental environment setup is shown in Table 1.

Table 1. Experimental environment information.

Experimental Environment Name	Configuration
Operating System	Win10
Programming Language	Python 3.7
Deep Learning Framework	Pytorch 2.0.1
Video Card	NVIDIA RTX 4060ti
Memory	32 GB

4.2. Data Introduction and Preprocessing

Data from 24,001 work records collected by a power grid company from January to July 2022 were obtained. After text preprocessing, 15,183 usable violation records were retained, including complete violation descriptions for field operations. Based on expert knowledge, five types of labels were determined: Action, Object, File, Person, and Surrounding. The YEDDA annotation tool was used to perform BIO labeling on the collected power grid field operation violation descriptions. The final BIO dataset for named entity recognition of the violation descriptions was annotated by two staff members, and any discrepancies in the annotations were discussed and re-labeled. In the BIO scheme, 'B' indicates the beginning of an entity (Begin), 'I' indicates inside an entity (Inside), and 'O' represents positions outside the entity (Outside). The specific YEDDA annotation process is shown in Figure 6.

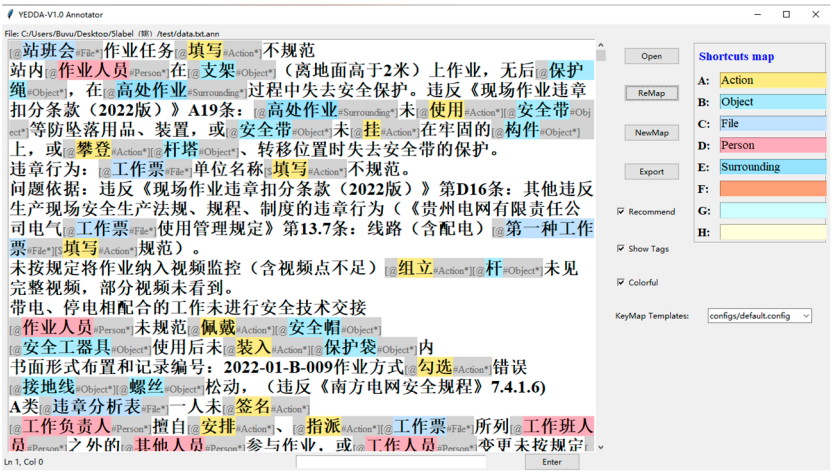


Figure 6. YEDDA Annotation Process.

The entity type labels in the power grid field operation violation description text data after BIO tagging can be found in Table 2. In the experimental validation, the annotated sample data was divided into training, validation, and test sets in a 7:2:1 ratio.

Table 2. Entity type distribution in electric grid dataset.

Entity Type	Number
Action	10,572
Object	9755
File	4120
Person	3232
Surrounding	4701

4.3. Model Evaluation Indicators

Named entity recognition is typically evaluated using three metrics: Recall, Precision, and F1-score, to assess the model's performance. The F1-score is a harmonic mean of Precision and Recall, used to measure the overall performance of the model. The formulas for each metric are shown in Equations (18)–(20):

$$Precision = \frac{TP}{(TP + FP)} \tag{18}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{19}$$

$$F1 - score = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \tag{20}$$

where TP (True Positive) is the number of actual positive samples correctly classified as positive, FP (False Positive) is the number of negative samples incorrectly classified as positive, and FN (False Negative) is the number of actual positive samples incorrectly classified as negative.

The values of Precision, Recall, and F1-score range from [0,1], with higher values indicating better model performance.

4.4. Results

To validate the effectiveness of the model proposed in this paper, we compare it with several commonly used entity recognition models: BiLSTM-CRF [16], LR-CNN [31], Lattice-LSTM [22], RoBERTa-BiLSTM-CRF [32], BERT-CRF [19], and BERT_wwm-BiLSTM-CRF [33], with BiLSTM-CRF as the baseline model. After several experiments, we determined that $batchsize = 32$, which may be related to the fact that most of the violation descriptions in the power grid field operation data contain fewer than 32 characters. In the comparison experiments, in order to be fair, the model constructed in this paper and other comparison models use Adam as the optimizer, the learning rate is set to $learningrate = 0.0001$, the number of training rounds is set to $epoch = 128$, and the specific comparison results are shown in Figure 7. The specific numerical results can be found in Table 3.

As shown in Figure 7 and Table 3, the model proposed in this paper outperforms all the compared models in terms of precision, recall, and F1-score. This demonstrates that the proposed model is more accurate in recognizing entities in the violation description text of grid field operations. The comparison between the BiLSTM-CRF and Lattice-LSTM models reveals that incorporating lexical information effectively enhances the recognition of entity boundaries by the named entity recognition model. Additionally, the comparison among the BiLSTM-CRF, BERT-CRF, and BERT_wwm-BiLSTM-CRF models shows that adding BERT embeddings or the BiLSTM module significantly improves the model's contextual understanding and its ability to handle out-of-vocabulary words, thereby greatly enhancing recognition accuracy and generalization capability. The comparison between

the RoBERTa-BiLSTM-CRF model and the BERT_wwm-BiLSTM-CRF model reveals that the advantage of the RoBERTa model primarily lies in its optimization for large-scale text understanding tasks. However, for the small-scale corpus of power grid violation descriptions constructed in this study, the limited data volume may not be sufficient to fully leverage the pretraining potential of the RoBERTa model, thereby constraining its performance improvement. Furthermore, when compared with the classical Lattice-LSTM model, which integrates Chinese word information, it is evident that the vocabulary fusion method and the use of the attention mechanism in the proposed model enable more effective learning of semantic features from the text context.

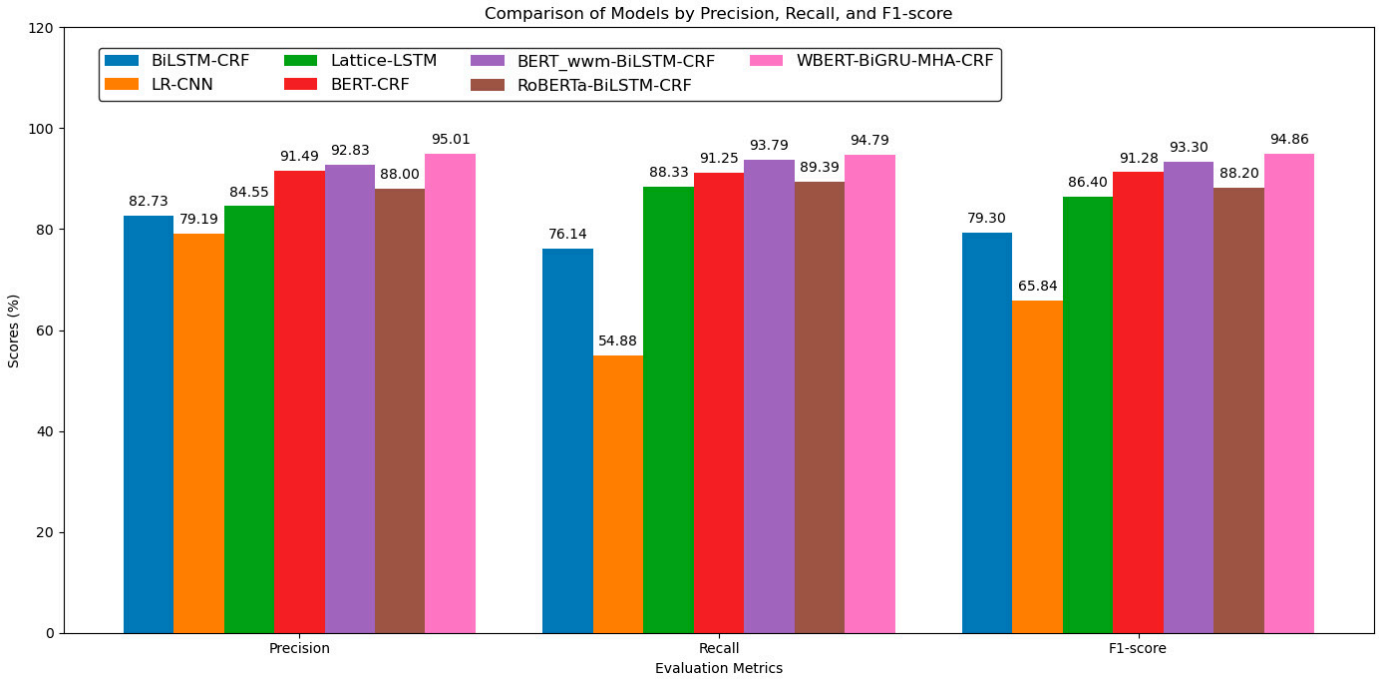


Figure 7. Model performance comparison on the test set for electric grid violation description recognition.

Table 3. Numerical results: model performance comparison on the test set for electric grid violation description recognition.

Model	Precision	Recall	F1-Score
BiLSTM-CRF [16]	82.73%	76.14%	79.30%
LR-CNN [31]	79.19%	54.88%	65.84%
Lattice-LSTM [22]	84.55%	88.33%	86.40%
BERT-CRF [19]	91.49%	91.25%	91.28.%
BERT_wwm-BiLSTM-CRF [32]	92.83%	93.79%	93.30%
RoBERTa-BiLSTM-CRF [33]	88.00%	89.39%	88.20%
WBERT-BiGRU-MHA-CRF (the proposed model)	**95.01%**	**94.79%**	**94.86%**

Ablation experiments were conducted on the grid operation record text entity recognition model proposed in this paper to evaluate the effectiveness of the Word2Vec lexical information embedding, BiGRU layer, and attention mechanism layer. The results are presented in Table 4.

By comparing test sequences 1 and 4 in Table 4, it is evident that the word fusion embedding, achieved by combining Word2Vec-based electrical power vocabulary embeddings trained on the power grid fieldwork corpus with BERT-based dynamic character embeddings, significantly improves the recognition of domain-specific terms in power grid operation records. Specifically, when a text sequence is input, BERT provides dynamic em-

beddings for each character, while Word2Vec captures the feature embeddings of electrical power terms identified within the sequence. These embeddings are then integrated and passed to the feature embedding encoding layer. Experimental results demonstrate that this method effectively enhances the ability to identify technical terms, playing a critical role in improving entity boundary detection.

Table 4. Results of ablation study on the test set.

	Model Structure			Evaluation of Indicators		
Test Number	**Lexical Embedding**	**Attention Mechanisms**	**GRU Module**	**Precision**	**Recall**	**F1-Score**
1	×	√	√	94.96%	94.79%	94.85%
2	√	×	√	94.43%	94.30%	94.34%
3	√	√	×	94.24%	93.81%	93.95%
4	√	√	√	95.01%	94.79%	94.86%

The comparison between test sequences 2 and 4 in Table 4 illustrates that multi-head attention, by processing in parallel across multiple attention heads, not only captures diverse semantic features in different projection subspaces but also models long-distance dependencies effectively. This mechanism allows for a more comprehensive focus on key information within complex text while mitigating the issue of over-reliance on local features, which is common in single-head attention.

When comparing the results of test sequences 3 and 4 in Table 4, where the BiLSTM model is used for comparison in the absence of the BiGRU model, it is found that the BiGRU model performs better. This may be attributed to the fact that, with a smaller dataset in the grid field operation violation text, the BiGRU model is more efficient in identifying and emphasizing key semantics due to its simpler structure, fewer parameters, higher training efficiency, and stronger robustness. As a result, BiGRU models generally outperform BiLSTM models. In summary, the lexical information embedding using the Word2Vec model, the BiGRU layer, and the attention mechanism layer in the named entity recognition model for grid field operation violation descriptions developed in this study enable better learning of contextual semantic information and entity boundaries, thereby improving the accuracy of named entity recognition.

5. Conclusions

Grid field operation violation description text named entity recognition is crucial for extracting information from vast amounts of historical grid inspection records. This paper investigates a deep learning-based method for named entity recognition (NER) of grid field operation violation description text to address current challenges in the domain of grid operation record NER. A labeled corpus for grid field operation violation description text NER is constructed, and a Chinese power violation description entity recognition model incorporating lexicon and attention mechanisms is proposed. The character-level dynamic embeddings of the input text sequences are obtained using the BERT pre-training model, while lexical information embedding is introduced through the Word2Vec model to handle the large number of specialized terms in the grid operation record text dataset. The global semantic features of the input sequences are captured using a BiGRU bi-directional gating unit, which reduces training time. The attention mechanism is employed to enhance the semantic information of feature vectors, addressing long-distance dependencies in feature extraction and improving entity recognition accuracy. Additionally, a CRF layer is utilized to handle label interdependencies, further improving the effectiveness of named entity recognition for grid field operation violation descriptions. Experimental results demon-

strate that, in the validation of the NER method for grid operation record text, the performance metrics—such as precision, recall, and F1-score—are significantly superior to those of the baseline model and other related NER models.

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

Funding: This research was funded by Guizhou Power Grid Co., Ltd., grant number GZKJXM20222320.

Data Availability Statement: Data are available upon request from the corresponding authors or works cited in the article.

Conflicts of Interest: Author Lingwen Meng was employed by Electric Power Research Institute of Guizhou Power Grid Co., Ltd. Yuanjun Huang and Dingli Ma were employed by the Southern Power Grid Digital Grid Group Co., Ltd. Guizhou Branch. 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. The authors declare that this study received funding from Guizhou Power Grid Co., Ltd. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

References

1. Ji, Z.; Wang, X.; Cai, C.; Sun, H. Power entity recognition based on bidirectional long short-term memory and conditional random fields. *Glob. Energy Interconnect.* **2020**, *3*, 186–192. [CrossRef]
2. Li, J.; Sun, A.; Han, J.; Li, C. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 50–70. [CrossRef]
3. Liu, P.; Guo, Y.; Wang, F.; Li, G. Chinese named entity recognition: The state of the art. *Neurocomputing* **2022**, *473*, 37–53. [CrossRef]
4. Brandsen, A.; Verberne, S.; Lambers, K.; Wansleeben, M. Can BERT dig it? named entity recognition for information retrieval in the archaeology domain. *J. Comput. Cult. Herit. (JOCCH)* **2022**, *15*, 1–18. [CrossRef]
5. Babych, B.; Hartley, A. Improving machine translation quality with automatic named entity recognition. In Proceedings of the 7th International EAMT Workshop on MT and Other Language Technology Tools, Improving MT Through Other Language Technology Tools, Resource and Tools for Building MT at EACL, Budapest, Hungary, 13 April 2003.
6. Hu, Z.; Ma, X. A novel neural network model fusion approach for improving medical named entity recognition in online health expert question-answering services. *Expert Syst. Appl.* **2023**, *223*, 119880. [CrossRef]
7. Abdoun, N.; Chami, M. Automatic text classification of pdf documents using nlp techniques. *INCOSE Int. Symp.* **2022**, *32*, 1320–1331. [CrossRef]
8. Liu, Y.; Huang, S.; Li, R.; Yan, N.; Du, Z. USAF: Multimodal Chinese named entity recognition using synthesized acoustic features. *Inf. Process. Manag.* **2023**, *60*, 103290. [CrossRef]
9. Wu, J.; Xu, X.; Liao, X.; Li, Z.; Zhang, S.; Huang, Y. Intelligent diagnosis method of data center precision air conditioning fault based on knowledge graph. *Electronics* **2023**, *12*, 498. [CrossRef]
10. Shaalan, K.; Raza, H. NERA: Named entity recognition for Arabic. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 1652–1663. [CrossRef]
11. Pradhan, A.; Yajnik, A. Parts-of-speech tagging of Nepali texts with Bidirectional LSTM, Conditional Random Fields and HMM. *Multimed. Tools Appl.* **2024**, *83*, 9893–9909. [CrossRef]
12. Yuan, T.; Qin, X.; Wei, C. A Chinese named entity recognition method based on ERNIE-BiLSTM-CRF for food safety domain. *Appl. Sci.* **2023**, *13*, 2849. [CrossRef]
13. Wu, X.; Wu, Z.; Jia, J.; Cai, L. Adaptive named entity recognition based on conditional random fields with automatic updated dynamic gazetteers. In Proceedings of the 2012 8th International Symposium on Chinese Spoken Language Processing, Hong Kong, China, 5–8 December 2012; pp. 363–367.
14. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
15. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.

16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
17. Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
18. Souza, F.; Nogueira, R.; Lotufo, R. Portuguese named entity recognition using BERT-CRF. *arXiv* **2019**, arXiv:1909.10649.
19. 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.
20. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Yang, Z. Pre-training with whole word masking for Chinese bert. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3504–3514. [CrossRef]
21. Zhang, Y.; Yang, J. Chinese NER using lattice LSTM. *arXiv* **2018**, arXiv:1805.02023.
22. Cong, L.; Cui, R.; Dou, Z.; Huang, C.; Zhao, L.; Zhang, Y.; Chen, C.; Su, C.; Li, J.; Qu, C. Named entity recognition for power data based on lexical enhancement and global pointer. In Proceedings of the Third International Conference on Electronic Information Engineering, Big Data, and Computer Technology (EIBDCT 2024), Beijing, China, 26–28 January 2024; Volume 13181, pp. 1345–1351.
23. Feng, J.; Wang, H.; Peng, L.; Wang, Y.; Song, H.; Guo, H. Chinese Named Entity Recognition Within the Electric Power Domain. In *International Symposium on Emerging Information Security and Applications*; Springer Nature: Singapore, 2023; pp. 133–146.
24. Du, B.; Fan, J.; Yuan, L.; Xie, C.; Song, D.; Luo, X.; Xie, W. Research on Construction Technology of Knowledge Graph for Power Grid Fault Handling Based on Attention Mechanism Bilstm-CRF Model. *Power Syst. Big Data* **2024**, *27*, 37–45.
25. He, L.; Zhang, X.; Li, Z.; Xiao, P.; Wei, Z.; Cheng, X.; Qu, S. A Chinese named entity recognition model of maintenance records for power primary equipment based on progressive multitype feature fusion. *Complexity* **2022**, *2022*, 8114217. [CrossRef]
26. Jiang, C.; Wang, Y.; Hu, J.; Xu, J.; Chen, M.; Wang, Y.; Ma, G. Power entity information recognition based on deep learning. *Power Syst. Technol.* **2021**, *45*, 2141–2149.
27. Mikolov, T. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
28. Cui, L.; Tian, X.; Wei, Q.; Liu, Y. A self-attention based contrastive learning method for bearing fault diagnosis. *Expert Syst. Appl.* **2024**, *238*, 121645. [CrossRef]
29. Voita, E.; Talbot, D.; Moiseev, F.; Sennrich, R.; Titov, I. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv* **2019**, arXiv:1905.09418.
30. Lafferty, J.; McCallum, A.; Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Int. Conf. Mach. Learn.* **2001**, *1*, 3.
31. Gui, T.; Ma, R.; Zhang, Q.; Zhao, L.; Jiang, Y.G.; Huang, X. CNN-Based Chinese NER with Lexicon Rethinking. In Proceedings of the International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019.
32. Liu, Y. Roberta: A robustly optimized bert pretraining approach [EB/OL]. *arXiv* **2019**, arXiv:190711692.
33. Huang, Q.; Tao, Y.; Wu, Z.; Marinello, F. Based on BERT-wwm for Agricultural Named Entity Recognition. *Agronomy* **2024**, *14*, 1217. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.