IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 14, 2022, accepted May 3, 2022, date of publication May 5, 2022, date of current version May 17, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3172970

A Joint Model for Hierarchical Nested Information Extraction

RUYANG YIN[1], ZHENCHENG ZHOU[2], AND ZONGHE GAO[2]
[1]Department of Civil Engineering, Institute of Transport Studies, Monash University, Clayton, VIC 3800, Australia
[2]State Grid Electric Power Research Institute Company Ltd., Nanjing 210003, China
Corresponding author: Zonghe Gao (gaozonghe0730@163.com)

ABSTRACT During the long-term power construction process, the power dispatching department has saved many notification texts related to adjustment of grid operation mode. There is an urgent need to study named entity recognition techniques to automatically recognize the power equipment and operation mode, in order to support automatic verification of grid operation mode. By analyzing the characteristics of notification texts, a classification method of hierarchical nested named entities is proposed for the first time in power domain. The entities are divided into two layers with nested relationships, and the corpus of grid operation mode is constructed. We further propose a joint model based on character-word feature fusion and attention mechanism. The model is based on the parameter sharing approach for joint recognition of hierarchical nested entities in the corpus and further introduces an attention mechanism to optimize the feature interaction between hierarchical nested entities. In addition, we splice embeddings of characters and words as feature input to obtain richer semantic features. Experimental results show that our model achieves state-of-the-art results. Eventually, the recognition results can be stored as a standardized verification information chain, providing effective data support for automatic verification of the grid operation mode and ensuring safe and stable operation of the grid.

INDEX TERMS Hierarchical nested named entity, joint model, attention mechanism, feature fusion, named entity recognition.

I. INTRODUCTION

The mode of power grid operation is determined by the actual situation of the power system in order to ensure the safe, high-quality and economic operation of the system. As the topology of power grid becomes increasingly complex, we must first ensure that the power grid is operating in the correct manner in order to maintain the safety and stability of the grid. When important power equipment is in three operation modes of put into production, out of service, and maintenance, the power dispatching agency will adjust the operation mode of the power grid and issue the corresponding notification texts. The notification text contains instructions for adjusting the operation mode of a large number of power equipment. At present, the management of grid operation mode mainly relies on manual verification. Experienced dispatchers understand and memorize the notification texts to ensure that the grid operation mode meets the corresponding requirements in notification texts. However, there are a lot of notification texts accumulated during the long-term construction of the power grid. When dispatchers are not able to obtain texts in time or face heavy maintenance tasks, omissions will inevitably occur, which will lead to unreasonable arrangements of grid operation mode and cause hidden dangers to the safe and stable operation of the power system. Therefore, it is an urgent problem in the process of power grid regulation and operation to automatically obtain the standard names of power equipment and its operation mode requirements in the notification texts through information extraction technologies, so as to support the automatic verification of the grid operation mode and assist the dispatcher in timely detection of operation modes that are inconsistent with the specified requirements and deal with them in a timely manner.

Named Entity Recognition (NER) is a key technology for information extraction, which aims to recognize predefined categories of entities from unstructured text.

Different from the text in general domain, there is a large number of specialist terminology in the notification texts of gird operation mode and the classification of entities is vague. For example, in some of the expressions of the

The associate editor coordinating the review of this manuscript and approving it for publication was Payman Dehghanian.

switch equipment, there is the expressions of bus-bar equipment, which will lead to the switch entity being incorrectly recognized as the bus-bar entity. Similar situations exist in abundance in the notification texts, and a reasonable entity classification method needs to be established for the characteristics of notification texts. Secondly, there is a hierarchical nested structure of entities in the notification texts. For example, the three fine-grained entities of plant-station, set-number, and voltage are nested in the expression of the generator equipment entity. The complex hierarchical nested entities in the notification texts related to adjustment of grid operation mode brings a challenge for NER.

Numerous approaches for nested NER have been proposed in recent years. One representative category is based on the sequence labeling method. Some works revised sequence models to support nested entities using different strategies [1]–[4] and some works adopt the hyper-graph to capture all possible entity mentions in a sentence [5], [6]. Ju *et al.* [1] introduce a LSTM-CRF-based model which dynamically stack flat NER layers to predict entities from inner to outer iteratively based on a deep neural network. However, the method suffers from error propagation, which result in recognition inaccuracy. Straková *et al.* [2] modeled the labels at different layers uniformly as joint labels and propose two models for nested NER task, where one is a multi-label classification model based on LSTM-CRF, another is a seq2seq model. The method exponentially increases the number of labels and also results in data sparsity, which may have a detrimental effect on performance. Besides, the method based on hypergraph needs a lot human efforts for designing unambiguous hypergraph. Another appealing category is based on span classification models that classify candidate spans based on span representations [7]–[10]. Tan *et al.* [7] introduce a mutitask framework for extract nested entities, but they train two subtasks of span boundaries detection and span type classification jointly instead of different application tasks. Shen *et al.* [8] recently implement a two-stage identifier to generate span proposals through a filter and a regressor, and then classify them into the corresponding categories. Although the span-based methods have the innate ability to handle nested NER, they suffer from high computational cost, ignorance of boundary information, under-utilization of the spans that partially match with entities, and difficulties in long entity recognition.

Considering that characters in the sentence belong to different entity types of different layers, we argue that the correlation intensities of the nested entities at different layers are different. For example, the entity of generator has higher correlation intensities wtih the three fine-grained entities of plant-station, set-number, and voltage than other fine-grained entities. Although there is a high dependency relationship between entities at different layers in the notification texts, there is a lack of systematic research on how to use the semantic correlation between entities at different layers to improve the performance of hierarchical nested named entity recognition.

In addition, the notification texts related to adjustment of power grid operation mode belong to Chinese text, which is used to express the meaning through the combination of characters and words. Existing Chinese NER research in the electric power domain can be divided into character-based and word-based approaches, but both of them have limitations. If only character feature is used, the NER model will lose semantic information of specialized words. However, if only the word feature is considered as feature input, the model performance will be limited by the accuracy of Chinese word segmentation [11]. General word separation tools are not applicable to the electrical power terminology within the notification texts. For example, "金家坝变" will be split into "金家" and "坝变", which will cause the NER results to be affected by the splitting error.

Finally, there are multiple different expressions for the same power equipment in the notification texts. For example, "Wujiang/220kV.I section bus-bar", "220kv Wujiang i bus-bar" and "220kV Wujiang (section I)" are different expressions for the same bus-bar equipment. After the power equipment has been correctly recognized, it needs to be further aligned to the unique equipment standard name in order to effectively support the automatic verification of the grid operation mode.

In order to solve the above problems, we first systematically propose a hierarchical nested structure and classification method of entities, and annotate a standard corpus. Subsequently, we propose a novel NER model. Finally, we store the entity recognition results as the standardized verification information chain. Our work can achieve efficient information extraction of the notification texts and the contributions are listed as follows.

1) Based on the characteristics of the notification text, a hierarchical nested structure and related classification method of entities is proposed for the first time in power domain. The hierarchical nested named entities are divided into two layers, where the layer-I entities include 5 types of power equipment entities and 2 types of operation mode attribute entities, and the layer-II entities include 8 types of fine-grained entities nested inside the power equipment. By manually annotating the entities in the notification texts, the corpus of grid operation mode is constructed.

2) Based on the hierarchical nested structure of entities in the corpus, a joint model based on character-word feature fusion and attention mechanism is proposed, which mainly contains a layer-I entity recognition module and a layer-II entity recognition module. Concretely, we introduce the parameter sharing approach for capturing the dependencies between the layer-I entity recognition task and the layer-II entity recognition task. In order to further extract the correlation between hierarchical entities at different layers, we introduce a novel attention mechanism. In addition, we splice embeddings of characters and words as feature input to obtain richer semantic features.

The experimental results show that our proposed NER model has achieved better results for the corpus of grid operation mode than the existing state-of-the-art methods. Eventually, the recognition results are stored as standardized verification information chains to provide a fast, accurate and reliable decision-making information basis for automatic verification of the grid operation mode.

II. RALATED WORK

The algorithmic models related to entity recognition have been developing rapidly and have gone through three stages: approaches based on dictionaries and rules, approaches based on statistical model, and approaches based on deep learning.

Early NER research mainly used manual construction of dictionaries and rules, which is not able to recognize entities outside the dictionary. Since then, a number of approaches based on statistical models have emerged, including Hidden Markov Model (HMM) [12] and Conditional Random Field (CRF) [13]. Such models rely heavily on manually extracted features for specific training data, which is costly in terms of labor and difficult to build models that can be adapted to other domains. With the accumulation of big data and the improvement of computing performance, researchers have further proposed to apply deep learning to named entity recognition. Classical neural networks, including Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), have strong learning ability and feature extraction capability to automatically extract key features of text without complex feature engineering, and have achieved remarkable recognize results in a series of NER tasks. Collobert *et al.* [14] used CNN as the feature extractor to model text sequences, and finally used CRF model to predict the labels of the sequences. Wu *et al.* [15] demonstrated the superiority of applying RNN to NLP tasks by using deep learning models CNN and RNN for entity recognition of text sequences and comparing the model results. Hammerton *et al.* [16] further improved entity recognition by using RNN and Long Short-Term Memory (LSTM) network to label text sequence, demonstrating that LSTM can establish long-term dependency in text sequences, which is more beneficial than RNN for text feature extraction. Since then, more and more NER research based on bi-directional LSTM have been proposed. Since Bi-LSTM is able to simultaneously extract features from contextual information in text sequences, the recognition performance of the model is further improved on the basis of the LSTM network [17]–[19]. Huang *et al.* [20] first proposed to use a combination of Bi-LSTM and CRF for information extraction, where Bi-LSTM simultaneously encodes text using contextual information and then models the transfer relationships between labels by CRF. Compared with degenerate models such as LSTM, Bi-LSTM, and LSTM-CRF, this model achieves better results in NER tasks and is widely used as the cornerstone of many subsequent improved NER models [21]–[24].

Attention mechanism has been widely used in natural language information extraction tasks in recent years. Liu *et al.* [25] alleviated the label inconsistency problem by adding an attention network based on the traditional Bi-LSTM-CRF model framework to effectively exploit the document-level information. Yang *et al.* [26] proposed a target attention network that can make full use of the target word information to effectively enhance the target word prediction in Neural Machine Translation (NMT). Ali *et al.* [27] constructed an efficient multilayer attention model by introducing attention mechanism in both the word embedding layer and the encoding layer, which effectively improved the accuracy of entity recognition results.

The above-mentioned end-to-end NER models built based on deep learning have the characteristics of convenient application and robustness and have now become a general solution for information extraction in various fields [28]. However, these general information extraction models can only recognize flat entities at the same layer in the text, which is not able to extract the overlapping entities with nested structures.

Various approaches for nested NER have been proposed in recent years, which mainly can be divided into two categories: sequence labeling based models and span-based models. For sequence labeling based models, the motivation is trying to transform the nested NER problem into a standard sequence labeling task inspired by the great success of sequence labeling in flat NER. Alex *et al.* [3] dynamically stacks flat NER layers to identify entities from inner to outer. Wang *et al.* [4] designs a pyramid structured tagging framework that uses CNN networks to identify entities from the bottom up. Lu *et al.* [5] is the first to propose the use of Mention Hypergraphs to solve the overlapping mentions recognition problem. Katiyar *et al.* [6] proposed hypergraph representation for the nested NER task and learned the hypergraph structure in a greedy way by LSTM networks. Instead of tagging each token by sequence labeling, span-based models classify spans based on span representation. Sohrab *et al.* [9] proposed an exhaustive model to exhaust all possible spans in a text sequence and then predicts their classes. Xu *et al.* [10] applies a supervised multi-head self-attention mechanism to construct the word level correlations for each type and fuse entity boundary detection and entity classification by a multitask learning framework to capture the dependencies between these two tasks.

Besides, with the successful application of pre-trained language models (PLM) to various tasks, many researchers incorporate pre-trained contextual embeddings into their NER models. PLM can achieve embedded representation of characters and words to further improve the performance of the NER model. In the field of pre-training, scholars have successively proposed Neural Network Language Model (NNLM) [29], Word2Vec [30], Elmo [31], Bidirectional Encoder Representation from Transformers (BERT) [32], and other pre-trained language models. Among them, BERT combines multiple layers of transformers [33] in series and dynamically generates token embeddings with global semantic information according to the context, which can effectively improve the result of downstream natural language processing tasks.

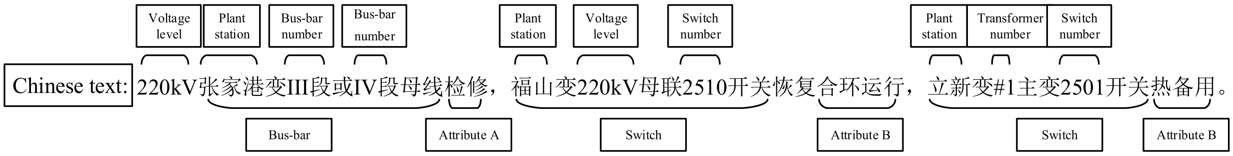


FIGURE 1. Example of notification text and entities.

Different from the aforementioned works, we employ a multi-task learning framework to joint Layer-I and layer-II NER. Simultaneously, we introduce an attention mechanism to consider the correlations between entities at different layers, which can further improve the performance. In addition, scholars have proposed to achieve the fusion of character-level and word-level feature by directly concatenating them, which can extract richer semantic information and obtain better information extraction results [11]. In our model, we use BERT to learn more expressive embeddings of words and characters and fuse them as feature input.

III. CONSTRUCTING CORPUS

We collected the notification texts issued by a provincial dispatching office of the State Grid Corporation in the past ten years, and further cleaned them to eliminate irrelevant information. Finally, the sentence structure of the corpus was divided by punctuation marks, and a complete corpus of 3620 texts related to the adjustment of grid operation mode was collated. Based on the characteristics of the notification texts, we proposed a hierarchical nested structure and a classification method of entities in texts. After the grid dispatching experts completed manual annotation, we constructed a standard corpus of grid operation mode.

A. A HIERARCHICAL NESTED STRUCTURE AND A CLASSIFICATION METHOD OF ENTITIES

Most of the existing work of information extraction in the electric power domain is to extract information at the same layer. For example, some scholars recognize information such as number, phenomenon, degree, cause, and treatment method in defect texts of power equipment, and all these named entities to be recognized are at the same layer. However, the notification texts contain a large number of important power equipment and operation mode requirements, in which the power equipment entities have a complex hierarchical nested structure. Figure 1 shows an example of the notification text and the entities in it. The corresponding English translation of the sample text in the figure is: When 220kV Zhangjiagang substation section III or IV bus-bar is in maintenance, Fushan substation 220kV bus tie 2510 switch resume closed-loop operation and Lixin substation #1 transformer 2501 switch is tuned to hot standby.

The primary entity to be recognized in the notification text is the power equipment entity, and it is easy to find that there are two classes of power equipment entities in the sample text, which are the bus-bar equipment and the switch equipment. However, there are multiple different expressions for the same equipment in the corpus. In order to provide effective data support for the automatic verification of the grid operation mode, the power equipment entity in the corpus needs to be further aligned to the unique standard name of the equipment in the energy management system after it has been recognized. It is not difficult to find that power equipment entities are composite entities. For example, the expression of the bus-bar entity in the sample text encapsulates fine-grained information such as voltage level, plant station, and the equipment number, which can align the recognized power equipment entity to the unique standard name. Therefore, based on the characteristics of the texts, we model two layers of entities, which are power equipment entities and fine-grained entities nested inside the equipment entities. There are five types of power equipment entities in the texts, which are labeled as bus-bar, set, switch, transformer and line. The fine-grained entities include 8 types, which are labeled as plant station, voltage level, line abbreviation, bus-bar number, switch number, line number, transformer number, and set number.In addition, in the notification texts, each power equipment entity is associated with a corresponding operation mode, which we model as the operation mode attribute entity. There are two classes of operation mode attribute entities, which are labeled as attribute A and attribute B respectively. The former indicates the operating mode of the equipment that causes the adjustment of grid operation mode, while the latter indicates the correct operation mode of the relevant equipment after the adjustment. In the notification text, each power equipment entity is directly associated with a subsequent operation mode attribute entity, and the power equipment associated with attribute A triggers the adjustment of operation mode of the power equipment associated with attribute B. In the sample text, for example, when the bus-bar equipment is in operation mode of maintenance, this will trigger the operation mode of the relevant switch to be adjusted to closed-loop operation and hot standby, respectively.

In summary, there exists a hierarchical nested structure of entities in the notification texts, including five categories of power equipment entities, eight categories of fine-grained entities nested in equipment, and two categories of operation mode attribute entities. The power equipment entities and operation mode attribute entities are defined as the layer-I entities, and the fine-grained entities are defined as the layer-II entities. The hierarchical nested structure and classification method of entities in the notification texts are shown in Figure 2, where the connection lines indicate the

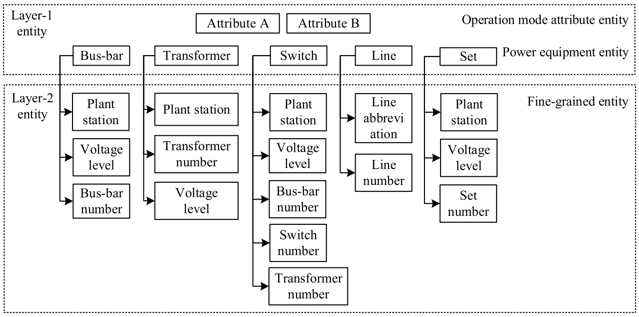


FIGURE 2. Hierarchical nested named entities in notification texts.

existence of corresponding fine-grained entities within the power equipment entities.

B. LABELING SCHEME AND STATISTICS OF ENTITIES

We use the above classification method to label the collected notification texts and classify the entities into 15 well-defined label classes. The BIO tagging strategy is used in this paper, where the beginning characters of entity sequence are tagged as B tags, the remaining characters of entity sequence are tagged as I tags, the characters that do not need to be recognized are uniformly tagged as O tags, and both the B tags and I tags are to be connected to specific entity categories. Named entities in the texts are manually tagged by grid scheduling experts through the YEDDA [34] tool, which segments and tag each character in the texts. The traditional BIO tagging method can only tag the named entities with flat structure and cannot deal with the entities in the corpus with a hierarchical nested structure due to the fact that each character corresponds to multiple entity tags at different layers. Therefore, we use a hierarchical labeling strategy to label complex hierarchical nested entities in the corpus, i.e., each text in the corpus is labeled twice to obtain the labels of the entities at all layers. In the first labeling, the operation mode attribute entities and power equipment entities are tagged, and in the second labeling, the fine-grained entities are tagged. After the labeling is completed, each notification text corresponds to two layers of labels.

The distribution of hierarchical nested entities in the corpus are summarized in Table 1, divided into the categories of entities at layer-I and layer-II. There are 3620 notification texts in the corpus, including 57832 entities, of which the numbers of lay-I entities and layer-II entities are 21,652 and 36,180, respectively. Meanwhile, there are 5.98 lay-I entities and 9.99 layer-II entities in each sentence on average, and the average length of entities is 4.54. In addition, there are a total of 46,988 entities with hierarchical nested structure, accounting for 81.25% of the total number of entities.

IV. JOINT MODEL BASED ON FEATURE FUSION AND ATTENTION MECHANISM

We propose a joint model based on feature fusion and attention mechanism, and the structure diagram of our model is shown in Figure 3. The model consists of three modules, namely the encoder module, the layer-I entity recognition module and the layer-II entity recognition module.

TABLE 1. Distribution of entities in the corpus.

Layer-I entities	Number	Layer-II entities	Number
Bus-bar	1582	Voltage Level	4273
Set	1149	Plant Station	9285
Switch	1428	Line Abbreviation	5062
Transformer	3625	Line Number	5046
Line	4016	Set Number	3573
Attribute A	2062	Transformer Number	3303
Attribute B	7790	Bus-bar Number	3699
/	/	Switch Number	1939

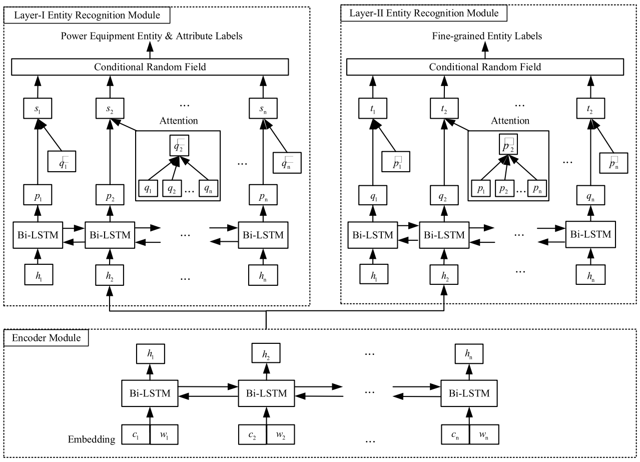
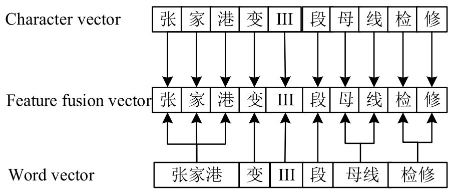


FIGURE 3. Structure diagram of entity recognition model.

Taking the feature fusion of characters and words as input, the encoder module utilizes a Bi-LSTM neural network to extract bidirectional sequence features and construct shared context representation. Afterward, the entity recognition modules of layer-I and layer-II take the shared context representation from the encoder module as input. A hierarchical nested entity correlated attention unit is further used to calculate the nested correlations between entities at different layers in each entity recognition module, which, along with the BiLSTM-CRF, predicts final tags related to hierarchical nested entities. In particular, the layer-I entity recognition module is used to recognize the power equipment entities and the corresponding operation mode attribute entities, while the layer-II entity recognition module is used to recognize the fine-grained entities nested inside the power equipment entities. In the following sections, we will describe each module in detail.

A. ENCODER MODULE

The first network layer of the encoder module is the embedding representation layer, which is used to convert each


Character vector
张 家 港 变 III 段 母 线 检 修
Feature fusion vector
张 家 港 变 III 段 母 线 检 修
Word vector
张家港 变 III 段 母线 检修

(Translation: Zhangjiagang substation section III bus-bar is in maintenance)

FIGURE 4. Feature fusion of character feature and word feature.

character in the corpus into a low-dimensional vector. In the pre-trained language models, BERT uses a bidirectional Transformer encoder that can dynamically generate character feature vectors and word feature vectors with global semantic information. Compared with the traditional pre-trained language models, BERT has stronger feature extraction ability, and the obtained feature embedding representations of characters and words have richer semantic information. Therefore, we choose BERT for text-to-vector conversion.

In the process of power system regulation and control operation, the power dispatching organization has accumulated various working documents, including notification texts related to the adjustment of grid operation mode, operation tickets, instruction tickets, dispatching logs, dispatching regulations, application forms of maintenance, etc. In this paper, we use the above corpus for self-supervised learning of the BERT model, and the required feature embedding representations of characters and words can be obtained by pre-trained BERT.

In order to enable the model to fully take into account the semantic information of characters and words, we construct the structure shown in Figure 4 to achieve feature fusion by directly concatenating the feature vectors of characters and words to which they belong.

We collect the professional words contained in the notification texts during the annotation process, and add them and the professional words in the standard names of power equipment together into the general word segmentation dictionary. Based on the word segmentation dictionary combined with professional words, we use the Jieba word segmentation tool to segment the corpus of grid operation mode. Each character feature vector corresponds to a word feature vector, for example, the character "张" corresponds to the word "张家港", and the feature vectors of the two are concatenated to obtain the feature fusion vector of the character "Zhang". In particular, there are words that contain only one character, such as the word "变", of which the feature fusion is performed in the same way. We use $X = [x_1, x_2, \ldots, x_n]$ to denote that each text sequence contains n characters, and $e_i = [c_i; w_i]$ to denote the feature fusion vector of the $i-th$ character in X, where c_i is the character vector and w_i is the word vector to which the character belongs. The feature fusion vector takes into account the features of characters and words, which can reduce the influence of word boundary delimitation errors and obtain relatively rich semantic features of specialized words.

In order to obtain contextual features shared by recognition modules of entities at different layers, the feature fusion vectors obtained from the embedding layer are further input into the Bi-LSTM network for training. The LSTM network introduces the input gate, forget gate, and output gate to control the information transmission process of text sequences based on RNN, which alleviates the gradient dispersion and gradient explosion problems easily generated by traditional RNN, and thus realizes the effective feature extraction of a long text sequence. The LSTM network state at moment t can be calculated as follows,

$$g_i^{(t)} = \text{sigmoid}\left(w_i h_{t-1} + u_i x_t + b_i\right) \tag{1}$$

$$g_f^{(t)} = \text{sigmoid}\left(w_f h_{t-1} + u_f x_t + b_f\right) \tag{2}$$

$$g_o^{(t)} = \text{sigmoid}\left(w_o h_{t-1} + u_o x_t + b_o\right) \tag{3}$$

$$C_t = g_{\mathrm{f}}^{(t)} C_{t-1} + g_{\mathrm{i}}^{(t)} \tanh\left(w h_{t-1} + u x_t + b\right) \tag{4}$$

$$h_t = g_{\mathrm{o}}^{(t)} \tanh\left(C_t\right) \tag{5}$$

where *sigmoid*$(\cdot)$ and *tanh*$(\cdot)$ are both activation functions, g denotes the gating unit of LSTM, and the subscripts i, f, and o denote the input gate, forget gate, and output gate, respectively. Eqs. (1)–(3) describe the calculation process of the feed-forward neural network of the above three gating units, w and u denote the weight matrix of the hidden layer vector h and the input vector x in turn, b denotes the bias vector, C_t denotes the textual information of the LSTM unit retained from the start moment to the current moment t, and h_t denotes the hidden state of the LSTM network at moment t.

The Bi-LSTM network consists of two LSTM networks with opposite directions, which are computed using the forward and backward order of text sequences, respectively, and thus can fully extract the contextual features in the text. Current research shows that Bi-LSTM networks have become a standard network layer for solving NER tasks. We use the Bi-LSTM to produce the forward state $\overrightarrow{h_i}$ and backward state $\overleftarrow{h_i}$, and concatenate them as the shared feature encoding of $i-th$ character in the text sequence, denoted as $h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}] \in \mathbf{R}^{2 \times d_h}$, where d_h indicates the dimension of $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$.

B. LAYER-I ENTITY RECOGNITION MODULE

Taking the shared features from the encoder module as input, another Bi-LSTM runs to further extract the internal features from $H = [h_1, h_2, \ldots, h_{\mathrm{n}}]$. The output representation sequence from the Bi-LSTM in the layer-I entity recognition module is denoted as $P = [p_1, p_2, \ldots, p_{\mathrm{n}}]$. Similarly, we use Bi-LSTM for internal feature extraction in the layer-II entity recognition module and the output sequence is denoted as $Q = [q_1, q_2, \ldots, q_{\mathrm{n}}]$.

Since there are nested correlations between entities at different layers in the corpus, we calculate the correlation coefficients between hierarchical nested entities through an attention network to extract the important features in the layer-II entity recognition module that contribute to entity recognition in this layer. In the layer-I entity recognition module, for the vector representation p_i of the $i-th$ character in the output sequence P, we obtain the attention scores between entities at different layers by calculating the correlation between this vector and all feature vectors in the sequence Q, and from this we obtain the attention weight matrix. A higher attention score represents a stronger correlation between entities at different layers and a corresponding higher attention weight. The computational process is listed as follows,

$$\operatorname{score}\left(p_i, q_j\right) = U^T \tanh\left(W_1 p_i + W_2 q_j\right) \tag{6}$$

$$A_{i,j} = \frac{\exp\left(\operatorname{score}\left(p_i, q_j\right)\right)}{\sum_{k=1}^{n} \exp\left(\operatorname{score}\left(p_i, q_k\right)\right)} \tag{7}$$

where W_1, W_2, U^T are trainable weight matrix, and $A_{i,j}$ is the attention weight of the feature vectors p_i and q_j. Through the attention weight matrix A, we can obtain the nested correlation features of each character based on the contextual feature sequence in the layer-II recognition module, which is denoted as $\widehat{Q} = [\widehat{q}_1, \widehat{q}_2, \ldots, \widehat{q}_n]$. Then, the output representation sequence of the attention layer is $S = [s_1, s_2, \ldots, s_n]$, where the computational process of s_i is listed as follows,

$$\widehat{q}_i = \sum_{j=1}^{n} A_{ij} q_j \tag{8}$$

$$s_i = \tanh\left(W_3\left[\widehat{q}_i; p_i\right]\right) \tag{9}$$

where W_3 is trainable weight matrix.

Bi-LSTM network and attention network can extract the contextual features of text and nested correlation features between hierarchical nested entities, respectively, but cannot model the transfer relationships between entity labels. Moreover, taking the output of attention network as input, a CRF is utilized to model the mappings between tokens and labels in the layer-I entity recognition module. The computational process is listed as follows,

$$\operatorname{Score}\left(X, \hat{Y}_{Layer-I}\right) = \sum_{i=1}^{n} V_{y_i, y_{i+1}} + \sum_{i=1}^{n} S_{i, y_i} \tag{10}$$

where $Score\left(X, \hat{Y}_{Layer-I}\right)$ is the score function, and $\hat{Y}_{Layer-I} = [y_1, y_2, \ldots, y_n]$ is the sequence of predict labels. Moreover, $V \in \mathbf{R}^{d_l \times d_l}$ is trainable transition score matrix, d_l is the dimension of layer-I entity type-based label space, $V_{y_i, y_{i+1}}$ denotes the transition score of labels y_i to its adjacent label y_{i+1}, and S_{i,y_i} is the state score matrix to model the mappings between i-th token and the corresponding label y_i. Then, the probability of $\hat{Y}_{Layer-I}$ is calculated as in Eq. (12), where $\hat{Y}_{Layer-I}$ denotes all true label sequences in the Layer-I Recognition Module of X. Eq. (13) describes the loss function of the Layer-I recognition module. The computational process is listed as follows,

$$P\left(\hat{Y}_{Layer-I} \mid X\right) = \frac{\exp\left(\operatorname{Score}\left(X, \hat{Y}_{Layer-I}\right)\right)}{\sum_{\tilde{Y} \in Y_{Layer-I}} \exp(\operatorname{Score}(X, \tilde{Y}))} \tag{11}$$

$$Loss_{Layer-I} = \log \sum_{\tilde{Y} \in Y_{Layer-I}} \exp(\operatorname{Score}(X, \tilde{Y})) - \operatorname{Score}\left(X, \hat{Y}_{Layer-I}\right) \tag{12}$$

Through this module, we can obtain the global optimal labeling results of the layer-I entities, including the entities of power equipment and the operation mode attribute.

C. LAYER-II ENTITY RECOGNITION MODULE

We use the same structure of the layer-I entity recognition module to extract internal features of the text sequence in this module. Taking the feature vector sequence output from the shared encoder module as input, this module first extracts contextual features through Bi-LSTM network, and then calculates the nested correlations between hierarchical nested entities through attention network. Moreover, we use the output vector sequence of the attention layer as the input of the CRF layer, and obtain the global optimal labeling result of the layer-II entities, i.e., the fine-grained entities.

D. JOINT LEARNING

As CRF is used as the entity decoder in both layer-I and layer-II recognition modules, we use Eq. (10) as the overall loss function of the proposed model:

$$Loss_{\text{ALL}} = Loss_{\text{Layer-I}} + Loss_{\text{Layer-II}} \tag{13}$$

where $Loss_{\text{Layer-I}}$ is the difference between the predicted entity type labels of power equipment entity and attribute entity and true labels of the Layer-I entity recognition module, while $Loss_{\text{Layer-II}}$ is the distance between the predicted entity type labels of fine-grained entity labels and true labels of the Layer-II entity recognition module. Our optimization goal is to make $Loss_{\text{Layer-I}}$ and $Loss_{\text{Layer-II}}$ as small as possible.

V. EXPERIMENTS AND ANALYSIS

A. EXPERIMENTAL PLATFORM AND PARAMETER CONFIGURATION

All experiments were performed on a computing device installed with the Ubuntu 16.04 operating system and Pytorch 1.5. The device was configured with Intel I9-9900KF CPU, NVIDIA 2080ti GPU, 128 GB DDR3 RAM, and a 2 TB disk. The vector dimension of character and word is 768. The length of a single text sequence of the corpus does not exceed 128, so the maximum sequence length of our model is 128. During model training, we use Early-stop [35] and Dropout [36] to avoid the model overfitting problem. Meanwhile, we use Adam [37] to update the model parameters to optimize its performance. Table 2 records the specific training parameters of our model in the experiment.

B. DATESETS

To evaluate the performance of our model, we conducted comprehensive experiments on the corpus of grid operation mode. We split dataset into training, development and testing with the ratio 8:1:1.

TABLE 2. Parameters of our model.

Parameters	Value
dimension of word embedding	768
dimension of character embedding	768
dropout rate	0.3
hidden unit of LSTM network	100
hidden unit of attention network	200
output unit of attention network	100
learning rate	0.0005
batch size	32

C. EVALUATION INDICATORS

The evaluation indicators of NER model are precision rate, recall rate and $F1$ value, which are calculated as follows,

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

$$F1 = \frac{2 \ \times Precision \times Recall}{Precision + Recall} \tag{16}$$

where TP is the number of correct named entities in the recognition results, FP is the number of incorrect named entities in the recognition results, and FN is the number of unrecognized named entities. The precision rate indicates the proportion of correct entities in the recognized entities, while the recall rate indicates the proportion of all entities in the sample to be correctly recognized, and the two are mutually constrained. $F1$ combines the precision rate and recall rate, and can analyze the entity classification performance more comprehensively.

D. BASELINES

In order to verify the superiority of our model for recognition of hierarchical nested entities in the corpus of grid operation mode, we compare it with several state-of-the-art models, including sequence labeling based and span-based models. In order to standardize the comparison and validate the effectiveness of our model, we use character-word fusion embedding and restrict the encoder of all models to be BERT.

Cascaded-Bi-LSTM-CRF applies several stacked LSTM-CRF layers to recognize nested entities at different layers in an inside-out manner [1].

Seq2seq is under a encoder-decoder framework to predict the joint label of entitie one by one [2].

BENSC is a span-based method that incorporates a boundary detection task and type classification task for multitask learning [7].

Locate and Label is a two-stage NER method to generate span proposals through a filter and a regressor, and then classify them into the corresponding entity types [8].

TABLE 3. Recognition result of our model.

Entities	Precision Rate (%)	Recall Rate (%)	F1 Value (%)
Bus-bar	94.15	92.34	93.24
Set	93.59	91.21	92.38
Switch	98.68	96.53	97.59
Transformer	93.38	90.92	92.13
Line	100.00	100.00	100.00
Attribute A	94.13	92.57	93.34
Attribute B	96.69	94.12	95.39
Voltage Level	98.12	98.36	98.24
Plant Station	98.23	96.78	97.50
Line Abbreviation	100.00	100.00	100.00
Line Number	100.00	100.00	100.00
Set Number	95.13	99.21	97.13
Transformer Number	97.16	98.22	97.69
Bus-bar Number	100.00	100.00	100.00
Switch Number	100.00	100.00	100.00

TABLE 4. Performance comparison of various models under the feature fusion embedding of characters and words.

Model	Precision Rate (%)	Recall Rate (%)	F1 Value (%)
Cascaded-Bi-LSTM-CRF (BERT)	94.7	93.0	93.8
Seq2seq (BERT)	94.9	93.5	94.2
BENSC (BERT)	94.7	94.2	94.4
Locate and Label(BERT)	96.5	96.6	96.5
Our model	98.2	96.9	97.5

E. EXPERIMENTAL RESULTS AND ANALYSIS

The recognition results of our model for hierarchical nested named entities are shown in Table 3.

It can be seen in Table 3 that our model performs well on the corpus of grid operation mode, and the $F1$ value of each entity is above 90%.

The lower $F1$ values for the power equipment entities at layer-I are partly due to the long names and various combinations of such entities, thus making them difficult to recognize correctly. In contrast, the $F1$ values of attribute entities at layer-I are higher because such entities generally have more distinctive indicative features. Similarly, the fine-grained entities at layer-II are recognized well as their expressions are uniform, standardized, and short in length.

The recognition results of the baseline models and our model for hierarchical nested named entities in the corpus under feature fusion are shown in Table 4.

According to Table 4, we can see that our model outperforms the best baseline (Locate and Label) by 1.0%

TABLE 5. Ablation study on our corpus.

Model	Precision Rate (%)	Recall Rate (%)	F1 Value (%)
Our model	98.2	96.9	97.5
-BERT	93.3	92.1	92.7(↓4.8)
-character-level embedding	97.3	95.1	96.2(↓1.3)
-word-level embedding	96.1	95.5	95.8(↓1.7)
-attention	95.8	94.0	94.9(↓2.6)

absolute F1 score on our datasets. Another, we can see that the Cascaded-Bi-LSTM-CRF baseline perform with worst F1 score. We think the model suffers from error propagation. It also can be observed that the Seq2seq baseline perform not well. We think the model suffers from exponentially increasing of the number of labels and data sparsity. In addition, we think that although BENSC utilize multi-task model for joint learning, they still suffer from error cascading between tasks because of working in pipeline manner. Based on the above experiments and results, our model achieves the expected performance, which can be attributed to the following factors.

1) We jointly extract hierarchical nested information based on the parameter sharing approach, and the layer-I entity recognition module and the layer-II entity recognition module perform feature extraction and interaction through a shared encoder module. At the same time, the interdependence of the layer-I entity recognition subtask and the layer-II entity recognition subtask is established, thereby improving the recognition performance.

2) By introducing the attention mechanism in the layer-I entity recognition module and the layer-II entity recognition module, it can make full use of the correlation between the hierarchical nested entities and effectively promote the mutual influence and information fusion between the two entity recognition modules.

F. ABLATION STUDY

Tabel 5 presents the results of an ablation experiment on our corpus showing each component of our model have various degrees of contributes to the effectiveness of our model.

We can see that ablation on pre-trained language model BERT significantly decreases the F1 scores by 4.8 percentage points, which indicates that BERT pre-trained with training data in the power domain can significantly improve the performance. At the same time, adding word-level embeddings and character-level embeddings can both improve the performance of our model, which proves that feature fusion embedding is effective and necessary. In addition, we also observe performance drop of 2.6 percentage points when we do not use attention mechanism which validates its effectiveness for learning the nested correlation between hierarchical entities at different layers.

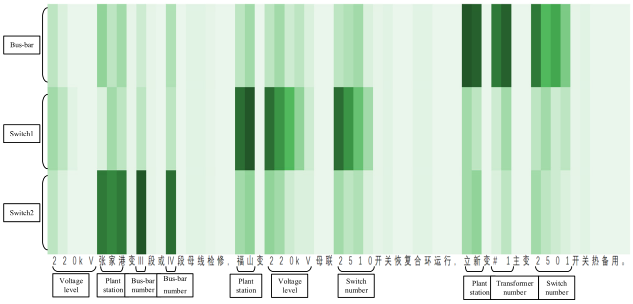


FIGURE 5. Attention of layer-I entities on layer-II entities in the sample notification text given above. The darker the color, the greater the attention weight.

G. VISUALIZATION

Each slice of tensor A is a correlation score matrix A_i, which indicates the correlation intensities between the vector representation P_i of i-th character in layer-I module and the vector representation Q of each character in layer-II module. We sum up over all the correlation score matrix Access-V5.docx corresponding to each character belonging to the same layer-1 entity and then normalizing. So we can get the normalized weight vectors that yields a general view of what characters in layer-II entities the layer-I entity mostly focus on. For interpreting the correlation between entities at different layers in sentence learned by attention mechanism, we plot a heap map of three normalized weight vectors for the sample notification text given above, corresponding to the three layer-I entities contained in the sample, as shown in Figure 5.

From the figure, we can see that the three layer-I entities in the sample notification text can focus on the relevant layer-II entities nested in themselves. We can conclude that the correlations between entities at different layers learned by our proposed attention mechanism are meaningful and in accord with common sense.

VI. STANDARDIZED VERIFICATION INFORMATION CHAIN

After the hierarchical nested entities in the grid operation mode corpus is recognized by our model, the recognized results include power equipment entities, fine-grained entities, and operation mode attribute entities. Since there are many different expressions of the same power equipment in the corpus of grid operation mode. For example, the following three expressions: “Wujiang/220kV.Section I Bus-bar”, “220kv Wujiang Station I bus-bar” and “220kV Wujiang (Section I)” all refer to the same power equipment. Therefore, in order to effectively support the automatic verification of the grid operation mode, it is necessary to align different expressions into the standard names of power equipment, so as to realize the unification of equipment information. The standard names of power equipment combined with the associated attribute entities can be store as the standardized verification information chain. The information chain can provide fast, accurate and reliable decision information basis for the automatic verification of the grid operation mode, thus ensuring the safe and stable operation of the power system.

Taking the sample notification text given above as an example, our model recognizes "Fushan 220kV bus tie 2510 switch" as a switch entity, and also recognizes the fine-grained entities nested within it, including station entity: Fushan, voltage level entity: 220kV and switch number entity: 2510. In this way, we have completely obtained all the information of power equipment: "equipment category: switch, plant station: Fushan, voltage level: 220kV, switch number: 2510". This information achieves a complete representation of professional power equipment based on fine-grained power equipment knowledge. Based on the word matching method, the fine-grained information is matched in the standard name set of the equipment category to which it belongs, i.e., to judge whether the corresponding fine-grained information exists in the standard names. If a standard name in the set can be successfully matched with all the fine-grained information, then we get the standard name of the power equipment. For example, the above power equipment belongs to the switch category, so the fine-grained information of this equipment is matched in the standard name set of switches. Since there is only one standard name "Jiangsu. Fushan substation/220kV. bus tie 2510 switch" in the set that can be successfully matched with all the fine-grained information, the standard name of the switch entity is "Jiangsu. Fushan substation/220kV. bus tie 2510 switch".

In addition, the expression of bus-bar equipment in the sample text "220kV Zhangjiagang substation section III or IV bus-bar" is standardized to "Jiangsu. Zhangjiagang/220kV.220kV section III bus-bar" and "Jiangsu. Zhangjiagang/220kV.220kV section IV bus-bar", and the expression of switch equipment "Lixin substation #1 transformer 2501 switch" is standardized to "Jiangsu. Lixin substation/220kV.#1 transformer 2501 switch".

After converting the power equipment entities and fine-grained entities into standard equipment names through the word matching method, we further combine the operation mode attribute entities associated with the equipment to store the recognition results as the standardized verification information chain of the grid operation mode. Based on the sample notification text given above, we obtain the standardized verification information chain of the operation mode as follows.

Standardized verification information chain 1:

{[Jiangsu. Zhangjiagang/220kV.220kV section III bus-bar] - operation mode attribute - [maintenance]} - trigger - {[Jiangsu. Fushan substation/220kV.bus tie 2510 switch] - operation mode attribute - [closed-loop operation]; [Jiangsu. Lixin substation/220kV.#1 transformer 2501 switch] - operation mode attribute - [hot standby]}

Standardized verification information chain 2:

{[Jiangsu. Zhangjiagang/220kV.220kV section IV bus-bar] - operation mode attribute - [maintenance]} - trigger - {[Jiangsu. Fushan substation/220kV.bus tie 2510 switch] – operation mode attribute - [closed-loop operation]; [Jiangsu. Lixin substation/220kV.#1 transformer 2501 switch] - operation mode attribute - [hot standby]}

The standardized verification information chain contains the standard names of the power equipment and the associated operation mode attributes, where the bus-bar equipment is directly associated with the operation mode attribute "maintenance" and the two-switch equipment are associated with the operation mode attributes "closed-loop operation" and "hot standby". The above information chain states that when the bus-bar equipment is in operation mode of "maintenance", the dispatcher has to adjust the operation mode of the two switches to "closed-loop operation" and "hot standby" respectively. By storing the result of our model as a standardized verification information chain, it can provide fast, accurate and reliable data support for the automatic verification of the grid operation mode, thereby effectively guaranteeing the safe and stable operation of the power system.

VII. CONCLUSION

This paper proposes the hierarchical nested structure and the classification method of entities based on the characteristics of notification texts for the first time in power domain, and independently constructs the corpus of grid operation mode. To effectively extract hierarchical nested information in the corpus, we propose the joint model based on character-word feature fusion and attention mechanism. The model jointly trains the recognition subtasks of hierarchical entities at different layers based on the parameter sharing approach. Different subtasks perform feature extraction and interaction by sharing the encoder module, and further mine the correlation between entities at different layers by introducing the attention mechanism. In addition, in order to fully consider the semantic features of notification texts, we integrate the feature of characters and words as the feature input, which provides richer semantic information for our model.

Compared with several classic NER models, our model achieves state-of-the-art recognition results for the hierarchical nested information in the corpus, which proves the superiority of our model. The model recognition results can be stored as the standardized verification information chain, providing structured data support for automatic verification of the grid operation mode. In the future, we plan to build a knowledge graph based on the standard names of power equipment in the next step and combine it to improve the performance of entity recognition in the corpus of grid operation mode.

REFERENCES

[1] M. Ju, M. Miwa, and S. Ananiadou, "A neural layered model for nested named entity recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2018, pp. 1446–1459.

[2] J. Straková, M. Straka, and J. Hajič, "Neural architectures for nested NER through linearization," 2019, *arXiv:1908.06926*.

[3] B. Alex, B. Haddow, and C. Grover, "Recognising nested named entities in biomedical text," in *Biological, Translational, and Clinical Language Processing*. Prague, Czech Republic: Association for Computational Linguistics, 2007.

[4] J. Wang, L. Shou, K. Chen, and G. Chen, "Pyramid: A layered model for nested named entity recognition," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 5918–5928.

[5] W. Lu and D. Roth, "Joint mention extraction and classification with mention hypergraphs," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 857–867.

[6] A. Katiyar and C. Cardie, "Nested named entity recognition revisited," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2018, pp. 1–11.

[7] C. Tan, W. Qiu, M. Chen, R. Wang, and F. Huang, "Boundary enhanced neural span classification for nested named entity recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 5, pp. 9016–9023.

[8] Y. Shen, X. Ma, Z. Tan, S. Zhang, W. Wang, and W. Lu, "Locate and label: A two-stage identifier for nested named entity recognition," 2021, *arXiv:2105.06804*.

[9] M. G. Sohrab and M. Miwa, "Deep exhaustive model for nested named entity recognition," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2843–2849.

[10] Y. Xu, H. Huang, C. Feng, and Y. Hu, "A supervised multi-head self-attention network for nested named entity recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 16, pp. 1–9.

[11] 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.*, Nov. 2019, pp. 2269–2272.

[12] 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.

[13] 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.

[14] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12 pp. 2493–2537, Aug. 2011.

[15] Y. Wu, M. Jiang, J. Xu, D. Zhi, and H. Xu, "Clinical named entity recognition using deep learning models," in *Proc. AMIA Annu. Symp.* Bethesda, MD, USA: American Medical Informatics Association, 2017, p. 1812.

[16] J. Hammerton, "Named entity recognition with long short-term memory," in *Proc. 7th Conf. Natural Lang. Learn. HLT-NAACL*, 2003, pp. 172–175.

[17] P. Wang, Y. Qian, F. K. Soong, L. He, and H. Zhao, "A unified tagging solution: Bidirectional LSTM recurrent neural network with word embedding," 2015, *arXiv:1511.00215*.

[18] V. Athavale, S. Bharadwaj, M. Pamecha, A. Prabhu, and M. Shrivastava, "Towards deep learning in Hindi NER: An approach to tackle the labelled data scarcity," 2016, *arXiv:1610.09756*.

[19] S. Almgren, S. Pavlov, and O. Mogren, "Named entity recognition in Swedish health records with character-based deep bidirectional LSTMs," in *Proc. BioTxtM*, 2016, p. 30.

[20] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*.

[21] 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*, Jun. 2016, pp. 260–270.

[22] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, vol. 1, Aug. 2016, pp. 1064–1074.

[23] 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.

[24] L. Liu, J. Shang, X. Ren, F. Xu, H. Gui, J. Peng, and J. Han, "Empower sequence labeling with task-aware neural language model," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI)*, Feb. 2018, pp. 5253–5260.

[25] B. Ji, R. Liu, S. Li, J. Yu, Q. Wu, Y. Tan, and J. Wu, "A hybrid approach for named entity recognition in Chinese electronic medical record," *BMC Med. Informat. Decis. Making*, vol. 19, no. S2, pp. 149–158, Apr. 2019.

[26] M. Yang, M. Zhang, K. Chen, R. Wang, and T. Zhao, "Neural machine translation with target-attention model," *IEICE Trans. Inf. Syst.*, vol. E103.D, no. 3, pp. 684–694, 2020.

[27] M. N. A. Ali, G. Tan, and A. Hussain, "Boosting Arabic named-entity recognition with multi-attention layer," *IEEE Access*, vol. 7, pp. 46575–46582, 2019.

[28] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 50–70, Jan. 2022.

[29] Y. Shi, W.-Q. Zhang, M. Cai, and J. Liu, "Efficient one-pass decoding with NNLM for speech recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 4, pp. 377–381, Apr. 2014.

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[31] H. Hassan, G. Sansonetti, F. Gasparetti, A. Micarelli, and J. Beel, "BERT, ELMo, USE and InferSent sentence encoders: The panacea for research-paper recommendation?" in *Proc. 13th ACM Conf. Recommender Syst.*, 2019, p. 1–5.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[34] J. Yang, Y. Zhang, L. Li, and X. Li, "YEDDA: A lightweight collaborative text span annotation tool," in *Proc. ACL*, 2018, pp. 31–36.

[35] 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.

[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*.

RUYANG YIN was born in Jiangsu, China, in 1995. She received the B.S. degree in civil engineering from Central South University, China, in 2017, and the M.S. degree in civil engineering from the University of Illinois Urbana–Champaign, USA, in 2018. She is currently pursuing the Ph.D. degree in civil engineering with the Department of Civil Engineering, Institute of Transport Studies, Monash University, Melbourne, Australia.

ZHENCHENG ZHOU was born in Jiangsu, China, in March 1993. He received the master's degree from the State Grid Electric Power Research Institute, Nanjing, China, in 2022. He is currently an Engineer with the State Grid Electric Power Research Institute. His research interests include smart grid, big data, and artificial intelligence.

ZONGHE GAO was born in Anhui, China, in April 1962. He received the master's degree from the State Grid Electric Power Research Institute, Nanjing, China, in 1989.

Currently, he is the Chief Expert of power system dispatch and control with the State Grid Electric Power Research Institute.