IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 1, 2021, accepted October 17, 2021, date of publication October 26, 2021, date of current version November 18, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3123154

Named Entity Recognition in Electric Power Metering Domain Based on Attention Mechanism

KAIHONG ZHENG[1,2], LINGYUN SUN[2,3], XIN WANG[2,3,4], SHANGLI ZHOU[1], HANBIN LI[4], SHENG LI[1], LUKUN ZENG[1], AND QIHANG GONG[1]

[1]Digital Grid Research Institute, China Southern Power Grid, Guangzhou, Guangdong 510663, China
[2]College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310058, China
[3]Zhejiang University-China Southern Power Grid Joint Research Centre on AI, Zhejiang University, Hangzhou, Zhejiang 310058, China
[4]College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, Zhejiang 310014, China

Corresponding author: Xin Wang (wangxin2009@zju.edu.cn)

This work was supported by the National Key Research and Development Program of China under Grant 2020YFB0906004 and Grant 2020YFB0906000.

ABSTRACT Named Entity Recognition (NER) is one key step for constructing power domain knowledge graph which is increasingly urgent in building smart grid. This paper proposes a new NER model called Att-CNN-BiGRU-CRF which consists of the following five layers. The prefix Att means the model is based on attention mechanism. A joint feature embedding layer combines the character embedding and word embedding based on BERT to obtain more semantic information. A convolutional attention layer combines the local attention mechanism and CNN to capture the relationship of local context. A BiGRU layer extracts higher-level features of power metering text. A global multi-head attention layer optimizes the processing of sentence level information. A CRF layer obtains the output tag sequences. This paper also constructs a corresponding power metering corpus data set with a new entity classification method. The novelties of our work are the five layer model structure and the attention mechanism. Experimental results show that the proposed model has high recall rate 88.16% and precision rate 89.33% which is better than the state-of-the-art models.

INDEX TERMS Power metering, attention mechanism, joint feature, named entity recognition.

I. INTRODUCTION

Knowledge graph in the domain of power metering can express the entities, relationships and concepts in the business activities of power metering in a structured form, and provide a more effective big data organization, management and cognitive ability [1]. In addition, knowledge graph can also play an important role in information retrieval and decision-making assistance in the electric power domain [2]–[4]. It is feasible and meaningful to construct a domain knowledge graph by extracting entity information from power metering literature which contains a large number of professional theoretical knowledge and cutting-edge technical methods.

Named Entity Recognition (NER) is a key technology in Natural Language Processing (NLP). It is also a basic task to construct knowledge graph. The purpose of NER is to extract the required entity elements from the texts which include various types of data, such as person names, place names, organization names, and professional terms. The quality of Chinese NER models based on neural networks is affected by different word embedding representations and usage of dictionary. NER technology in the domain of power metering mainly has the following difficulties:

1) The generic domain entity classification methods are not applicable in the power metering domain. Especially some complicated terms are difficult to determine. Such as "限流保护 (over current protection)", "过压保护 (over voltage protection)" and "低压电流互感器 (low-voltage current transformer)". As there doesn't exist entity classification standard in power metering domain, it is necessary to establish a set of entity classification methods with reasonable boundary delineation according to the domain requirements.

2) The corpus usually contains abbreviations of professional technical terms in the power metering domain.

The associate editor coordinating the review of this manuscript and approving it for publication was Manuel Rosa-Zurera.

These terms are not easily recognized as their format is simplified. For example, ''Alternating Current'' is also called ''AC''.

3) Compared with English NER, Chinese NER requires additional word segmentation processing [5]. The precision of generic word segmentation methods doesn't satisfy the requirement of specialized domain words.

In order to solve the above problems, this paper proposes a new entity classification method for the power metering domain and a new NER model together to improve the recognition effect of power metering entities and the construction effect of knowledge graph. The contributions of this paper are as follows:

1) A new classification method for power metering entities is proposed. Using this method, a basic complete power metering corpus is constructed by labeling the entities in the corpus data. At the same time, the professional vocabularies obtained during the labeling process have been integrated into the word segmentation dictionary, which further improves the precision of word segmentation.

2) An embedding representation method that combines character-level and word-level information is proposed. It use an attention mechanism to assign weights to both types of embedding. These weights are used as inputs to the model. Which makes it easy to capture word structure features in Chinese text and identify Chinese abbreviations effectively.

3) A NER model combining attention mechanism named Att-CNN-BiGRU-CRF is proposed. The commonly used LSTM is replaced by GRU and the performance is improved. In addition, a local attention layer is added to the CNN layer and a global multi-headed attention layer is added after the GRU. The purpose of these model structure modification is to calculate the relevance weights of the text sequences and to obtain more character-level, word-level and sentence-level features.

The experimental results show that our model has better recognition effects than previous models, such as BiLSTM-CRF and CNN-BiGRU-CRF et al.

II. RELATED WORK

In the early days, NER techniques mainly used the method of writing rules manually. Later, many methods based on statistical models emerged. Most of these methods train linear statistical models on a corpus with certain annotations. These statistical models are typically Hidden Markov Model (HMM) [6], Support Vector Machine (SVM) [7] and Conditional Random Field (CRF) [8]. These methods are more heavily dependent on manual features and specific training data. They are expensive to develop, and work well only for small datasets.

Later improvements were using deep neural networks to optimize entity recognition. Some of the representative works are surveyed as follows: Lample *et al.* [9] developed a method of combining Bi-directional Long Short-Term Memory (Bi-LSTM) with CRF in English sequence tag prediction tasks. The results are identical to the best statistical methods. There is no need to manually define features. In order to detect both word-level and character-level features, Chiu and Nichols [10] used CNNs for extracting character-level features to form a BiLSTM-CNNs model. Which effectively improved recognition while eliminating most feature engineering requirements. Ma and Hovy [11] proposed a BiLSTM-CNN-CRF model to construct word-level and character-level representations. It do not rely on specific domain data. Its experimental results based on 7 languages show that character-level features are better than word-level features.

In recent years, the attention mechanism has been rapidly developed in computer vision. Some researchers attempt to use it in entity recognition tasks as its ability to efficiently extract features from high-dimensional data. Attention mechanism is modeled after the selective attention mechanism used by the human body in visual observation, when more attention is applied to the key target parts to be observed to obtain more detailed information. In the case of this paper's task, the target becomes a sequence of power metering texts, where the character, the word, the sentence mentioned at the current moment are firstly associated with a corresponding parts of the power metering technology. Its relevance changes continuously as the text sequence advances. By applying the attention mechanism to the Encoder-Decoder model structure [12], the model can obtain the information related to the next word to be processed in the input text sequence, and select the key information of the task requirement from the large amount of power metering text information. As the attention mechanism weakened the influence of irrelevant information, the overall effect was improved. Rei *et al.* [13] proposed a method for combining word vectors and character vectors based on attention mechanism. The method used the traditional method that feature vectors were learned with English vocabulary as the basic element, and proposed that character-level information from the vocabulary is necessary to participate in the training process. Andrej *et al.* [14] used the multi-headed Encoder-Decoder structure to achieve better results in multi-label text classification tasks. Yin *et al.* [15] combined BiLSTM and soft attention model to deeply extract character-level features to identify named entities. BiLSTM was used to extract global text features. Soft attention model was used to extract local text features. That work shows that the attention mechanism can improve the performance of Chinese NER model.

The above representative methods are basically oriented to generic domains. The languages of these test dataset were mainly English. Inspired by the idea of combining multiple neural networks, Chinese researchers have also proposed some improved NER methods for Chinese language or specific domains. Such as Wang *et al.* [16], Xu *et al.* [17], Wu *et al.* [18], and Luo *et al.* [19] have proposed NER models using attention mechanisms in social media domain and medical domain.

Noting the fact that Chinese language use character and word combinations to express their meanings. If we use

character vector or word vector separately, the NER model will lose certain semantic information. So, Shan *et al.* [20], Xu *et al.* [21], and Dong *et al.* [22] combine word-level, character-level and radical-level information in Chinese characters to extract more semantic information and enhance the Chinese entity recognition. Compared with these methods which directly concatenate each feature embedding, Yu *et al.* [23] and Jia *et al.* [18] applied an attention mechanism to assign different weights to the union of each embedding to obtain more optimized results. Yang *et al.* [24] use the BERT model to pre-train word embeddings. Their input data has richer word information than other models.

In the power domain, Feng *et al.* [25] and Tian *et al.* [26] both used the BiLSTM-Attention neural network method to classify text. Although these methods have good precision on specific content data sets, they cannot specifically identify entity information and are difficult to be directly used in the construction of knowledge graph. Fan *et al.* [27] used semantic tagging information such as clauses, word segmentation, and part-of-speech tagging as a preprocessing method in NER for the entire business domain of the power grid. Zhao *et al.* [28] used a BiLSTM-CRF model on two categories of power data. Jiang *et al.* [29] proposed a NER model combining BERT, Bi-LSTM, and CRF for power domain.

Among the various previous works, these models have some shortcomings for the NER task in power metering domain.

1) BiLSTM-CRF The BiLSTM-CRF proposed by Lample *et al.* [9] is one of the most fundamental and commonly used models for NER task. Many new entity recognition models are improved from it. As a generic NER model, BiLSTM-CRF is weak in its ability to capture and understand the semantic relationships of word contexts in power metering domain.

2) BiLSTM-CNN-CRF Ma *et al.* [11] added a convolutional neural network to the BiLSTM-CRF.That can improve the model's ability to extract contextual features of words. However, the main drawback of this model is its low recognition precision of complicated words.

3) NER model using attention mechanism The research work of Rei *et al.* [13], Andrej *et al.* [14], and Yin *et al.* [15] involve attention mechanism. The attention mechanism gives more weights to important words that can optimize the priority of the model for selection of lexical contextual features to a certain extent and improve the recognition ability for complicated words. However the input vector of these model contains less semantic information which affects the recognition effect.

In the work of Shan *et al.* [20], Xu *et al.* [21] and Dong *et al.* [22], they found a joint feature embedding vector which is constructed from combining character-level features with word-level features. The joint feature embedding vector can obtain richer semantic information. So using joint feature embedding may overcome the above shortcomings.

Now the NER methods using joint feature embedding and attention mechanisms are still facing many challenge problems. This paper proposes an innovational NER model which combine these two factors with neural networks such as CNN, Bi-GRU, and CRF for the power metering domain.

III. CONSTRUCTING CORPUS DATASETS

A. POWER METERING CORPUS PREPARATION

In order to verify the effect of our NER model, we need to train and test on the corpus of the power metering domain. However, there doesn't exist an open dataset for the power metering domain which can be directly used for the training and testing of NER. Furthermore, the distribution of power metering knowledge resources on the web is scattered, so a power metering corpus in a certain scale needs to be constructed.

The power metering corpus is collected from a wide range of data sources. They are mainly various encyclopedic knowledge data in the web, literature and books of the power metering domain, and internal business information of relevant power grid company. Most of these data are unstructured and include definitions of concepts, technical principles and applications of various terms.

For the power metering encyclopedic knowledge data, a crawler script program is constructed using the SCRAPY web crawler framework to crawl data from all kinds of power websites. The crawled data should be filtered to remove duplicated data and non-electricity metering related data.

For literature data and internal data of electric power companies, we can use many better classification information with little manual work.

In our work, data pre-processing is carried out for various types of data. Such as standardizing and unifying the format, deleting irrelevant text corpus, and dividing statements in the corpus with the period as separator.

B. POWER METERING ENTITY CLASSIFICATION METHOD

Due to the strong professionalism in the domain of electric power metering, it makes the classification of entities more diverse and complex. There always exist some difficult problems with blurred boundaries of named entities and inconsistent labeling. For example, ''power meter stop'' can be regarded as a type of electric power phenomenon, and it can also be regarded as ''power meter'' as a type of electric power equipment, and ''stop'' as a type of electric power phenomenon.

We design a new entity classification method by referring to the electrical terminology section of Chinese National Standards which include industry standard terms related to power metering technology, power metering instruments, power supply and power consumption, as well as the metering business information from power grid company. The method can better delineate various types of terminology and reduce the occurrence of blurred boundaries. We divide the collected corpus into five categories of entities which are metering data, metering technology, power equipment, power organization,

and power phenomena. We use that classification method to annotate and construct a power metering corpus for training and testing. The details are as follows:

Metering data: We label the parameter names of the power data monitored by the power metering system and the index data names obtained by statistical analysis as metering data entities, such as ''coverage'', ''total power consumption'', ''average load'', etc.

Metering technology: We label the technical methods or technical behavior related to power metering as metering technology entities, such as ''voltammetry'', ''automated meter reading'', ''installation inspection'', etc.

Power equipment: We label the equipment and devices used for power metering work as power equipment entities, such as ''voltmeter'', ''transformer box'', ''junction box'', etc.

Power organization: We label the people, regions and institutions involved in power metering as power organization entities, such as ''power maintenance and repair workers'', ''power supply districts'', ''SGCC'', etc.

Power phenomena: We label the phenomena generated in the process of power metering as power phenomena entities, such as ''phase sequence abnormalities'', ''voltage overruns'', ''power theft'', etc.

By judging the objects described by the domain specific words and the meanings they expressed, the entity types can be decided. That can reduce the inconsistency of backward and forward labeling and improve the precision of NER.

C. STATISTICS OF ENTITY LABELING OF OUR CORPUS

We use the above classification method to label the corpus data from multiple data sources and divide the entities into five defined categories of labels. We adopt BIO labeling method and use the tool YEDDA [30] for entity labeling which can segment each character in the labeled corpus. Each character and the corresponding label are on a separate line, with a blank line at the end of each sentence, which is consistent with the corpus format of the public data set Conll2003, so that the program can read the corpus correctly in standard manner. There are total of 18,665 sentences, including 22,874 entities. The specific number of each entity type in the data set is shown in Table 1.

TABLE 1. Corpus entity count.

Type of Entities	Number of Entities (Proportion/%)
Metering Data	5528 (24.2)
Metering Technology	4076 (17.8)
Power Equipment	7421 (32.4)
Power Organization	2164 (9.5)
Power Phenomenon	3685 (16.1)

D. CHARACTER VECTORS AND WORD VECTORS INITIALIZATION

The training data should be firstly converted to the input format required by the Word2Vec Tool released by Google Inc. The parameters of the Word2Vec tool are set up properly, such as the dimension of the word vector, the size of the window, the minimum number of occurrences of the word to be trained, and the number of threads, etc.

The Word2Vec tool can generate character embedding for each word in the sentence. We use the Skip-Gram word2vec model to transform each word into a corresponding vector. The Skip-Gram model obtains the probability distribution of each character in the corresponding position in the word list by learning the weight matrix between layers. The input sentence can be represented as $s = [w_1, w_2, \ldots, w_n]$. The n denotes the number of words in the sentence. The one-hot vector of the i-th character is represented by W_i. The output of this transformation is a sequence of character vectors $[x_1, x_2, \ldots, x_n]$.The character vector of the i-th word in the sentence is represented by x_i.

IV. ATTENTION MECHANISM ENTITY RECOGNITION MODEL

The structure of our attention mechanism entity recognition model is shown in Figure 1. The model contains a joint feature embedding layer, a convolutional attention layer, a BiGRU layer, a global multi-head attention layer and a CRF layer. In this paper, we combine CNN with local attention mechanism to form a convolutional attention layer, and apply global attention mechanism in the middle of BiGRU and CRF layers. The model achieves good results on multi-level contextual feature extraction. The purpose of this model is to generate a learning function that maps the input sentences to a sequence of sentence labels.

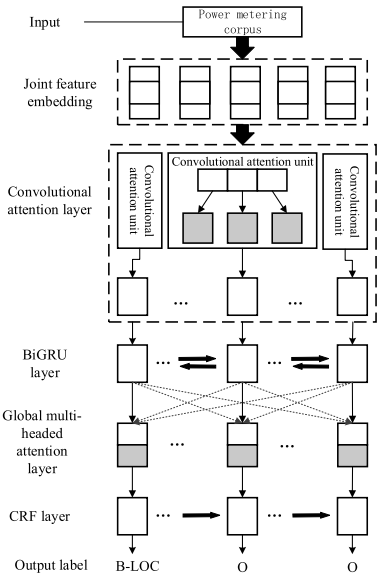

FIGURE 1. Structure diagram of entity recognition model.

A. CHARACTER AND WORD JOINT FEATURE EMBEDDING LAYER

We construct character feature embedding and word feature embedding separately by BERT, which uses a bidirectional Transformer as an encoder that fuses textual information from the left and right sides of the current character [31]. When training the character vector, instead of encoding sentences from left to right or right to left to predict characters, the encoder randomly hides or replaces some characters according to a certain ratio and predicts the original characters based on the context.

In this paper, we use the above power metering corpus for self-supervised learning of the BERT model, and the required character embedding and word embedding can be obtained by pre-training the acquired parameters. Compared with the traditional Word2Vec model, BERT has stronger generalization ability and the obtained character embedding and word embedding have richer semantic information.

Many words in power terminology have similar structures, and the character feature vectors can be used to accurately identify the entity boundaries. For example, "低压电流互感器 (low-voltage current transformer)" may be wrongly recognized as "低压电流 (low-voltage current)" and "互感器 (transformer)", instead of being one technical term.

In order to enable the model to dynamically select the semantic information of word feature vectors and character feature vectors [32], we construct the structure as shown in Figure 2. Each character embedding corresponds to the word embedding to which it belongs, e.g., "监" corresponds to "监测" as the first set of inputs, "测" corresponds to "监测" as the second set of inputs. The words, "用户 (customers)" and "用电量 (electricity consumption)", are processed in the same way. Specially the word "的" consists of one single character corresponds directly to the character "的". For the word vector sequence $\boldsymbol{x_{voc}}$ and the character vector sequence $\boldsymbol{x_{ch}}$ of the input sentence $\boldsymbol{X_i}$, we use bi-linear operations to capture the correlation information between them. We add the Relu activation function into the method proposed in the literature [32] to optimize the range of θ value. The final embedding representation is obtained by weighted summation of the $\boldsymbol{x_{voc}}$ and $\boldsymbol{x_{ch}}$.

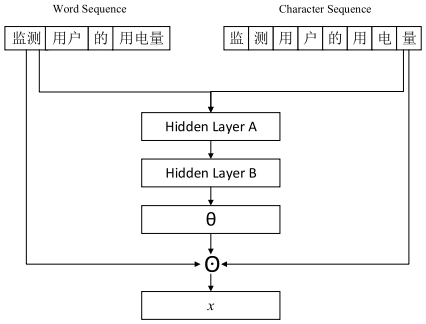

FIGURE 2. Attention based joint embedding.

In the formula (1) (2), $\boldsymbol{W_t}$, $\boldsymbol{W_r}$ and $\boldsymbol{W_c}$ are different weight matrices that are continuously tuned through training. Larger θ indicates higher importance and weight of word features, and vice versa. The model can be used in different texts to obtain critical semantic information as needed. We denote the dimensionality of character embedding and word embedding by d_{ch} and d_{voc}. The joint feature embedding dimension of character embedding and word embedding is denoted by d_e, and $d_e = d_{ch} = d_{voc}$.

$$\theta = \text{Relu}(W_t \tanh(W_r x_{voc} + W_c x_{ch})) \tag{1}$$

$$x = \theta \odot x_{voc} + (1 - \theta) \odot x_{ch} \tag{2}$$

B. CONVOLUTIONAL ATTENTION LAYER

The convolutional attention layer is using local attention mechanism on a convolutional neural network. After the input joint feature embedding, this layer can extract local contextual features by encoding the input character sequence and implicitly grouping meaningful related characters in the local context.

In this paper, we set the size of the sliding window in the CNN layer to k, and then merge the positional embedding $\boldsymbol{x_{pos}}$ with the joint embedding of word features so that the local contextual order information within the sliding window is preserved. $\boldsymbol{x_{pos}}$ is represented by one-hot vector, meaning the specific position of the character in the window, and its dimension d_{pos} has the value of k. The final input embedding dimension can be expressed as $d_{input} = d_e + d_{pos}$.

We apply local attention to the sliding window of the CNN layer to maximize the extraction of the semantic relationship between the center character of the window and the characters before and after it. The hidden dimension of this layer is represented by d_h. Assuming that the current input data is a n-th character, the input data in the window is the joint embedding $\mathbf{x}_{[\mathbf{n}-(\mathbf{k}-\mathbf{1})/\mathbf{2}]:[\mathbf{n}+(\mathbf{k}-\mathbf{1})/\mathbf{2}]}$ and the output data is the hidden vector $\mathbf{h}_{[\mathbf{n}-(\mathbf{k}-\mathbf{1})/\mathbf{2}]:[\mathbf{n}+(\mathbf{k}-\mathbf{1})/\mathbf{2}]}$. The number of both types of data is k. The hidden vector is calculated as follows:

$$h_m = \alpha_m x_m \tag{3}$$

And $m \in E$, $E = \{n-(k-1)/2, \ldots, n, \ldots, n+(k-1)/2\}$, α_m is the attention weight, which is calculated as:

$$\alpha_m = \frac{\exp s(x_n, x_m)}{\sum_{i \in E} \exp s(x_n, x_i)} \tag{4}$$

The attention score function s is used to measure the correlation between $\boldsymbol{x_n}$ and $\boldsymbol{x_m}$. A higher score means that elements near position m are more important and should be assigned a higher weight. The calculation formula is as follows:

$$s(x_n, x_m) = v^T \tanh(W_1 x_n + W_2 x_m) \tag{5}$$

And $v \in R^{d_h}$, and $\boldsymbol{W_1}, \boldsymbol{W_2} \in R^{d_h, d_{input}}$. The CNN layer sets d_h filters in a context window of size k. This layer uses multiple filters with different window size to learn contextual features. The results are expressed as follows.

$$h_j^c = \sum_k (W^c * h_{[n-\frac{k-1}{2}]:[n+\frac{k-1}{2}]} + b^c) \tag{6}$$

For the above equation, $\boldsymbol{W^c} \in R^{k \times d_h \times d_{input}}$, and $\boldsymbol{b^c} \in R^{k \times d_h}$. $\mathbf{h_{[n-(k-1)/2]:[n+(k-1)/2]}}$ denotes the joint set of hidden states $h_{i \in E}$. These data are firstly multiplied by elements in the same dimension, and then a summation pooling operation is performed on them in that dimension to accumulate the variables within each pooling window, and finally the output of the convolutional attention layer is obtained. The final output is $h_n^c = [h_1, h_2, \cdots, h_n]$, and $h_i \in R^{d_h}$.

C. BIGRU LAYER

The output of the convolutional attention layer are character-based contextual features, which are then used as input to the BiGRU layer for further computation as a way to capture long-term dependencies in the text sequence.

Gated Recurrent Unit (GRU) is a gating mechanism in Recurrent Neural Networks (RNN), which has excellent results in many sequence labeling tasks. Compared with LSTM, which uses multiple gating, GRU can complete the operation of forgetting and filtering memory by using only one gating, with simpler structure and fewer parameters, which makes it better in aspects of time cost and hardware requirement. Therefore, in this paper, BiGRU is chosen to replace the commonly used BiLSTM layer.

The BiGRU layer consists of two unidirectional GRUs with opposite directions. The BiGRU associates the output of the current moment with the state of both the previous moment and the next moment. Its purpose is to extract deep textual features from the input continuous sentence information. The formula of the BiGRU layer is as follows:

$$r = \sigma(W^r \cdot [x^t, h^{t-1}]) \tag{7}$$

$$z = \sigma(W^z \cdot [x^t, h^{t-1}]) \tag{8}$$

$$h' = \tanh(W \cdot [x^t, h^{t-1} \odot r]) \tag{9}$$

$$h_t = (1 - z) \odot h^{t-1} + z \odot h' \tag{10}$$

The reset gate and update gate are represented by r and z. The output of the BiGRU layer can be represented by Equation 11, where $\boldsymbol{h_n^c}$ is the output of the convolutional attention layer, $\boldsymbol{h_{n-1}^g}$ is the previous hidden state of the BiGRU layer. $\boldsymbol{W^r}$ and $\boldsymbol{W^z} \in R^{d_h \times d_h}$ are parameters.

$$h_n^g = BiGRU(h_{n-1}^g, h_n^c; W^r, W^z) \tag{11}$$

D. GLOBAL MULTI-HEADED ATTENTION LAYER

In this paper, we use a global multi-headed attention layer in order to better handle information at the sentence level. The self-attentive mechanism maps the query to a series of key-value pairs, on top of which the multi-headed attention mechanism is needed because the self-attentive mechanism cannot capture important features from multiple perspectives and levels. This mechanism maps the input character into multiple vector spaces and calculates the contextual representation of the character in the vector space. The model repeats this step several times, and finally stitches the results together to obtain a comprehensive character contextual feature. For example, the sentence "用电负荷过大会导致电力计量装置烧坏 (Excessive electrical loads can cause electrical metering devices to burn out)" is a common form of entity A will cause entity B for power metering professional expressions. Therefore, when there exists "导致 (cause)" in the input sequence, we should focus on the left and right components and give them more weight.

The formula for calculating the multi-headed attention mechanism is as follows

$$MulitiHead(h_n) = concat(score_1(h_n), score_2(h_n), \ldots, score_h(h_n))W^O \tag{12}$$

$\boldsymbol{h_n}$ is the output of the hidden layer of BiGRU. $\boldsymbol{score_i}$ is the output of the i-th layer self-attention mechanism, $score_i$ is calculated as follow:

$$score_i(h_n) = attention(h_n W_i^Q, h_n W_i^K, h_n W_i^V) \tag{13}$$

$\boldsymbol{W_i^Q}$, $\boldsymbol{W_i^K}$, $\boldsymbol{W_i^V}$ and $\boldsymbol{W^O}$ are parameter matrices used to map the input sequence h_n to different vector spaces. The parameter matrix sizes are $\boldsymbol{W_i^Q}, \boldsymbol{W_i^K}, \boldsymbol{W_i^V} \in R^{2d_h \times d_k}$ and $\boldsymbol{W^O} \in R^{2d_h \times 2d_h}$. The $d_k = 2d_h/h$, d_h is the dimension of the hidden layer OF BiGRU and h is the number of attention heads. The attention function is the expanded dot product of the self-attention mechanism operation. The formula is defined as follows:

$$attention(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V \tag{14}$$

$\sqrt{d}$ has a scaling adjustment function to avoid too large inner product. The output of the hidden layer of the BiGRU layer at j is defined as $\boldsymbol{h_j^g} = \boldsymbol{MulitiHead(h_n)}$, where $j = 1, \ldots, \tau$ denotes all character instances in the sentence.

E. CRF LAYER

We use the global multi-headed attention layer to obtain global feature information and then pass it as input data to the CRF layer for entity recognition of text sequences. CRF works well for sequence labeling problems and generally uses a linear chain conditional random field to solve the given input sequence by a maximum entropy model to predict the model one to one corresponding with the input sequence labeling.

In this paper, we denote the CRF layer by $\boldsymbol{H_t} = [\boldsymbol{h_t^r}; \boldsymbol{h_t^g}]$. Assuming that the predicted label sequence is $Y = \{y_1, y_2, y_3, \ldots, y_\tau\}$, the probability of the true label sequence is calculated by the following equation:

$$P(Y|X) = \frac{\exp(\sum_{k=1}^{K}(W_k H_i + b_k))}{\sum_{y'} \exp(\sum_{k=1}^{K}(W_k H_i + b_k))} \tag{15}$$

y' denotes an arbitrary tag sequence, $\boldsymbol{W_k}$ and b_k are trainable parameters. Finally, the decoding is performed by the Viterbi algorithm to obtain the recognition results of entity labels.

F. MODEL TRAINING

We fully exploit various types of semantic information based on characters, words and sentences in power metering text by using multi-feature joint embedding, CNN feature extractor with local attention and global multi-headed attention mechanism.

The log-likelihood function is used as the loss function in training the model. Given a set of training examples $\{(\boldsymbol{X_i}, \boldsymbol{Y_i})\}|_{i=1}^{K}$, the loss function L can be defined as follows:

$$L = \sum_{i=1}^{K} \log P(Y_i|X_i) \tag{16}$$

The training process is continuously iterated and the order of the training samples is randomized to gradually update the input samples to the model in batches of a certain size. Finally, the AdaDelta [33] algorithm is used to optimize all parameters to achieve the final goal.

V. EXPERIMENTS AND ANALYSIS

A. DATA SETS

We construct a corpus of power metering based on multiple data sources. This corpus includes 356 documents of power metering domain, with a total of 18,665 sentences. There are 22,874 technical terms that are classified into five categories. We divide the power metering corpus into training set, test set, and verification set according to the ratio of 8:1:1. The training set has 284 documents, the test set has 36 documents, and the verification set has 36 documents. The evaluation indicators used in the experiment include precision rate P, recall rate R, and F.

We use the jieba word segmentation tool to divide the corpus text and add the counted entity names to the dictionary. The word segmentation results obtained have an precision of 96.2%. If the length of the vocabulary is long, it is more prone to word segmentation errors. However, these errors will attenuate the weight under the adjustment of the joint embedding and attention mechanism. Only part of the vocabulary will be biased, so the impact is small.

The experiment environment is set up as follows. the operating system is Windows 10, the processor is Intel Xeon E5-2695, and the memory is 32GB. The programming language is Python 3.5, and Pytorch is used as the deep learning framework. We used Early-stop [34], [35] and Dropout [36] to avoid the model overfitting problem. At the same time, we use Adam [37] to update the model parameters to optimize its performance. Table 2 records the details of the experiment parameters.

B. ANALYSIS OF RESULTS

In order to compare the effects of different parameters on the experiment results and analyse the best experiment results from different models, the following experiments are carried out. The training set and test set used in each experiment are the same.

TABLE 2. Parameters of Att-CNN-BIGRU-CRF model.

Name of Parameters	Value of Parameters
embed_dim	100
dropout	0.5
GRU_hiddens	200
learning_rate	0.006
Batch_size	32
decay_rate	0.05

1) EXPERIMENT 1

The recognition effect of each model on different entity category. The final data is the average of the results of three experiments as shown in Table 3.

TABLE 3. Recognition result of each models on different entity category.

Entity Category (F-Value)	BiLSTM-CRF	CNN-BiLSTM-CRF	Att-CNN-BiGRU-CRF
Metering Data	86.17	89.22	92.86
Metering Technology	84.54	86.08	88.79
Power Equipment	86.82	87.13	90.31
Power Organization	80.66	81.74	84.65
Power Phenomenon	85.35	84.61	87.42

According to the data in Table 3, we can see that the F-values of metering data category are higher because the expression of this category data is uniform and standardized. The F-value of power organization category is lower, partly because the number of entities of this category is small, and partly because some power organizations have long names and various combinations, which are difficult to be fully recognized. The number of power equipment category is the highest, so the recognition effect is good. Metering technology and power phenomenon categories have medium recognition effect because there exists many specialized words which are difficult to be recognized and the recall rate is not high.

2) EXPERIMENT 2

The impact of learning rate and attention mechanism on power metering entity recognition.

In order to explore the best performance of this model, we adjust and compare the core parameters of learning rate and attention mechanism to obtain the optimal results.

The data in Table 4 shows that the best performance of this model is achieved with a learning rate of 0.005 and the

TABLE 4. Influe of learning rate and attention mechanism on entity recognition results.

Learning Rate	No Attention	Attention
0.01	85.33	88.17
0.005	86.29	**88.62**
0.001	86.08	88.24

TABLE 5. Performance comparison of each model under joint feature embedding.

Model	Recall Rate R(%)	Precision Rate P(%)	F Vaule (%)
BiLSTM-CRF	83.72	84.36	84.04
BiGRU-CRF	83.68	84.34	84.01
CNN-BiGRU-CRF	85.37	86.02	85.69
CNN-BiGRU(Att)-CRF	87.23	87.54	87.38
Att-CNN-BiGRU-CRF	**88.16**	**89.33**	**88.74**

addition of an attention mechanism. Then, we select the value of learning rate near 0.005 for more detailed experiments, and finally use 0.006 as the optimal experiment parameter as shown in Table 2.

3) EXPERIMENT 3

Recognition effect of different neural network models on power metering entities.

Table 5 shows the recognition results of five models. They are BiLSTM-CRF, BiGRU-CRF, CNN-BiGRU-CRF, CNN-BiGRU(Att)-CRF only combined with global attention mechanism and Att-CNN-BiGRU-CRF proposed in this paper. Each model uses joint feature embedding. Table 6 shows the experimental results of our model for the three input cases of character embedding, word embedding, and joint embedding.

The results of the above experiments are taken as the average of three results.

From Table 5 and 6, it can be seen that the Att-CNN-BiGRU-CRF model proposed in this paper is better than the other five models in terms of three metrics: precision rate, recall rate and F-value on the same input features.

It can also be found from the experimental results:

1) The CNN-BiGRU-CRF model has a higher F-value than the BiGRU -CRF model. It shows that the use of CNN to learn local contextual features of words is helpful for entity recognition.

2) BiGRU is more efficient compared to BiLSTM algorithm.

TABLE 6. Performance comparison of our model under different feature embedding.

Model	Input Feature	Recall Rate R(%)	Precision Rate P(%)	F Value(%)
Att-CNN-BiGRU-CRF	Character	87.21	88.42	87.81
	Word	88.02	87.95	87.98
	Character+Word	**88.16**	**89.33**	**88.74**

3) Compared with the CNN-BiGRU-CRF model, the F-value of our model is improved by 3.05%. The improvement is mainly in some complicated entity terms, such as "低压电流互感器 (low-voltage current transformer)" which is correctly recognized as "power equipment" by our model. However, in the CNN-BiGRU-CRF model, "低压电流 (low-voltage current)" is recognized as "power phenomenon" and "互感器 (transformer)" as "power equipment". In addition, in the BiLSTM-CRF model, only "低压电流 (low-voltage current)" is identified, and "互感器 (transformer)" is incorrectly identified as a non-entity. This shows that applying local and global attention mechanisms can optimize key features and discard useless features. The precision of entity recognition can be significantly improved.

4) Our model has similar effects when character embedding and word embedding are used as input features, and the F value is significantly improved after the two are jointly embedded, which proves that the joint feature embedding process is effective and necessary.

VI. CONCLUSION

In order to improve the effect of Chinese named entity recognition in power metering domain. In this paper, we propose an Att-CNN-BiGRU-CRF model. To make full use of the semantic information in the Chinese character structure, we use the attention mechanism to dynamically assign weights to character embedding and word embedding to form joint embedding as the input of our model. For improving the computational performance we use GRU to replace the common LSTM. We use local and global multi-headed attention mechanisms to assign different proportions of weights to word-level and sentence-level feature vectors to increase the importance of critical information. In our experiments, we compare our model with other entity recognition models and demonstrate that the attention mechanism has good results in mining the textual contextual semantic connections, which indicates that the model proposed in this paper has better entity recognition performance and has higher application value. In addition, this paper also designs a new entity

classification method in the power metering domain to make the boundary of specialized terms clearer, and constructs a corresponding power metering corpus.

Our model has relatively strong generality. The limitations is caused by corpus construction and subsets of custom dictionaries. That leads to only the power metering domain is identified currently instead of the whole power domain. That is not a defect of our model. The reasons are as follows:

1) Our model is based on BiGRU-CRF which has good generality.

2) The attention mechanism can optimize the accuracy of the model for capturing contextual features which is also highly generalized.

3) The joint feature embedding and the CNN both have generality. As each layer of our model has generality, it can be applied to the whole power domain with good enough generality.

In the future, we plan to apply this model to the complete power domain. In addition, we also plan to make improvements of our model. Such as to combine with dictionary features in power domain, to add lexical analysis, and to expand the power metering corpus etc.

Our work has the following implications:

1) Provides an effective method for entity identification tasks in power metering domain.

2) Our model combines joint feature embedding and attention mechanism to achieve good experiment results. Which proves the advantage and usability of joint feature embedding and attention mechanism.

3) Our model has good generality and can be extended to do entity recognition tasks in other domain such as medical domain.

Our work also has many practical applications in smart power grid, such as:

1) A Question Answering intelligence voice system of power metering domain based on knowledge graph can use our model to provide AI customer service. The more better NER results, the more higher quality of the AI system.

2) When constructing a vertical semantic search system in power metering domain, our model can be used to extract semantic knowledge and do semantic retrieval based on knowledge graph. It will achieve great search results based one the advantage of our model.

REFERENCES

[1] T. J. Pu, Y. P. Tan, and G. Z. Peng, "Construction and application of knowledge map in power field," *Power Grid Technol.*, vol. 45, no. 6, pp. 2080–2091, 2021.

[2] Z. Q. Liu and H. F. Wang, "Retrieval method for defect records of power equipment based on knowledge graph technology," *Autom. Electr. Power Syst.*, vol. 42, no. 14, pp. 158–164, 2018.

[3] X. H. Zhang, W. Xu, and F. Wu, "Intelligent method for characteristic event tracing and prediction of cascading failures in AC/DC hybrid power grid," *Automat. Electr. Power Syst.*, vol. 45, no. 10, pp. 17–24, 2021.

[4] X. Shan, X. Lu, and M. Y. Zhai, "Analysis of key technologies for artificial intelligence applied to power grid dispatch and control," *Autom. Electr. Power Syst.*, vol. 43, no. 1, pp. 49–57, 2019.

[5] F. Wu, J. Liu, C. Wu, Y. Huang, and X. Xie, "Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation," in *Proc. World Wide Web Conf.*, 2019, pp. 3342–3348.

[6] S. Morwal, N. Jahan, and D. Chopra, "Named entity recognition using hidden Markov model (HMM)," *Int. J. Natural Lang. Comput.*, vol. 1, no. 4, pp. 15–23, 2012.

[7] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[8] C. Sutton, K. Rohanimanesh, and A. McCallum, "Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data," in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, 2004, pp. 1–31.

[9] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 260–270.

[10] J. P. C. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 357–370, Dec. 2016.

[11] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 1064–1074.

[12] A Vaswani, N Shazeer, and N. Parmar, "Attention is all you need," presented at the NIPS Annu. Meeting, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[13] M. Rei, G. K. O. Crichton, and S. Pyysalo, "Attending to characters in neural sequence labeling models," 2016, *arXiv:1611.04361*.

[14] A. Zukov-Gregoric, Y. Bachrach, P. Minkovsky, S. Coope, and B. Maksak, "Neural named entity recognition using a self-attention mechanism," in *Proc. IEEE 29th Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2017, pp. 652–656.

[15] J. Z. Yin, S. L. Luo, Z. T. Wu, and L. M. Pan, "Chinese named entity recognition with character-level BiLSTM and soft attention model," *J. Beijing Inst. Technol.*, vol. 29, no. 1, pp. 60–71, 2020.

[16] B. Wang, Y. Chai, and S. Xing, "Attention-based recurrent neural model for named entity recognition in Chinese social media," in *Proc. 2nd Int. Conf. Algorithms, Comput. Artif. Intell.*, Dec. 2019, pp. 291–296.

[17] K. Xu, Z. Yang, P. Kang, Q. Wang, and W. Liu, "Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition," *Comput. Biol. Med.*, vol. 108, pp. 122–132, May 2019.

[18] G. Wu, G. Tang, Z. Wang, Z. Zhang, and Z. Wang, "An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition," *IEEE Access*, vol. 7, pp. 113942–113949, 2019.

[19] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, and J. Wang, "An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition," *Bioinformatics*, vol. 34, no. 8, pp. 1381–1388, 2017.

[20] Y. D. Shan, H. J. Wang, and H. Huang, "Study on named entity recognition model based on attention mechanism," *Comput. Sci.*, vol. 46, no. S1, pp. 111–114, 2019.

[21] C. Xu, F. Wang, J. Han, and C. Li, "Exploiting multiple embeddings for Chinese named entity recognition," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, 2019, pp. 2269–2272.

[22] C. Dong, J. Zhang, C. Zong, M. Hattori, and H. Di, "Character-based LSTM-CRF with radical-level features for Chinese named entity recognition," in *Proc. Int. Conf. Comput. Process. Oriental Lang.*, 2016, pp. 239–250.

[23] J. Yu, X. Jian, H. Xin, and Y. Song, "Joint embeddings of Chinese words, characters, and fine-grained subcharacter components," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 286–291.

[24] R. Y. Yang, Q. He, and N. S. Du, "Chinese named entity recognition based on gated multi-feature extractors," *Comput. Eng. Appl.*, early access. [Online]. Available: https://kns.cnki.net/kcms/detail/11.2127.TP.20210203.1003.008.html

[25] F. Bin, Z. Youwen, T. Xin, G. Chuangxin, W. Jianjun, Y. Qiang, and W. Huifang, "Power equipment defect record text mining based on BiLSTM-attention neural network," *Proc. CSEE*, vol. 40, pp. 1–10, Aug. 2020.

[26] Y. Tian and W. Ma, "Attention-BiLSTM-based fault text classification for power grid equipment," *Comput. Appl.*, vol. 40, no. S2, pp. 24–29, 2020.

[27] H. Fan, H. C. Huang, and X. Wang, "Research on knowledge extraction technology of grid text data based on semantic annotation," in *Proc. 3rd Smart Grid Conf.*, 2018, pp. 140–144.

[28] Z. Zhao, Z. Chen, J. Liu, Y. Huang, X. Gao, F. Di, L. Li, and X. Ji, "Chinese named entity recognition in power domain based on Bi-LSTM-CRF," in *Proc. 2nd Int. Conf. Artif. Intell. Pattern Recognit.*, 2019, pp. 176–180.
[29] C. Jiang, Y. Wang, J. H. Hu, J. Q. Xu, M. Chen, Y. W. Wang, and G. M. Ma, "Power entity information recognition based on deep learning," *Power Grid Technol.*, vol. 45, no. 6, pp. 2141–2149, 2021.
[30] J. Yang, Y. Zhang, L. Li, and X. Li, "YEDDA: A lightweight collaborative text span annotation tool," in *Proc. ACL*, 2018, pp. 31–36.
[31] J. Devlin, M.-W. Chuang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805.*
[32] Y. Jia and X. Ma, "Attention in character-based BiLSTM-CRF for Chinese named entity recognition," in *Proc. 4th Int. Conf. Math. Artif. Intell. (ICMAI)*, 2019, pp. 1–4.
[33] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," 2012, *arXiv:1212.5701.*
[34] R. Caruana, S. Lawrence, and C. L. Giles, "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping," in *Proc. 13th Int. Conf. Neural Inf. Process. Syst.*, 2000, pp. 402–408.
[35] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.
[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980.*

KAIHONG ZHENG received the master's degree from Zhejiang University, China, in 2016. He is currently an Engineer with the Digital Grid Research Institute, China Southern Power Grid. His research interests include electric energy measurement, electrical energy measurement automation systems, electricity information acquisition systems, and electricity technology.

LINGYUN SUN is currently a Professor with Zhejiang University. He is also the Deputy Director of the National Institute of Design, Zhejiang University, where he is the Deputy Director of the Modern Industrial Design Institute. He is also actively engaged in AI design, information and interactive design, and theory and method of creation design.

XIN WANG received the Ph.D. degree from Zhejiang University, in 2009. He is currently an Associate Professor with the College of Computer Science, Zhejiang University. His research interests include smart grid, big data, and artificial intelligence.

SHANGLI ZHOU received the bachelor's degree from Tsinghua University, China, in 1991. He is currently a Professor Level Senior Engineer with the Digital Grid Research Institute, China Southern Power Grid. His research interests include electric energy measurement, electrical energy measurement automation systems, electricity information acquisition systems, and electricity technology.

HANBIN LI received the master's degree from the Zhejiang University of Technology, China, in 2021. His research interests include knowledge graph, natural language processing, and electricity technology.

SHENG LI received the master's degree from Hunan University, China, in 2019. He is currently an Engineer with the Digital Grid Research Institute, China Southern Power Grid. His research interests include electric energy measurement, electrical energy measurement automation systems, electricity information acquisition systems, and electricity technology.

LUKUN ZENG received the B.S. degree in communication engineering and the Ph.D. degree in electrical engineering from North China Electric Power University, Beijing, China, in 2013 and 2019, respectively. He is currently working with the Digital Grid Research Institute, China Southern Power Grid. His research interests include electrical energy measurement automation systems and data mining technology.

QIHANG GONG received the master's degree from Zhejiang University, China, in 2020. He is currently an Engineer with the Digital Grid Research Institute, China Southern Power Grid. His research interests include electric energy measurement, electrical energy measurement automation systems, electricity information acquisition systems, and electricity technology.