*Article*

# Named Entity Recognition in Power Marketing Domain Based on Whole Word Masking and Dual Feature Extraction

Yan Chen [1,2,*], Zengfu Liang [1], Zhixiang Tan [1] and Dezhao Lin [1]

[1] School of Computer and Electronic Information, Guangxi University, Nanning 530004, China
[2] Guangxi Intelligent Digital Services Research Center of Engineering Technology, Nanning 530004, China
* Correspondence: cy@gxu.edu.cn

**Abstract:** With the aim of solving the current problems of low utilization of entity features, multiple meanings of a word, and poor recognition of specialized terms in the Chinese power marketing domain named entity recognition (PMDNER), this study proposes a Chinese power marketing named entity recognition method based on whole word masking and joint extraction of dual features. Firstly, word vectorization of the electricity text data is performed using the RoBERTa pre-training model; then, it is fed into the constructed dual feature extraction neural network (DFENN) to acquire the local and global features of text in a parallel manner and fuse them. The output of the RoBERTa layer is used as the auxiliary classification layer, the output of the DFENN layer is used as the master classification layer, and the output of the two layers is dynamically combined through the attention mechanism to weight the outputs of the two layers so as to fuse new features, which are input into the conditional random field (CRF) layer to obtain the most reasonable label sequence. A focal loss function is used in the training process to alleviate the problem of uneven sample distribution. The experimental results show that the method achieved an F1 value of 88.58% on the constructed named entity recognition dataset in the power marketing domain, which is a significant improvement in performance compared with the existing methods.

**Keywords:** power marketing; named entity recognition; RoBERTa; dual feature extraction; attention mechanism

## 1. Introduction

With the reform of the electric power system, power grid enterprises have completed the initial construction of smart grids [1–3]. As a result, they have accumulated a substantial amount of unstructured business data [4] through the implementation of smart grid information systems. This includes information about electric power marketing [5,6] systems. To understand this huge amount of marketing data, there is a need for classification and keyword positioning. Deep semantic relationship mining has become a hot research direction in the field of natural language processing [7,8] and electric power marketing. Named Entity Recognition (NER) [9–11] is a fundamental task in natural language processing; its main task is to identify meaningful nouns or phrases from unstructured text and classify them. In the case of power marketing, named entity recognition is mainly used to identify entities in power marketing texts. For example, by extracting the unstructured information in the complaint opinion work order, the complaint time, faulty equipment, station line, belonging station, etc., can be understood quickly. This will enable effectively improving the efficiency of marketing personnel, as well as improving the quality of power marketing services. This will provide the basis for the construction of knowledge graph [12] or power customer service intelligent question and answer systems [13] in the field of power marketing afterwards.

Different from the traditional named entities, power marketing text contains a large number of proprietary terms in the field of electricity, which has the characteristics of complexity, specificity, and a strong domain. This will enhance the difficulty of named entity

recognition to a certain extent. If the accuracy rate of named entity recognition is low, it will have a serious impact on downstream tasks, so how to improve the accuracy rate of named entity recognition in power marketing field is an urgent problem to be solved. At present, NER technology in the Chinese power marketing domain mainly faces the following challenges:

(1) Lack of publicly available labeled datasets.
(2) Electricity entities are highly specialized and difficult to identify in comparison with general-purpose fields, and there are problems such as multiple meanings of words in electricity entities. For example, both entities "60 kV镇龙线" (60 kV Zhenlong Line) and "10 kV镇龙站" (10 kV Zhenlong Station) contain the word "镇龙" (Zhenlong). The word "镇龙" (Zhenlong) can mean either on the line or on the station.
(3) The feature extraction of traditional models is generally limited to contextual information and annotation correctness, which is more suitable for general entity recognition datasets. Some local features specific to entity words in the power marketing domain dataset were not extracted, which had an impact on the recognition effect of this type of dataset.

In order to solve the above problems, this paper proposes a new method for named entity recognition in the power marketing domain. First, word vectorization is performed on the power marketing text data using the RoBERTa pre-training model. Then, the vectors are fed into the constructed DFENN to obtain both local and global features of the text and to fuse them. The output of the RoBERTa layer of the model is used as the auxiliary classification layer and the output of the DFENN layer is used as the master classification layer. The outputs of the two layers are dynamically weighted and fused to obtain new features through the attention mechanism; they are then input into the CRF layer to obtain the most reasonable label sequence. Through experimental verification, the PMDNER model proposed in this paper can better utilize the features of entities and find the key influential information for the output more precisely, thus improving the results of entity recognition and comprehensively characterizing the semantic information of the electricity marketing text.

The contributions of this paper are as follows:

(1) By processing the power marketing data provided by the Guangxi Power Grid, the unstructured data from it are extracted, and the entities are classified with reference to the actual needs of power companies for annotating the data, so as to finally construct a named entity identification dataset in the field of power marketing.
(2) Using RoBERTa pre-training as the embedding layer, the whole word-masking strategy is adopted to mask and predict the words in the power marketing text according to the characteristics of the Chinese language, so that the model can better characterize the Chinese semantics. DFENN as an intermediate layer can accurately obtain both the global features and local features of the text, which is conducive to characterizing the text features more comprehensively and improving the recognition effect of the named entities.
(3) In this paper, the auxiliary classification layer is introduced; the output of the RoBERTa layer and the output of the DFENN layer are applied as the auxiliary classification layer and the master classification layer, respectively; and the attention mechanism function is used for weighted fusion, which can properly perform the annotation of power marketing data and verify the effective recognition of the named entities in the power marketing domain through a comparison test.
(4) Through the use of the focal loss function, the problem of uneven sample distribution is alleviated and the model's ability to recognize difficult samples is improved.

## 2. Related Work

At present, the common methods for named entity recognition include dictionary and rule-based methods, statistical machine-learning-based methods, and deep-learning-based methods. The dictionary and rule-based named entity recognition methods usu-

ally achieve good results in a specific domain, but are time-consuming, laborious and less portable. With the development of machine learning technology, the machine learning method represented by CRF [14] can effectively reduce the labor cost by combining dictionary and rules, but they still have drawbacks, such as inefficient feature extraction and slow training.

In recent years, deep learning methods based on neural networks have achieved great success in fields such as computer vision and speech recognition, and good results have been achieved in the research of named entity recognition, a key fundamental task in natural language processing. Li et al. [15] proposed a named entity recognition model for BiLSTM-CRF power defect text based on selectable word vectors. The model can identify the category information of specialized named entities in the power domain from a large number of power defect texts, thus constructing a large amount of power defect text data and facilitating the management of text data. Guo et al. [16] used an Att-BiLSTM-CRF model combined with a self-attention mechanism to achieve excellent results on clinical named entity recognition (CNER). Li et al. [17] proposed a dynamic attention mechanism-based approach by stitching the character vector of raw text information and the word vector of domain information to further enhance the model performance. Ling et al. [18] proposed a bidirectional long and short-term memory method based on attention and conditional random field layer (Att-BiLSTM-CRF) for chemical-named entity recognition to achieve document-level chemical memory, the method obtains document-level global information through the attention mechanism, enforces marker consistency among multiple instances of the same token in the document, and achieves good performance on small feature engineering, which is widely cited by researchers. Kai et al. [19] proposed a neural network method called Dic-Att-BiLSTM-CRF (DABLC), which uses an efficient exact string matching method to match entities with dictionaries and constructs a dictionary attention layer to enhance the model performance by combining the dictionary matching method with a document-level attention mechanism. In response to the complexity and lack of context of medical terminology on Twitter, Erdenebileg et al. [20] used a bidirectional long-short-term memory model (BiLSTM) to learn a large amount of contextual information, and then used convolutional neural networks (CNN) to generate character-level features, and labeled information incorporated into the CRF model to detect health-related entities from Twitter messages. Hao et al. [21] used the attention mechanism to improve the vector representation in BiLSTM, designed several different attention weight assignment methods and combined them to effectively prevent significant information loss during feature extraction, and finally, combining BiLSTM with the CRF layer can effectively solve the problem of strong label dependence in sequences. The self-attentive mechanisms utilized in the above studies have few parameters and low complexity. Although they enhance the computational efficiency, they all suffer from the inability to learn the sequential relationships within the sequence. Furthermore, most of the embedding layers utilize the Word2Vec model, resulting in context-independent shallow feature vectors that are insufficient for characterizing the multiple meanings of words.

Pre-trained language models using deep learning not only overcome the limitations of machine learning methods, but also address the shortcomings of self-attentive mechanisms. Rong et al. [22] used a pre-trained XLNet model to extract sentence features, and then combined it with BiLSTM and CRF to demonstrate the superiority of XLNet in the NER task, which resulted in good recognition results on both the CoNLL-2003 English dataset and WNUT-2017 dataset. Jacob Devlin et al. from the Google team [23] proposed a language pre-trained model based on Bidirectional Encoder Representation from Transformers (BERT) to characterize word vectors in 2018. Meng et al. [24] proposed a novel model, BERT-BiLSTM-CRF, designed specifically for identifying and extracting power equipment entities from pre-processed Chinese technical literature. Li et al. [25] proposed a new strategy to incorporate dictionary features into the model, and then pre-trained the model on unlabeled clinical data using BERT, which can exploit unlabeled domain-specific knowledge, and then extracted text features and predictive labels using long short-term mem-

ory (LSTM) and CRF. Wu et al. [26] employed the RoBERTa-BiLSTM-CRF model for clinical named entity recognition (CNER) and achieved favorable recognition results on the CCKS2017 and CCKS2019 datasets. Lee et al. [27] used BioBERT, a domain-specific language representation model that had been pre-trained on a large-scale biomedical corpus, for named entity recognition, with boosting results substantially better than BERT and previous state-of-the-art models. He et al. [28] proposed an entity recognition method based on progressive multi-type feature fusion. This approach utilized the BERT preprocessing model to obtain word vectors with contextual information and conducted named entity recognition on power maintenance datasets. Tong et al. [29] proposed a named entity method for power communication planning based on Transformer and BiLSTM-CRF models to address challenges associated with long text and low efficiency of information extraction in power communication planning reports.

The studies mentioned above primarily focused on single word vector features for named entity recognition, overlooking the domain-specific text in an integrated manner. In modeling, the model tends to prioritize global features of text sequences, neglecting the modeling of local semantic information within the text sequences. However, the inclusion of local semantic information plays an equally critical role in the recognition of named entities, and ignoring it may lead to sparse and incomplete feature extraction, ultimately leading to lower accuracy in the named entity model. In this paper, simultaneous extraction of global and local features is proposed for named entity recognition in the electric power marketing domain. This approach ensures that local features are adequately considered alongside global features when text sequences are extracted. Through the attention mechanism, the word vector output from the RoBERTa layer and the features extracted from the DFENN layer are dynamically weighted and fused, which helps in better utilization of entity features and addresses challenges such as low utilization of entity features, multiple meanings of words and inadequate recognition of specialized terms in the entity recognition process.
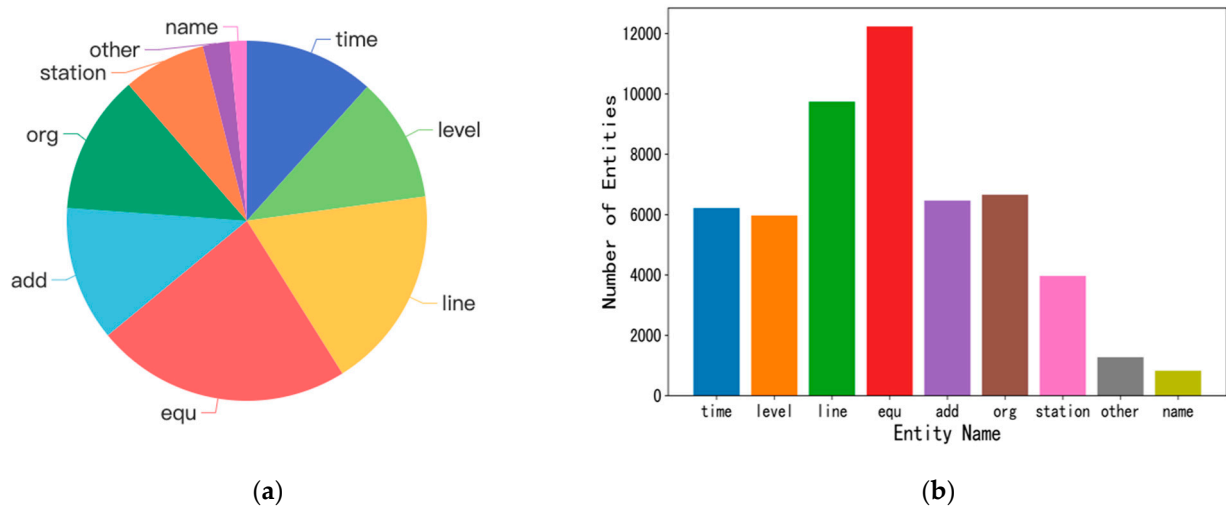
## 3. Constructing the Dataset

A large amount of unstructured data is recorded in the electric power marketing system of the Guangxi district. In order to study these data, this paper selects the marketing system data from the system in the past two years, which contain textual data such as accident investigation information, user feedback, device operation, etc., and constructs an electric power marketing corpus. The current smart grid system has a low utilization rate of these data, and only supports simple text queries, without conducting deep mining. As a result, the behavioral knowledge contained in them remains untapped. In this paper, we use the power marketing corpus to build a named entity recognition dataset in the power marketing domain, employing deep learning methods for analysis and modeling.

The power marketing corpus is derived from the actual operation system of the power grid, and through screening, sentences with unknown meaning, structural mutilation and semantic repetition are eliminated, and finally, 6909 high-quality data are extracted from the corpus for training and testing. By analyzing the characteristics of the power marketing corpus and considering the actual needs of the power company, the entity types are classified into nine categories, such as time, transmission equipment and voltage level (see Table 1 for details). During the subsequent recognition process, the entities will be classified into their respective categories. The labeled dataset comprises 53,378 entities. It is divided into three sets: the training set, the validation set and the test set. The distribution ratio among these sets is 8:1:1.

The distribution of entity categories is shown in Figure 1. The largest number of entities belongs to the "equ" type and the "line" type categories, while the "other" type and the "name" type categories have the smallest number of entities. The remaining entity types are distributed relatively evenly, and there is the problem of uneven sample distribution.

**Table 1.** Experimental dataset.

| Entity Type | Entity Example | Number of Entities |
|---|---|---|
| time | 2022年03月30日 (30 March 2022) | 6219 |
| level | 60 kV | 5969 |
| line | 60 Kv 镇龙线 (60 kV Zhenlong Line) | 9743 |
| equ | 901开关 (901 switch) | 12,234 |
| add | 玉林市玉州区仁东镇木根村 (Mugen Village, Rendong Town, Yuzhou District, Yulin City) | 6465 |
| org | 河口村第二经济合作社 (Hekou Village Second Economic Cooperative) | 6660 |
| station | 沙坡站 (Shapo station) | 3988 |
| other | 杂草,树木 (Weeds, trees) | 1273 |
| name | 李四 (Si Li) | 827 |



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

**Figure 1.** This is the distribution of entity types in the dataset. (**a**) Pie charts; (**b**) bar charts.

In this paper, we utilize the label-studio annotation platform, which offers a visual interface, to annotate entities within sentences. The BIO annotation method is employed for this purpose. In this method, "B" represents the begin of an entity, "I" represents the inside part of an entity, and "O" represents the outside part of an entity. If a word in the dataset is labeled as B/I-XXX, it indicates that the word represents either the beginning or inside part of a named entity, with XXX denoting the type of the named entity. Examples of the labeling are shown in Table 2 and Figure 2.
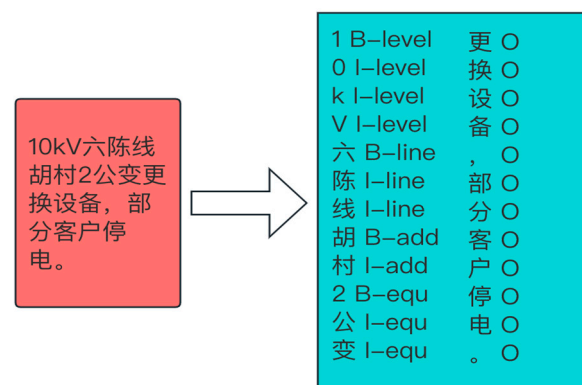


**Figure 2.** Example of corpus labeling. "10 kV六陈线胡村2公变更换设备, 部分客户停电" corresponds to "10 kV Liuchen line Hucun 2 transformer replacement equipment, part of the customer power outage".

**Table 2.** Entity annotation example. The meaning of the Text line in the table is "Failure of Public Transformer No. 5".

| Text | **5** | 号 | 公 | 变 | 出 | 现 | 故 | 障 |
|---|---|---|---|---|---|---|---|---|
| Annotation | B-equ | I-equ | I-equ | I-equ | O | O | O | O |

## 4. Methods

In this paper, we propose a named entity recognition method for the electric power marketing domain, which adopts the strategy of whole-word masking and joint dual feature extraction. The model consists of an input layer, RoBERTa layer, DFENN layer, Attention layer and CRF layer. The RoBERTa layer is responsible for converting the power text data into word vectors. The DFENN layer extracts both local and global features simultaneously and combines them through fusion techniques. The attention layer dynamically weighs the word vector output from the RoBERTa layer and the features output from the DFENN layer. The CRF layer is used to output the global optimal annotation sequence. The PMDNER model architecture diagram is shown in Figure 3.
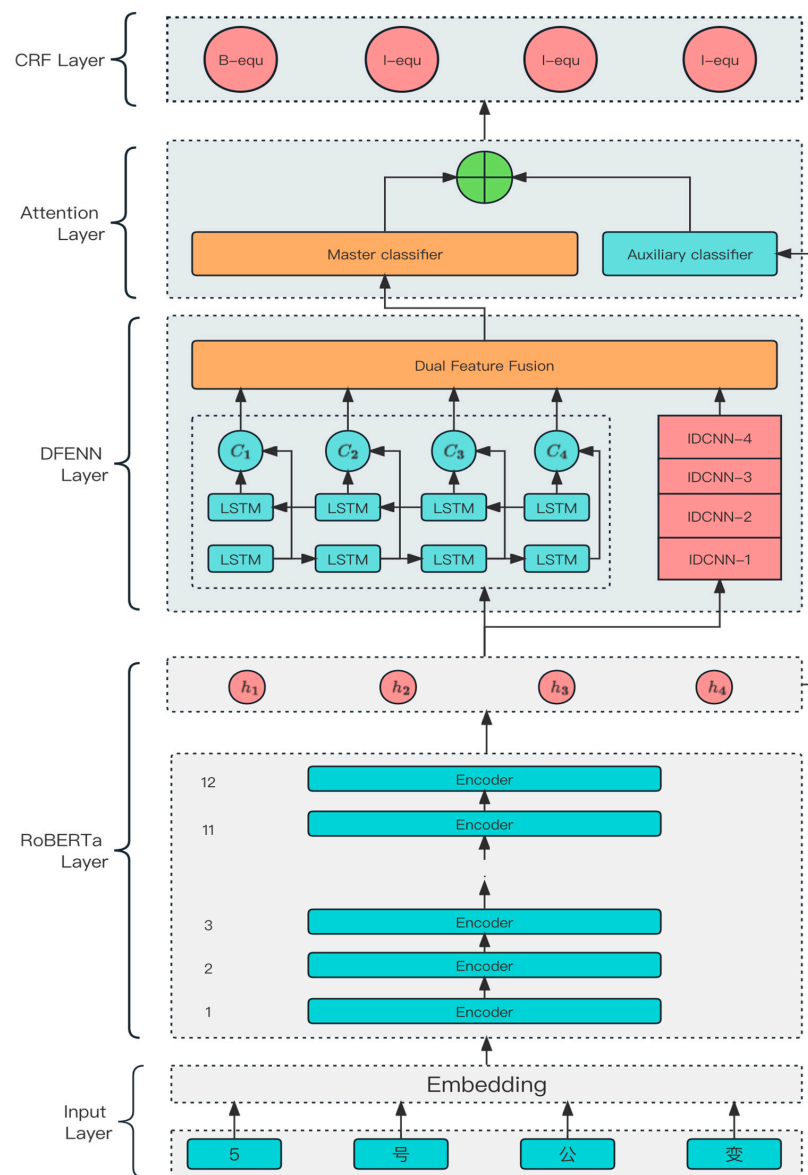


**Figure 3.** PMDNER model structure diagram. The input text means "Transformer No. 5".

*4.1. RoBERTa Pre-Training Model*

In this paper, we adopt the RoBERTa-wwm-ext pre-training model proposed by the joint lab of Harbin Institute of Technology and iFLYTEK, which is based on Chinese data and combines the advantages of Chinese Whole Word Masking (WWM) technology and RoBERTa model. The schematic of its coding layer is shown in Figure 4, where the sum of word embedding, segment embedding and positional embedding is used as the input of the model to better identify entity information, which can effectively solve the problem of multiple meanings of words that cannot be solved by traditional Word2Vec model. Furthermore, it solved the problem of complex and variable terminology found in Chinese datasets within the power marketing domain. The structure of the RoBERTa-wwm-ext model includes 12 layers of Transformer, after a network framework consisting of stacked encoder parts of multiple bidirectional Transformer models, it can capture the bidirectional relationship in the text more thoroughly, and the output word vector contains the a priori semantic knowledge acquired by RoBERTa-wwm-ext in the pre-training phase. When training the RoBERTa-wwm-ext model, its parameters are fine-tuned with the changes of the training set, and its values are continuously updated to better learn the semantic knowledge in the training set. Compared with the BERT model, RoBERTa-wwm-ext improves the pre-training method in three main aspects:

(1) The masking strategy uses whole word masks instead of single-character masks.
(2) Dynamic masking is used instead of static masking in the model task.
(3) The next sentence prediction (NSP) task in the pre-training phase is removed.

The difference between the RoBERTa-wwm-ext model and the BERT model in terms of masking strategy is shown in Figure 5. In the Chinese language context, the BERT model is a character-level masking strategy, whereas the RoBERTa-wwm-ext model is a word-level masking strategy. After the completion of the pre-training phase, the word vector output from RoBERTa-wwm-ext can be input to the next layer of DFENN for simultaneous extraction of global and local information.
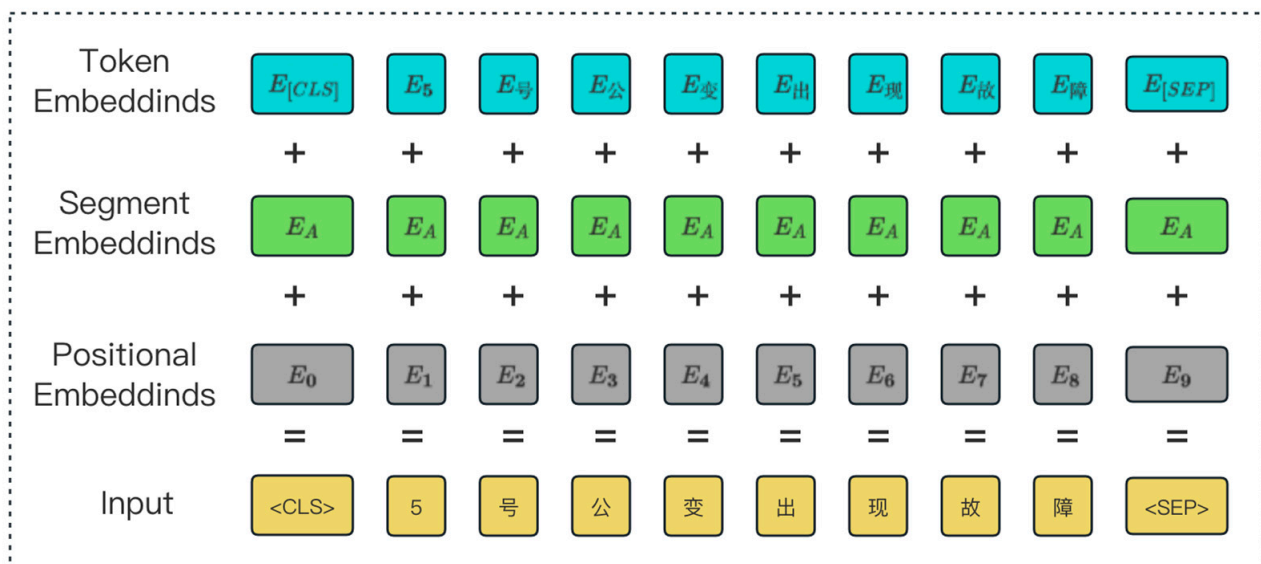


**Figure 4.** Schematic diagram of coding layer, RoBERTa generates the input vector. The token embeddings subscript means "Failure of Public Transformer No. 5".
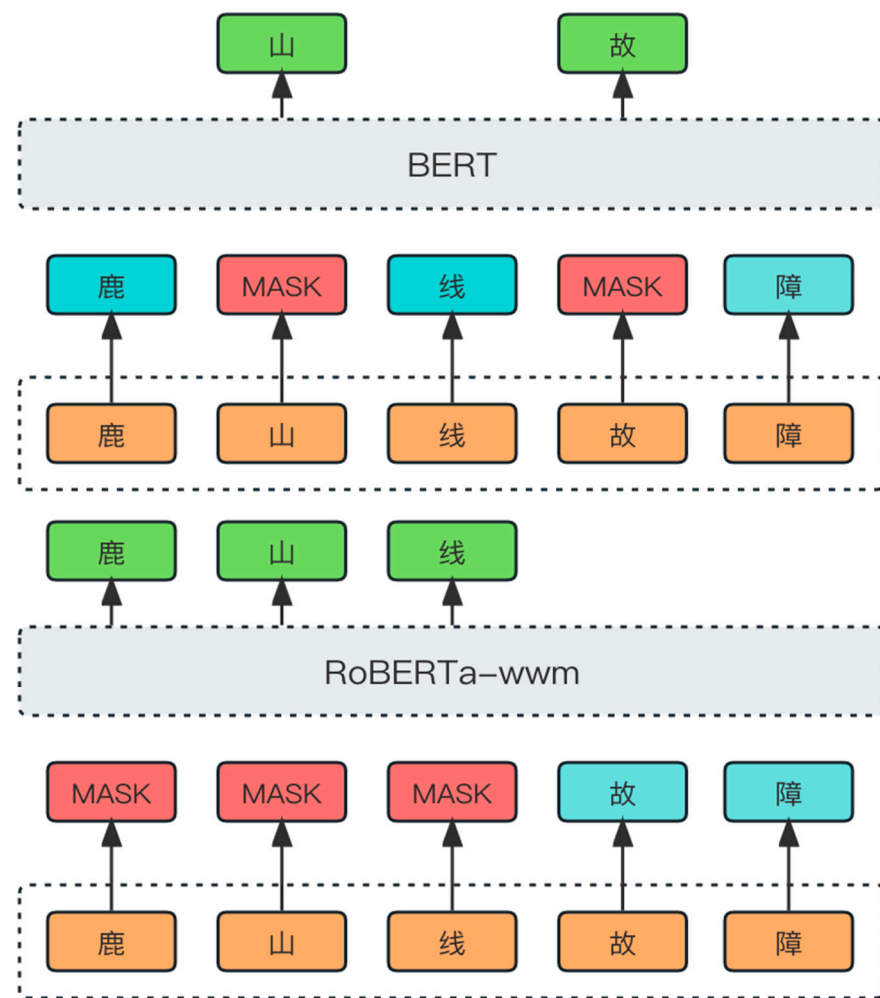
**Figure 5.** Mask strategy difference of RoBERTa-wwm and BERT model, MASK indicates that the current character is masked. "鹿山线故障" corresponds to "Lushan line failure".

### 4.2. Dual Feature Extraction Neural Network

In named entity recognition tasks, BiLSTM is typically used to capture the global semantic information of text sequences, but this can also lead to the loss of local semantic information. The local semantic information of the text sequence is also crucial for named entity recognition, and if the local semantic information is neglected, the final extracted features may be sparse and the semantic information may not be complete, thus reducing the accuracy of named entity recognition. The iterated dilated convolutions neural network (IDCNN) has a larger acceptance domain than the traditional convolutions neural network (CNN) and is more inclined to capture local information in text sequences, which is the exact opposite of BiLSTM. In the power marketing dataset, entity types with longer lengths, such as addresses or organizations, exhibit greater contextual long-range dependencies. On the other hand, entity types with more distinct features around them, such as electrical equipment and stations, require stronger local features. Therefore, DFENN is constructed to extract text features in order to utilize them more comprehensively. The word vectors output from RoBERTa are input to the BiLSTM network and IDCNN network, respectively, and the global semantic information and local semantic information of the text features are extracted in a parallel manner. Since the features are extracted using a parallel approach, there will not be much difference in time consumption. Finally, the features extracted by BiLSTM and IDCNN are spliced and fused, and the fused features contain not only global contextual semantic and syntactic information, but also local semantic information, which improves the utilization rate of entity features and thus can further improve the efficiency

of named entity recognition for power marketing. The layer network consists of two sub-modules, BiLSTM and IDCNN, which are described in detail below.

### 4.2.1. BiLSTM

Long Short-Term Memory (LSTM) is a special kind of Recurrent Neural Network (RNN). Unlike the traditional RNN, LSTM can better solve the problems of gradient disappearance and gradient explosion when processing sequential data. The LSTM cell structure is shown in Figure 6. The LSTM contains an internal state called "cell state", it can control the flow and forgetting of information to better handle long sequences of data. Among them, the calculation method for each step is as shown in Equations (1)–(6).
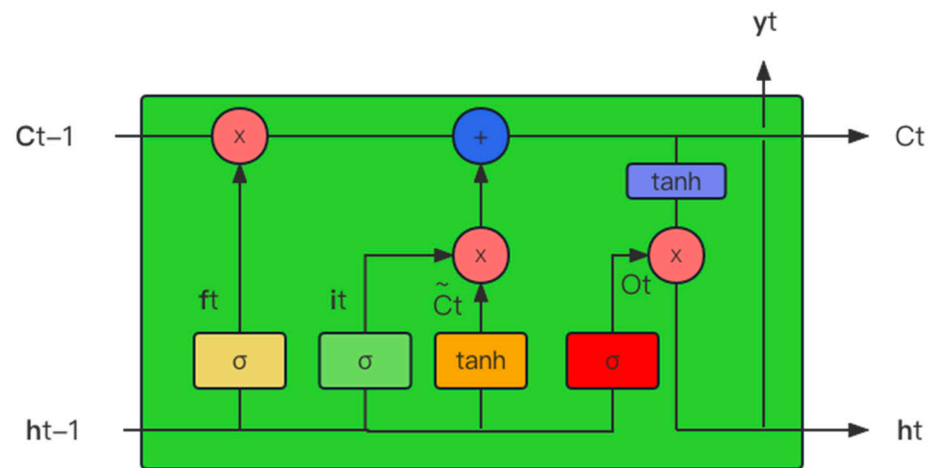


**Figure 6.** LSTM cell structure diagram.

$$f_t = c\left(W_f[h_{t-1}, x_t] + b_f\right) \tag{1}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{2}$$

$$\widetilde{C}t = tanh(W_c[h_{t-1}, x_t] + b_c) \tag{3}$$

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C}t \tag{4}$$

$$O_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = O_t \times tanh(C_t) \tag{6}$$

Based on the calculation steps mentioned above, the LSTM model selectively discards some unnecessary information and improves neuron memory, which can better solve the long-time dependency problem. However, LSTM fails to fully leverage the contextual information of power marketing data. BiLSTM is a bi-directional LSTM. The uni-directional LSTM model can only capture the information passed from front to back, while the bi-directional network can capture both forward and backward information. BiLSTM can simultaneously utilize the contextual information of the power marketing text, learn the long sequence semantic features of the text, and enhance the recognition ability of the model, which is calculated as in Equations (7)–(9).

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(x_t) \tag{7}$$

$$h_t = LSTM(x_t) \tag{8}$$

$$h_t = <\overrightarrow{h_t}, h_t> \tag{9}$$

The final hidden layer state $h_t$ is calculated by the above equation to capture the long-range context dependence of the power marketing data, and the output of BiLSTM is $(h_1, h_2, \ldots, h_n)$.

### 4.2.2. IDCNN

In this paper, we choose IDCNN to extract local semantic information from electricity marketing data. IDCNN is composed of four equal-sized expanded convolutional blocks stacked together, with each expanded convolutional block consisting of three expanded convolutional layers. The number of layers in each expanded convolutional block cannot be too many, as too many layers in the stack will result in an excessive number of parameters and eventually lead to model overfitting. In this paper, three convolutional layers and four iterations are set up, and the output of each layer after affine transformation is used as the input for the next layer of expansion convolution, and the same convolutional kernel size and filter size are set for these three expansion convolutional layers. The expansion widths of the three layers are set to 1, 2, and 4. This setting allows the network to take into account the information of each character while extracting text features, and the perceptual domain of the network can increase exponentially with the increase in the expansion width. Ultimately, the semantic information can be extracted at a relatively long distance.

In the dual feature extraction network, IDCNN's first layer of expansion width is 1, and the size of the convolution kernel is set to $3 \times 3$, then the output of the previous inflated convolutional layer is activated using the relu activation function to obtain the output of the next inflated convolutional layer. Finally, the output of the last layer of convolution is obtained, which is calculated as in Equations (10)–(12).

$$h_t^1 = D_1^{(0)} x_t \tag{10}$$

$$h_t^n = relu\left(D_{L_{n-1}}^{n-1} h_t^{n-1}\right) \tag{11}$$

$$H_t^n = tanh(w_t^n h_t^n + b_t^n) \tag{12}$$

In the equation, $D_j^{(i)}$ denotes the i-th expansion convolution layer with the expansion width set to j. $D_1^{(0)}$ denotes the first layer, and $h_t^1$ is the output of $x_t$ after the first expansion convolution. $h_t^n$ denotes the output of the n-th inflated convolutional layer, and $L_n$ denotes the number of layers of each inflated convolution. $w_t^n$ denotes the weight matrix and $b_t^n$ denotes the bias term.

Since an expanded convolution block consists of three expanded convolution layers, the above three layers can be treated as one expanded convolution block, represented by *B(i), i* represents the i-th convolution block, and the output of the previous expanded convolution block is used as the input of the next expanded convolution block. In this paper, the parameters are shared among the expanded convolution blocks, and the input of the m-th expanded convolution block is the output of the m-1st expanded convolution block. After completing the iteration, we finally obtain $b_t^m$. The final output local feature sequence is $(b_1, b_2, \ldots, b_n)$, which is calculated as in Equations (13) and (14).

$$b_t^{(1)} = B\left(D_1^{(0)} x_t\right) \tag{13}$$

$$b_t^{(m)} = B\left(b_t^{(m-1)}\right) \tag{14}$$

After the dual feature extraction in a parallel way, the local feature sequences $(b_1, b_2, \ldots, b_n)$ obtained by IDCNN and the global feature sequences $(h_1, h_2, \ldots, h_n)$ obtained by

BiLSTM are spliced and fused, and obtain a new feature sequence with both local and global features, which is calculated as shown in Equation (15).

$$F = \{(h_1 \oplus b_1), (h_2 \oplus b_2), \ldots, (h_n \oplus b_n)\} \tag{15}$$

### 4.3. Attention Layer

In named entity recognition tasks, the inputs to the model are often long sequences, but not all information in the sequence is useful for the entity recognition task. Therefore, only the important and useful information needs to be preserved. For this reason, we introduce the attention mechanism into the named entity recognition task. The attention mechanism enables the model to assign different weights to the parts of the input and extract the more critical and important information, so that the model can make more accurate judgments.

In this paper, the output of RoBERTa is used as the auxiliary classifier, while the DFENN output layer is used as the master classifier. After training, the word vectors output from the RoBERTa-wwm layer contain rich contextual semantic information, which can be learned as long-range global feature information and local feature information after inputting them into the DFENN model. Finally, the output vectors of both will be calculated the weights through the attention mechanism function, and after weighted fusion, they contribute to better sequence annotation of power marketing data. A function is used as a score function to measure the magnitude of correlation coefficients between the output vectors of RoBERTa-wwm layer and the output vectors of BiLSTM layer. The calculation of the function is shown in Equation (16),

$$similarity(h_t, f_s) = W \frac{\left(h_t - \overline{h_t}\right)\left(f_s - \overline{f_s}\right)}{\sqrt{\left(h_t - \overline{h_t}\right)^2}\sqrt{\left(f_s - \overline{f_s}\right)^2}} \tag{16}$$

where $h_t$ represents the output result of the RoBERTa-wwm layer, which is the auxiliary classifier, and $f_s$ represents the output result of the DFENN layer, which is the main classifier. $W$ represents the weight matrix, represents the classifier mean, and $S$ is the covariance matrix. The feature weights of the two layers are obtained using the function, and then the new features are passed into the CRF layer by multiplying the vector features of these two granularities.

### 4.4. CRF Layer

When performing named entity recognition tasks in the field of power marketing, the DFENN model, although capable of extracting global and local features of text sequences, does not handle the dependencies between adjacent tags. To solve this problem, Conditional Random Fields (CRF) are introduced, and CRF can obtain the globally optimal tag sequence by considering the relationship between neighboring tags. For the input sequence $X = (X_1, X_2, \ldots, X_n)$, extracting features can obtain the output matrix $P = (P_1, P_2, \ldots, P_n)$, and for the prediction sequence $Y = (Y_1, Y_2, \ldots, Y_n)$, its score function is calculated as shown in Equation (17).

$$S(X, Y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \tag{17}$$

In the equation, $A_{y_i, y_{i+1}}$ represents the fraction of $y_i$ transferred to $y_{i+1}$, and $P_{i, y_i}$ represents the fraction of the character predicted to be the first label. Firstly, the softmax layer is used to count the probability of all possible labels identified, and finally the label sequence with the highest probability is output.

When building the corpus in the field of power marketing, due to the large difference in the frequency of some entity types in the power marketing data, it leads to the problem of unbalanced sample distribution in the corpus. For example, the number of entity types "equ" far exceeds the number of "name" entity types, which leads to an imbalance in the

distribution of the loss function during the training process, and eventually the model will prefer the labels with a large number of samples, and for the samples. The recognition performance of labels with fewer samples is poor. In this paper, we use the focal loss function to alleviate the problem caused by the unbalanced distribution of samples, and the predicted entities are passed through the focal loss function to obtain the loss values, and the algorithm is shown in Equation (18).

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \tag{18}$$

In the equation, $p_t$ represents the probability that the sample is correctly identified, and the larger its value represents the easier the sample is identified. When the sample is correctly identified, $p_t \to 1$, whereas when the sample is incorrectly identified, $1 - p_t \to 1$. In this paper, by adjusting for different values of $\alpha$ and $\gamma$ in the focal loss function, $\alpha$ set to 0.25 and $\gamma$ set to 2 can achieve the best results. The model reduces the loss for accurately identified samples during the training process and increases the weight of the hard-to-identify samples in the loss function, thus, the model's recognition performance improves.

## 5. Experiments and Results Analysis

This experiment uses GPU to train the model, taking advantage of its high-performance parallel computing to improve the computing efficiency. The hardware and software environment used for the experimental configuration is shown in Table 3. During the training process, the parameters are continuously optimized based on the training results, and the parameter settings of the model in this paper are detailed in Table 4, and the original parameters are used for the RoBERTa pre-trained model.

**Table 3.** Software and hardware configuration table.

| Type | Configuration Items | Configuration Parameters |
|------|---------------------|--------------------------|
| Software | developing language<br>Software Environment<br>Experimental platform | Python 3.8<br>PyTorch 1.11.0<br>Ubuntu 20.04 |
| Hardware | Graphics card<br>Memory | RTX3090<br>24 G |

**Table 4.** Model parameter setting.

| Parameter | Value |
|-----------|-------|
| Learn rate | $1 \times 10^{-5}$ |
| Batch size | 4 |
| Epoch | 50 |
| Lstm_embedding_size | 1024 |
| Hidden size | 768 |
| Bert model | RoBERTa-wwm-ext |
| Embedding size | 512 |
| Optimizer | AdamW |
| Dropout | 0.5 |
| Number of layers of IDCNN | 3 |
| Expansion width | 1, 2, 4 |
| Number of IDCNN blocks | 4 |

### 5.1. Evaluation Indicators

This experiment uses common metrics for named entity recognition to measure the performance of the model, which are precision (*P*), recall (*R*) and *F*1 value. The specific evaluation formulas are as follows:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{19}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{20}$$

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\% \tag{21}$$

*P* represents the number of correctly identified entities as a percentage of the total number of predicted entities. *R* represents the number of correctly identified entities as a percentage of the total number of entities labeled in the sample. And *F*1 value combines the precision and recall rates, which can be evaluated for the model as a whole. Meanwhile, *TP* indicates the number of entity labels in the test set correctly identified by the model, *FP* indicates the number of entity labels in the test set incorrectly identified by the model, and *FN* indicates the number of entity labels in the test set not identified by the model.

*5.2. Analysis of Results*

This paper uses the constructed named entity recognition dataset in the field of power marketing for training evaluation, and uses precision, recall and *F*1 value as the evaluation criteria to measure the performance of the model. Meanwhile, four sets of comparison experiments are set up to verify and analyze the effectiveness of the named entity recognition method proposed in this paper for the power marketing domain.

(1)   Comparison of the performance of different masking strategies

In order to verify the effectiveness of the whole word masking strategy in improving named entity recognition capability in the field of power marketing, comparative experiments were carried out on the model DFENN-CRF with no masking strategy, the model BERT-DFENN-CRF using a word-level masking strategy, and the model RoBERTa-DFENN-CRF using a whole word masking strategy. The experimental results are shown in Table 5.

**Table 5.** Comparison experiments of different masking strategies.

| Masking Strategy | Precision/% | Recall/% | *F*1 Value/% |
| :---: | :---: | :---: | :---: |
| No Masking | 82.91 | 81.52 | 82.21 |
| Word-level Masking | 85.72 | 83.63 | 84.66 |
| Whole Word Masking | 88.23 | 86.56 | 87.39 |

As can be seen from Table 5, the model that utilizes the whole word masking exhibits the best performance with a precision rate of 88.23%. In addition, when comparing whole word masking to word-level masking, the accuracy, recall and *F*1 values of the model improved by 2.51%, 2.93% and 2.73%, respectively. Moreover, compared to the no masking strategy, the accuracy, recall and *F*1 values of the model improved by 5.32%, 5.04% and 5.18%, respectively. Without using the masking strategy, the model outputs word vectors lack contextual semantic information, which makes it difficult to solve the problem of words with multiple meanings, leading to relatively poor recognition performance with a precision rate of 82.91%. By implementing the word-level masking strategy, the model randomly masks the characters during pre-training, and then allows the model to predict the masked characters, so that the encoder can retain the contextual semantic representation of each character, to some extent, which solves the problem of multiple word meanings by using the contextual information with the precision rate of 85.72%. Whole word masking is compared to word-level masking by first splitting the power marketing text and then randomly masking the words, and then allowing the model to predict the masked words, so that the model can learn the complete word-level semantic information and improve the inference and representation of Chinese semantics, thus the performance of the model is further improved with the precision rate of 88.23%.

(2)  Comparison of the performance of different feature extraction methods in the middle layer of the model

In order to verify the effectiveness of different feature extraction approaches in the middle layer of the model for improving the recognition of named entities in the power marketing domain, comparative experiments were conducted on RoBERTa-CRF, RoBERTa-BiLSTM-CRF, RoBERTa-IDCNN-CRF, and RoBERTa-DFENN-CRF. The results of the experiments are shown in Table 6.

**Table 6.** Comparison experiments of different feature extraction methods in the middle layer.

| Number | Model | Precision/% | Recall/% | $F1$ Value/% |
|--------|-------|-------------|----------|--------------|
| 1 | RoBERTa-CRF | 84.34 | 83.68 | 84.01 |
| 2 | RoBERTa-BiLSTM-CRF | 85.55 | 84.92 | 85.23 |
| 3 | RoBERTa-IDCNN-CRF | 85.32 | 84.83 | 85.07 |
| 4 | RoBERTa-DFENN-CRF | 88.23 | 86.56 | 87.39 |

From the experimental results in Table 6, we can see that the intermediate layer has the highest $F1$ value of 87.39% using the DFENN model. This indicates that adding a dual feature extraction neural network in the middle layer can improve the effectiveness of named entities. Model 1 achieved the lowest $F1$ value because it did not add any intermediate layer network, and although the encoder was able to retain the contextual semantic representation of each character, the absence of an intermediate layer for global semantic information and local semantic information extraction led to the worst final results. Compared with model 3, model 2 has improved all indicators, but the difference is not significant, which is because the intermediate layers of these two models can further extract semantic information. The former is better at learning the features of the whole sentence and capturing the long-distance context dependence of the text, which is better for the recognition of entities with longer length; while the latter focuses more on the information and features around the entities, which can better distinguish the entity boundaries. For example, there are some obvious features around such entities as voltage level, such as the number in front of "KV", and the entity type of station usually ends with "站", so the obvious features can be captured for correct labeling. Model 4 adds a dual feature extraction neural network after the pre-trained model, which can make up for the deficiencies of BiLSTM and IDCNN, and thus achieves the best results with the precision rate of 87.39%.

(3)  Comparative analysis of the performance of different models

To verify the performance of the PMDNER model in the recognition of named entities in the field of power marketing, comparative experiments were conducted with BiLSTM-CRF, BERT-BiLSTM-CRF, RoBERTa-BiLSTM-CRF, RoBERTa-IDCNN-CRF, RoBERTa-DFENN-CRF and RoBERTa-DFENN-Att-CRF models. The results of named entity recognition in the field of power marketing are shown in Table 7 and Figure 7. From the experimental results, it can be seen that the proposed model 7 has better recognition performance than other models, with an $F1$ value of 88.58%. When compared to the BiLSTM-CRF benchmark model without the use of pre-trained models, the $F1$ value has an improvement of 7.36%.

**Table 7.** Comparative experiments with different models.

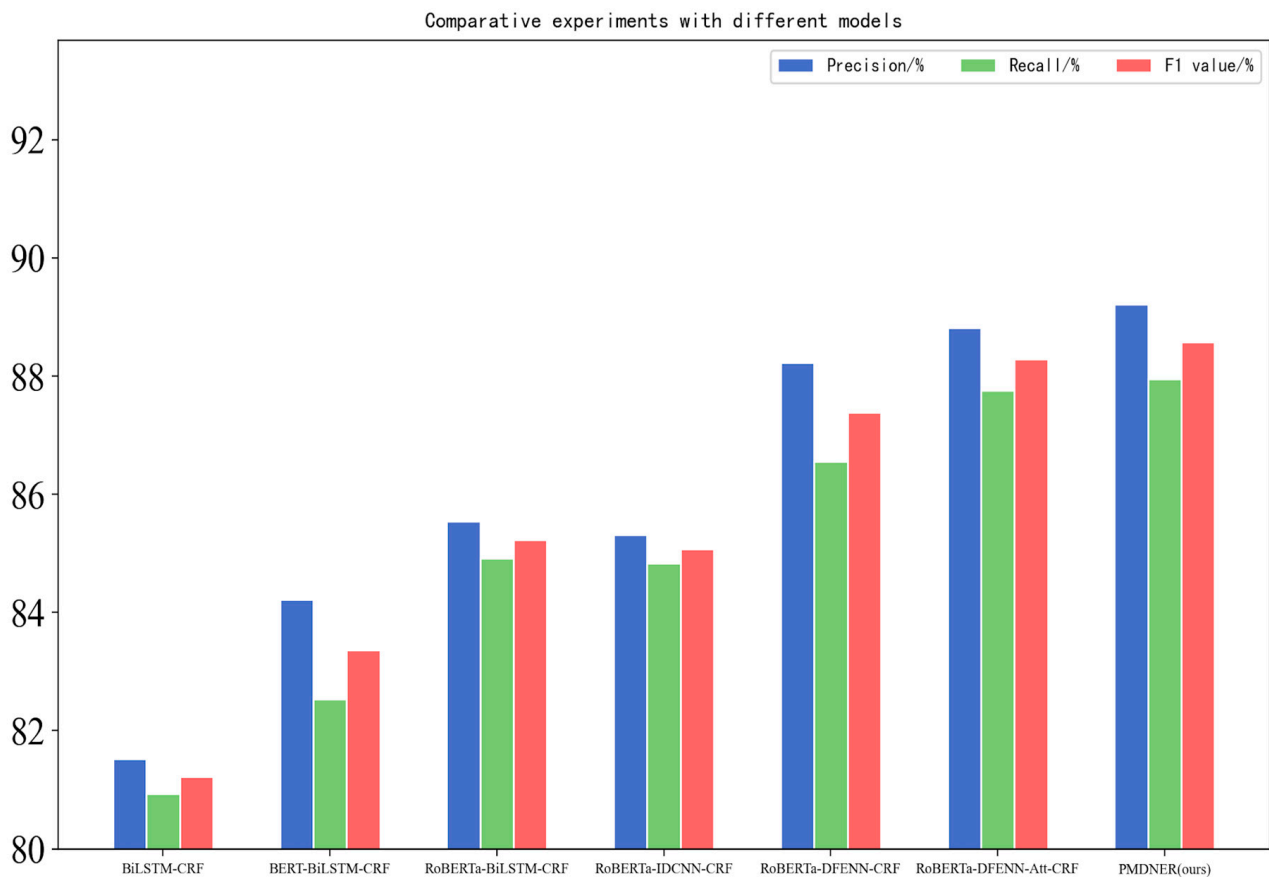| Number | Model | Precision/% | Recall/% | $F1$ Value/% |
|--------|-------|-------------|----------|--------------|
| 1 | BiLSTM-CRF | 81.52 | 80.93 | 81.22 |
| 2 | BERT-BiLSTM-CRF | 84.22 | 82.53 | 83.37 |
| 3 | RoBERTa-BiLSTM-CRF | 85.55 | 84.92 | 85.23 |
| 4 | RoBERTa-IDCNN-CRF | 85.32 | 84.83 | 85.07 |
| 5 | RoBERTa-DFENN-CRF | 88.23 | 86.56 | 87.39 |
| 6 | RoBERTa-DFENN-Att-CRF | 88.82 | 87.76 | 88.29 |
| **7** | **PMDNER (ours)** | **89.22** | **87.95** | **88.58** |

**Figure 7.** Comparative experiments with different models.

After using the BERT pre-trained model, model 2 can more fully consider the location information and contextual semantic information of the characters, and improves in recognition with a 2.15% improvement in the $F1$ value compared to model 1.

Regarding the selection of pre-trained models, model 3, which utilizes RoBERTa, demonstrates further improvement with a 1.86% increase in the $F1$ value compared to model 2. This could be attributed to the whole word masking strategy employed by RoBERTa pretrained model, which can better characterize the Chinese semantics and is more applicable to Chinese named entities.

The $F1$ value of model 5 after using the DFENN proposed in this paper is also improved by 2.16% compared with model 3 and 2.32% compared with model 4. This is because the dual feature extraction neural network proposed in this paper can obtain the global feature information and local feature information of the input text in parallel, which makes up for the BiLSTM that only focuses on the extraction of full-text information but not local information, and also makes up for the shortcomings of IDCNN, which can only obtain local features but not long-range global features, and finally achieves good results in the named entity task in the field of power marketing.

Model 6 is improved from model 5 by using the output of the RoBERTa layer as the auxiliary classification layer and the output of the DFENN layer as the master classification layer with weighted fusion using the attention mechanism function. Finally, model 6 proposed in this paper improves 0.9% compared with model 5. It is because the word vector output of the RoBERTa layer incorporates rich contextual semantic information, which can learn the global features of text and the local feature information of text after feeding it into the DFENN neural network model. Finally, the output vectors of both are used to calculate the weights by the attention mechanism function, and then after the weighted fusion can be better for the sequence annotation of power marketing data.

Model 7 is based on model 6, which introduces a focal loss function to alleviate the problem of unbalanced sample distribution. By increasing the weight of the types with a small number of entities in the loss function, the model focuses more on the hard-to-identify samples during the training process and improves the recognition ability of the model for hard-to-identify samples, and the recognition effect of the model is better than all the above models, with an $F1$ value of 88.58%, which is the most ideal recognition effect.

## 6. Conclusions

In this paper, a named entity recognition dataset is constructed in the electric power marketing domain. With the data annotated by using the BIO annotation method, nine entity types are classified, and a PMDNER model is established for named entity recognition in the electric power marketing domain. The RoBERTa pre-training model is employed to implement word vectorization of the electric power text data, and then the semantic representation vector of RoBERTa output is input to the constructed DFENN neural network to obtain both local and global features of the text in parallel for further fusion. Utilizing the output of the RoBERTa layer of the model as the auxiliary classification layer and the output of the DFENN layer as the master classification layer, the weights of the two layers are calculated using the attention mechanism, and new features are fused based on the weights to be input to the CRF layer to predict the labels. Additionally, a focal loss function is used to alleviate the problem of uneven sample distribution during training. This method can better tackle the challenges such as multiple meanings of words, incomplete feature extraction and poor recognition of specialized terms in the process of entity recognition. The experiments demonstrate that the recognition effect of the model is significantly improved, and the accuracy of the model is 89.22%, the recall is 87.95%, and the $F1$ value is 88.58%, which is a 7.36% improvement compared with the baseline model BiLSTM-CRF.

In future work, more suitable feature fusion methods can be explored to further strengthen the recognition effect of named entity identification in the power marketing domain, and the correlation between named entities and entity relationship extraction can be investigated to scientifically and precisely mine the potential knowledge in the field of power marketing. In addition, it is also expected to apply the proposed model to data from other domains to verify the generalization ability of the model.

## References

1. Massaoudi, M.; Abu-Rub, H.; Refaat, S.S.; Chihi, I.; Oueslati, F.S. Deep learning in smart grid technology: A review of recent advancements and future prospects. *IEEE Access* **2021**, *9*, 54558–54578. [CrossRef]
2. Gunduz, M.Z.; Das, R. Cyber-security on smart grid: Threats and potential solutions. *Comput. Netw.* **2020**, *169*, 107094. [CrossRef]
3. Raza, M.A.; Aman, M.M.; Abro, A.G.; Tunio, M.A.; Khatri, K.L.; Shahid, M. Challenges and potentials of implementing a smart grid for Pakistan's electric network. *Energy Strategy Rev.* **2022**, *43*, 100941. [CrossRef]

4.  Amalina, F.; Hashem, I.A.T.; Azizul, Z.H.; Fong, A.T.; Firdaus, A.; Imran, M.; Anuar, N.B. Blending big data analytics: Review on challenges and a recent study. *IEEE Access* **2019**, *8*, 3629–3645. [CrossRef]
5.  Guo, H.; Davidson, M.R.; Chen, Q.; Zhang, D.; Jiang, N.; Xia, Q.; Kang, C.; Zhang, X. Power market reform in China: Motivations, progress, and recommendations. *Energy Policy* **2020**, *145*, 111717. [CrossRef]
6.  Mier, M.; Weissbart, C. Power markets in transition: Decarbonization, energy efficiency, and short-term demand response. *Energy Econ.* **2020**, *86*, 104644. [CrossRef]
7.  Lauriola, I.; Lavelli, A.; Aiolli, F. An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing* **2022**, *470*, 443–456. [CrossRef]
8.  Shaik, T.; Tao, X.; Li, Y.; Dann, C.; McDonald, J.; Redmond, P.; Galligan, L. A review of the trends and challenges in adopting natural language processing methods for education feedback analysis. *IEEE Access* **2022**, *10*, 56720–56739. [CrossRef]
9.  Kaplar, A.; Stošović, M.; Kaplar, A.; Brković, V.; Naumović, R.; Kovačević, A. Evaluation of clinical named entity recognition methods for Serbian electronic health records. *Int. J. Med. Inform.* **2022**, *164*, 104805. [CrossRef]
10. Gligic, L.; Kormilitzin, A.; Goldberg, P.; Nevado-Holgado, A. Named entity recognition in electronic health records using transfer learning bootstrapped neural networks. *Neural Netw.* **2020**, *121*, 132–139. [CrossRef]
11. Qiao, B.; Zou, Z.; Huang, Y.; Fang, K.; Zhu, X.; Chen, Y. A joint model for entity and relation extraction based on BERT. In *Neural Computing and Applications*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 1–11.
12. Guo, L.; Yan, F.; Li, T.; Yang, T.; Lu, Y. An automatic method for constructing machining process knowledge base from knowledge graph. *Robot. Comput.-Integr. Manuf.* **2022**, *73*, 102222. [CrossRef]
13. Lin, T.H.; Huang, Y.H.; Putranto, A. Intelligent question and answer system for building information modeling and artificial intelligence of things based on the bidirectional encoder representations from transformers model. *Autom. Constr.* **2022**, *142*, 104483. [CrossRef]
14. Patil, N.; Patil, A.; Pawar, B.V. Named entity recognition using conditional random fields. *Procedia Comput. Sci.* **2020**, *167*, 1181–1188. [CrossRef]
15. Li, J.; Fang, S.; Ren, Y.; Li, K.; Sun, M. SWVBiL-CRF: Selectable Word Vectors-based BiLSTM-CRF Power Defect Text Named Entity Recognition. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data). IEEE, Atlanta, GA, USA, 10–13 December 2020; pp. 2502–2507.
16. Wu, G.; Tang, G.; Wang, Z.; Zhang, Z.; Wang, Z. An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition. *IEEE Access* **2019**, *7*, 113942–113949. [CrossRef]
17. Li, Y.; Du, G.; Xiang, Y.; Li, S.; Ma, L.; Shao, D.; Wang, X.; Chen, H. Towards Chinese clinical named entity recognition by dynamic embedding using domain-specific knowledge. *J. Biomed. Inform.* **2020**, *106*, 103435. [CrossRef]
18. Luo, L.; Yang, Z.; Yang, P.; Zhang, Y.; Wang, L.; Lin, H.; Wang, J. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics* **2018**, *34*, 1381–1388. [CrossRef]
19. Xu, K.; Yang, Z.; Kang, P.; Wang, Q.; Liu, W. Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition. *Comput. Biol. Med.* **2019**, *108*, 122–132. [CrossRef]
20. Batbaatar, E.; Ryu, K.H. Ontology-based healthcare named entity recognition from twitter messages using a recurrent neural network approach. *Int. J. Environ. Res. Public Health* **2019**, *16*, 3628. [CrossRef]
21. Wei, H.; Gao, M.; Zhou, A.; Chen, F.; Qu, W.; Wang, C.; Lu, M. Named entity recognition from biomedical texts using a fusion attention-based BiLSTM-CRF. *IEEE Access* **2019**, *7*, 73627–73636. [CrossRef]
22. Yan, R.; Jiang, X.; Dang, D. Named entity recognition by using XLNet-BiLSTM-CRF. *Neural Process. Lett.* **2021**, *53*, 3339–3356. [CrossRef]
23. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
24. Meng, F.; Yang, S.; Wang, J.; Xia, L.; Liu, H. Creating knowledge graph of electric power equipment faults based on BERT–BiLSTM–CRF model. *J. Electr. Eng. Technol.* **2022**, *17*, 2507–2516. [CrossRef]
25. Li, X.; Zhang, H.; Zhou, X.H. Chinese clinical named entity recognition with variant neural structures based on BERT methods. *J. Biomed. Inform.* **2020**, *107*, 103422. [CrossRef] [PubMed]
26. Wu, Y.; Huang, J.; Xu, C.; Zheng, H.; Zhang, L.; Wan, J. Research on named entity recognition of electronic medical records based on roberta and radical-level feature. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 2489754. [CrossRef]
27. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef]
28. 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]
29. Weiyue, T. Named entity recognition of power communication planning based on transformer. In Proceedings of the 2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). IEEE, Chongqing, China, 17–19 June 2022; Volume 10, pp. 588–592.