scientific reports

Check for updates

OPEN

A multi-level feature enhancement framework for named entity recognition in power system texts

Ziming Wei[1], Hongchao Gao[2], Shaocheng Qu[1]✉, Li Zhao[1], Qianqian Shi[1] & Chen Zhang[1]

Power equipment maintenance work orders, a critical type of power system texts, are rich in operational details such as faulty components and maintenance procedures. However, automated information extraction from these orders is impeded by complex domain-specific terminology and intricate semantic structures. This paper proposes a novel multi-level feature enhancement framework to overcome these challenges. The framework's contributions are threefold: Firstly, a Hierarchical Knowledge-Driven Data Completion method is proposed to construct the Power Equipment Maintenance Named Entity Recognition (PEM-NER) dataset, leveraging raw data from the State Grid Corporation. Secondly, a Position-Aware Global Attention mechanism is developed and integrated within the transformer architecture. This mechanism effectively captures relative positional information and dataset-scale features, significantly enhancing contextual understanding for NER tasks. Thirdly, a Fine-Grained Information Enhancement Module is designed to refine character-level dependency analysis, thereby improving the precision of entity boundary detection. Extensive evaluations on the PEM-NER dataset and three public benchmarks demonstrate the proposed model's superior performance, especially in recognizing entities within power system texts. The framework exhibits promising applications in knowledge graph construction and question-answering systems within the field of power equipment maintenance.

Keywords Named entity recognition, Power system, Attention mechanism, Deep learning

With the rapid growth of global industrialization and increasing power demand, power systems face great pressure to operate safely and stably, which has led to a large increase in power equipment maintenance work. The data generated from these frequent and intensive maintenance activities is important for evaluating equipment operating conditions, predicting potential failures, and developing data-driven smart grids[1,2]. Therefore, how to use large amounts of maintenance data efficiently has become an urgent challenge for power companies seeking to improve system performance and sustainability. It is important to note that a large part of power system text comes from maintenance work orders, which record detailed information about equipment and the complete process of defect repairs. However, as shown in Table 1, most existing work orders are recorded as unstructured text, and extracting accurate information from these unstructured records is both challenging and time-consuming.

Named Entity Recognition (NER) is a key task in Natural Language Processing (NLP) that aims to identify entities with specific meanings in natural language text and classify their types accurately[3]. It is a basic step for building structured knowledge bases and information retrieval systems[4,5]. Although deep learning has made significant progress in NER, its application in the power sector remains limited. Specifically, extracting entity information from power equipment maintenance work orders faces three major challenges: First, complex language features. The text contains many domain-specific terms and abbreviations, and Chinese text lacks clear word boundaries, which makes semantic understanding more difficult. Second, varying recording styles. Different maintenance personnel have very different writing habits, which leads to multiple ways of expressing the same type of entity. This variation in recording style makes entity recognition more difficult. Third, incomplete data. Because recording standards are not well-developed, the raw data often has missing key information. This not only damages the semantic completeness of sentences but also seriously affects the training performance of supervised learning models.

Existing general-purpose NER models often struggle when processing such complex power domain data. Traditional sequence labeling models and methods based on general pre-trained language models often fail

[1]Department of Electronics and Information Engineering, College of Physical Science and Technology, Central China Normal University, Wuhan, China. [2]OPT Machine Vision Tech Co., Ltd, Dongguan, China. ✉email: qushaocheng@mail.ccnu.edu.cn

Case	Description
Example 1	35kV Baling Substation added 35kV main transformer door frame A-type pole top plate welding, added 35kV main transformer neutral point arrester equal diameter pole top plate welding and grounding lead wire welding.
Example 2	Xi#1 main substation body, 110kV side neutral arrester pre-test, Xi#1 main substation 10kV bus bridge, Xi#1 main substation 35kV side neutral arrester, Xi31 lightning arrester, Xi#1 main substation 10kV bus bridge routine test.
Example 3	AC 220kV Yuwang Substation: 220kV equipment area, 110kV equipment area weed removal, main transformer fire sandbox debris cleaning; 10kV high-voltage room, protection room, battery room, safety equipment room hygiene cleaning.
Example 4	Liuqiao substation Liu #2 main transformer, Liu 35KV arc extinguishing coil, Liu 107, 302, 502 switching unit equipment pre-test and maintenance overhaul, Liu #2 main transformer gas relay calibration and on-load voltage regulator core lifting inspection.

Table 1. Example of raw data sample. (English translation version).

to accurately identify complex entity boundaries and lack the ability to effectively capture global information from the dataset. Furthermore, most existing methods overlook the negative impact of missing data on model training performance. To tackle the aforementioned challenges, we develop a multi-level feature enhancement framework focusing on effectively extracting critical information units from textual records. This framework combines domain knowledge-driven data completion, global attention mechanisms, and fine-grained feature enhancement to form an integrated system. The proposed framework demonstrates superior performance compared to existing approaches, with validation conducted on multiple benchmark datasets.

The main contributions of this paper are as follows:

- We present the Hierarchical Knowledge-Driven Data Completion method, through which we constructed the Power Equipment Maintenance Named Entity Recognition (PEM-NER) dataset using maintenance records from the State Grid Corporation of China. This dataset encompasses seven entity classes pertinent to power equipment maintenance processes, comprising 6,371 sentences and 235,796 characters.
- We propose a novel Position-Aware Global Attention mechanism. This attention mechanism emphasizes relative positional information during computation and employs two global memory units to capture feature information at the scale of the entire dataset. Within this mechanism, we propose a novel Learnable Double Normalization (L-DNorm) method.
- We designed a Fine-Grained Information Enhancement Module to capture character-level local dependencies, enhancing the model's entity boundary detection capabilities.
- Through extensive experimental validation, our proposed model demonstrates superior recognition performance on both the PEM-NER dataset and three public benchmark datasets. This work not only provides new solutions for power system information extraction, but also creates new ideas for data research in related industrial scenarios.

Related work

Named Entity Recognition has made significant progress since its introduction at the Sixth Message Understanding Conference (MUC-6) in 1995. This section reviews the development of NER, from general deep learning models to domain-specific applications, and analyzes the limitations of existing methods in processing power system texts.

Deep learning-based NER model

Early NER research was mainly based on rule-based and statistical models, but deep learning techniques have become the main approach in recent years[6]. Huang et al.[7] were the first to propose a model that integrates bidirectional long short-term memory networks (BiLSTM) with conditional random fields (CRF) for NER. This model demonstrated high accuracy and robustness, establishing itself as a classical approach in the field of NER. Chiu et al.[8] introduced convolutional neural networks (CNNs) to capture character-level features, combining them with BiLSTM to efficiently enhance NER performance.

In recent years, the emergence of pretrained models such as Transformer and BERT has revitalized the field of NER. Yan et al.[9] proposed TENER, an NER model based on the Transformer encoder architecture, which effectively captures both character-level and word-level features. Building on the multi-head attention mechanism within the Transformer framework, Liu et al.[10] developed a multimodal Chinese NER model named USAF, which integrates textual and acoustic features to achieve superior performance in Chinese NER tasks.

Although these models perform well on general datasets, they usually focus on local or long-distance dependencies within sentences. When processing power maintenance work orders, these models often ignore global dataset features across samples, and standard attention mechanisms cannot accurately handle complex relative position information in power texts. This leads to limitations in recognizing dense and structurally complex technical terms.

Domain-specific NER research

With the rapid advancement of deep learning, its applications have expanded across a growing number of research domains[11–14]. This pattern is especially noticeable in the area of NER, where domain-specific NER tasks, tailored to various types of textual data, have become a focal point of academic research. Yu et al.[15] proposed a mineral named entity recognition model based on deep learning, leveraging BERT combined with CRF to effectively identify seven categories of entities within mineralogical texts. Liu et al.[16] employed a CNN to filter a corpus of railway faults caused by electromagnetic interference and further integrated BiLSTM and BERT to

construct an NER model for extracting fault-related entities. Nath et al.[17] modeled the NER task for clinical texts as a multi-label supervised annotation problem, introducing three multi-label entity annotation frameworks aimed at simultaneously identifying entities and their associated attributes. Chu et al.[18] proposed a NER model based on a multi-feature fusion Transformer, which innovatively combines three features, character, word, and radical, to enhance the model's recognition accuracy in aerospace domain texts.

However, NER research for the power industry is relatively limited. Compared to other domains, power equipment maintenance work orders have unique challenges: inconsistent text recording styles and many missing fields (such as missing equipment names or fault locations). Most existing domain-specific models focus on feature fusion or multi-label classification, but have not proposed effective strategies to address the common industrial problem of incomplete data. They also lack fine-grained processing mechanisms for handling unclear entity boundaries in power texts. Therefore, developing a framework that can address missing data, capture global information, and accurately identify complex entity boundaries is important for power system text mining.

Dataset construction

Initial construction

Data annotation

The data used in this study were sourced from maintenance work orders recorded in the Production Management System (PMS) of the State Grid, covering the period from 2016 to 2023. Each work order consists of unstructured textual data in Chinese, as illustrated in Table 1. After initial data cleaning and transformation, a raw dataset containing 6371 samples totaling 205,375 characters was constructed.

The primary objective of the NER task for power equipment maintenance work orders is to extract key information to facilitate further management and analysis by power enterprises. This includes applications such as equipment life prediction, defect correlation analysis, and smart grid construction. To this end, we identified seven entity categories of highest analytical value: EN(Equipment Name), VL(Voltage Level), Line, Sub(Substation), DP(Damage Part), MS(Maintenance Status), and Time. The dataset annotation process utilized the BMES (Begin-Middle-End-Single) sequence labeling framework, which categorizes entity boundaries as follows: B designates the initial position of multi-character entities, M and E sequentially identify subsequent and terminal segments of such entities, while S exclusively labels standalone single-character entities.

Quality control

The annotation process for this study was conducted by a team of four annotators to address the demands of processing a large-scale sample set. To ensure annotation quality, we implemented a strict quality control process: first, all annotators underwent systematic training based on equipment maintenance specification documents from power grid companies; subsequently, quantitative assessment of inter-annotator consistency was conducted using the Fleiss Kappa coefficient[19]. Specifically, 10% of samples were randomly selected from the entire dataset for independent annotation by all four annotators. Calculations revealed a Fleiss Kappa coefficient of 0.83 among annotators, which according to the classification standards of Landis and Koch (1977)[20], falls within the "almost perfect agreement" range (0.81–1.00), indicating high reliability and consistency in the annotation process of this study. To further ensure reliability, any inconsistent annotations identified during the process were reviewed and resolved through group discussions and final adjudication by a senior domain expert.

Hierarchical knowledge-driven data completion

Early maintenance work orders for power equipment suffered from incomplete documentation rules and heterogeneous recording practices across personnel, resulting in substantial missing data fields. Critical omissions included equipment name, damage parts, maintenance status, and location information (e.g., grid lines or substations). These data deficiencies significantly compromised the training efficacy and practical utility of NER models. Conventional imputation methods, such as Lagrange interpolation, K-nearest neighbors (KNN), and Markov models, rely on statistical correlations to estimate missing values, prioritizing global data distribution repair over discrete textual field reconstruction. Such approaches fail to address semantic interdependencies in unstructured text. Meanwhile, deep learning-based methods (e.g., generative adversarial networks, GANs) exhibit opaque decision-making processes and risk generating entities violating equipment-type constraints, potentially contravening safety protocols. To address this issue, we propose a Hierarchical Knowledge-Driven Data Completion (HKDC) method, which achieves targeted data completion by integrating domain prior knowledge and hierarchical relationship constraints.

Hierarchical knowledge base construction

To ensure the semantic and rationality of the operation and maintenance work orders after data completion, we constructed two hierarchical dictionaries based on the power equipment operation and maintenance specification documents of the power grid company and historical operation and maintenance work orders:

- *Equipment Maintenance Dictionary* A three-tier tree-structured entity constraint repository, with equipment names as root nodes, damage parts as secondary nodes, and maintenance status as leaf nodes. Figure 1 shows some of the contents in this dictionary.
- *Grid Topology Dictionary* A hierarchical system designating substations as parent nodes and their affiliated power lines as child nodes, enabling bidirectional substation-line mapping.

Data deficiency detection and completion

After annotating entities in original work order texts, rule matching is used to locate the missing fields. If the sentence appears to contain the name of the equipment but no defective part, or contains the line but no

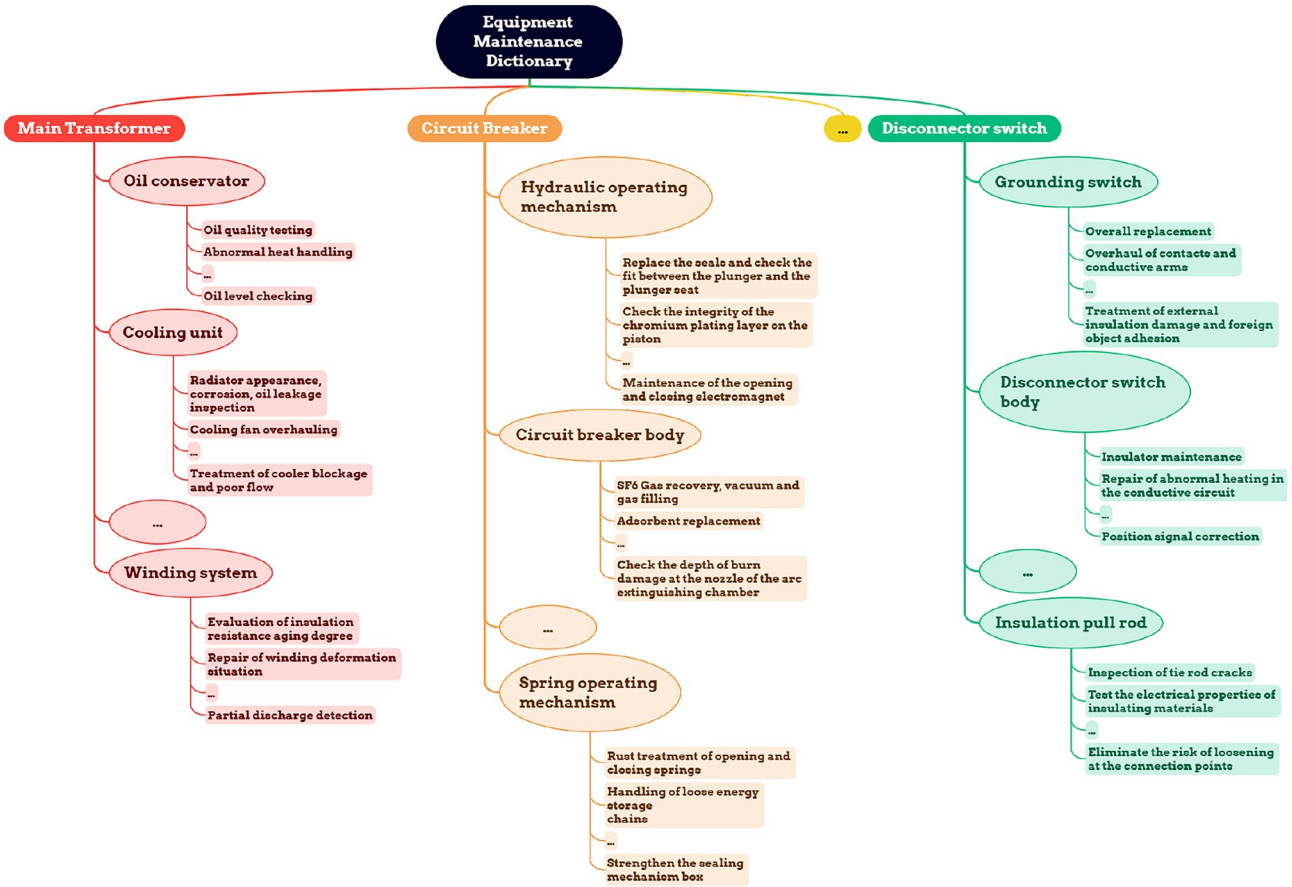

Fig. 1. Example of equipment maintenance dictionary content.

substation, etc., it is judged as a missing entity. Aiming at the samples with missing entities, a dynamic inference strategy is used to make up the data. First, the existing entity category information in the current sample is extracted based on the labeled BMES labels. Then, query the hierarchical knowledge base according to the existing entities to obtain the set of candidate entities, and finally randomly select the corresponding entities to be inserted into the current sample and annotate the inserted entities. For instance, if a sentence includes the terms "main transformer" (EN) and "oil conservator" (DP), the corresponding operation and maintenance statuses (MS), such as "oil quality testing" and "abnormal heat handling," are retrieved from the equipment operation and maintenance dictionary. Subsequently, one of these statuses is randomly selected and incorporated into the current sentence, followed by the completion of the BMES annotation for the entire sentence.

Considering the existence of incomplete information of certain equipment operation and maintenance records in real industrial scenarios, excessive data complementation may introduce noise or bias, which in turn affects the accuracy of the model in practical use. In order to avoid this situation, we only complemented the data in the training set, and in addition, we complemented only 40% of the data sample, so that the model has the ability to handle incomplete data.

In conclusion, our proposed HKDC method solves the pain points of missing logic and semantic conflicts in textual completion for power operation and maintenance through hierarchical knowledge-driven approach. Compared with the traditional interpolation method and deep generative model, it has significant advantages in text-based data processing, domain rule compliance, and result interpretability. After data annotation and data completion, we constructed the Power Equipment Maintenance NER (PEM-NER) dataset. The dataset contains a large number of typical maintenance data of various types of power equipment, covering 7 types of entities, a total of 6,371 sentences and 235,796 characters. The specific number and percentage of each entity category is shown in Fig. 2.

Dataset feature analysis

The PEM-NER dataset, specific to the power sector, presents four distinct characteristics that also pose challenges for named entity recognition:

- *Text complexity in Chinese* The dataset is in Chinese, which lacks obvious morphological changes and spaces between words, making it difficult to identify word boundaries.
- *Inconsistent writing styles* Different staff members record maintenance work orders in unique writing styles, so recording styles and completeness of records vary widely.

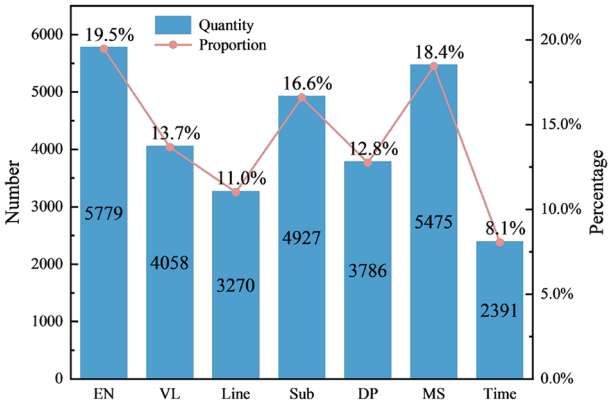

Fig. 2. Number and percentage of each entity category in PEM-NER dataset.

- *Variable text length* The varying complexity of faults and maintenance procedures across different types of equipment results in substantial differences in the length of maintenance work order texts. Based on statistical analysis of the existing data, the shortest work order contains 8 characters, while the longest extends to 362 characters.
- *Domain-specific terminology and abbreviations* The dataset contains numerous specialized terms and acronyms unique to the power industry, with a wide variety of entities and complex semantics, increasing the difficulty of accurate entity recognition.

Proposed method

Figure 3 depicts the general structure of our proposed model. Based on the transformer framework, we propose a novel position-aware global attention(PAGA) mechanism, and construct a fine-grained information enhancement module(FIEM) as another branch to better capture local information. Additionally, we introduce ELECTRA[21] for character embedding to enhance contextual representations and employ CRF to obtain optimal label sequences.

Following a brief overview of transformer fundamentals, this section provides detailed descriptions of each component within our proposed model.

Preliminary of self-attention and position encoding

The Transformer model's success stems from its attention mechanism[22], which excels in capturing long-range dependencies and enabling parallel processing[23].

The self-attention mechanism enables bidirectional contextualization across sequence positions through inter-token dependency modeling. For an input sequence $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ comprising n tokens, three parameterized projection heads derive the fundamental operators:

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_q, \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_k, \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_v, \tag{1}$$

where $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v$ denote trainable projection matrices. The contextual aggregation is then formulated as:

$$\text{Attention} = \phi\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V}, \tag{2}$$

with $\phi(\cdot)$ representing the row-wise softmax normalization and d_k corresponding to the latent subspace dimensionality of attention heads.

While parallel processing improves computational efficiency, it eliminates sequential information[24]. Position encodings address this by encoding token positions into fixed-length vectors using sine and cosine functions:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right), PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right), \tag{3}$$

where *pos* stands for the position, *d* is the encoding dimension and *i* is the dimension index.

Character embedding

Character embeddings constitute a fundamental component in named entity recognition systems, transforming discrete symbolic tokens into dense numerical embeddings that capture latent semantic features.[25]. This process effectively decreases the dimensions of the lexical representation, generating a compact and continuous embedding. By incorporating character-level information, the model gains a richer understanding of word semantics within the surrounding text, improving its ability to interpret context and identify entities accurately.

Numerous NER models rely on approaches like Word2Vec or CNN to achieve character embedding, though these techniques come with certain limitations[26,27]. Notably, they often fail to recognize distant connections in

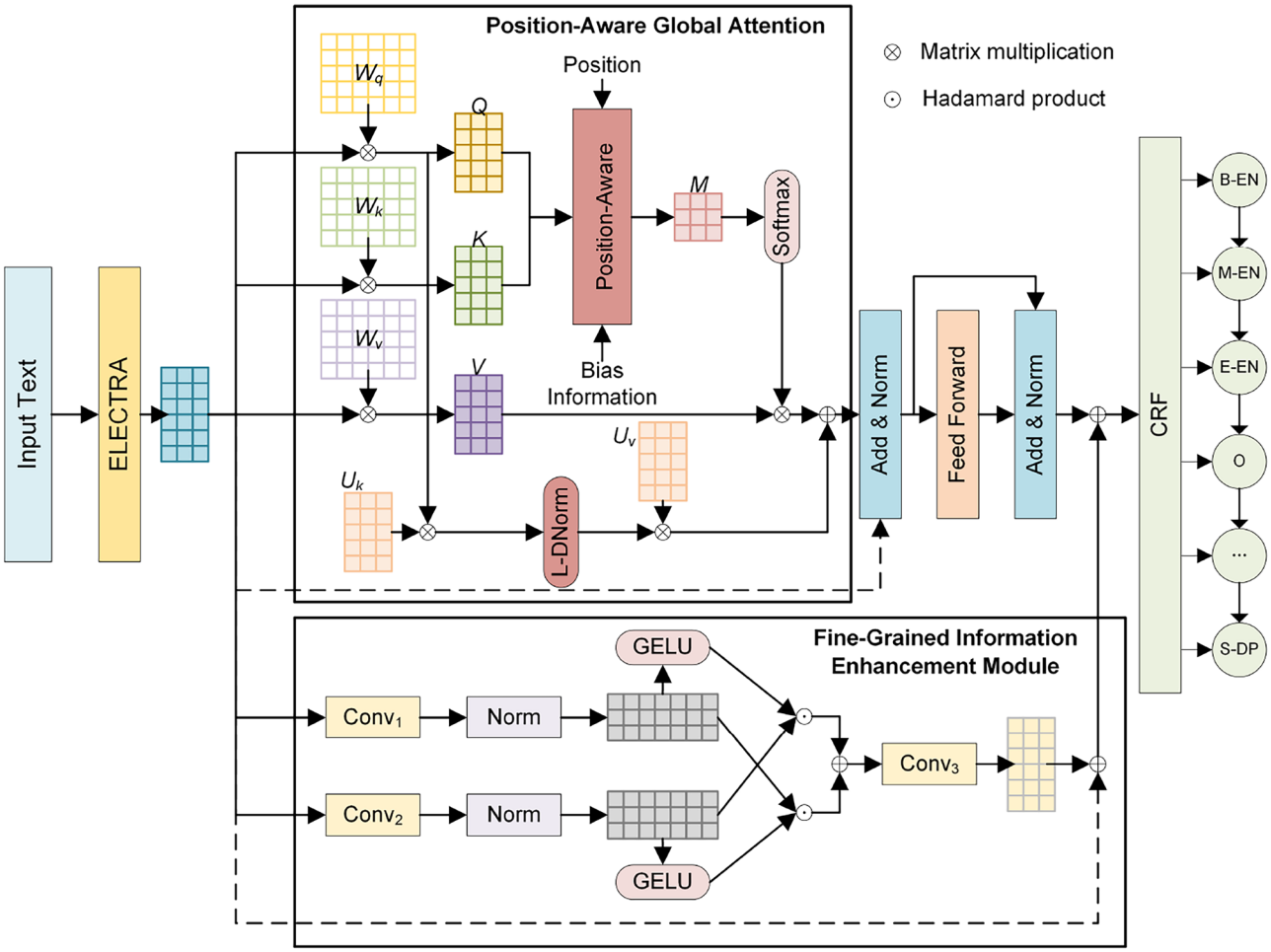

Fig. 3. The overall architecture of our proposed model.

string-based inputs because their learning depends on short segments of context, causing important structural patterns to be overlooked[28]. The emergence of pre-trained language models such as BERT, XLNet and RoBERTa has effectively addressed prior limitations, establishing them as the predominant methods for character embedding.

In this study, motivated by industrial deployment requirements that necessitate enhanced computational efficiency, we implemented ELECTRA for character embedding of input text. ELECTRA's innovation lies in its Replaced Token Detection task, which supplants BERT's Masked Language Modeling (MLM) approach. This architecture employs a Generator to produce replacement tokens and a Discriminator to determine whether original tokens have been substituted, thereby utilizing all input tokens for learning and substantially improving training efficiency. With the same number of parameters, ELECTRA has better performance compared to models such as Bert and RoBERTa, and is computationally efficient, e.g., only 1/4 of the training volume is needed to reach the RoBERTa level.

Position-aware global attention mechanism

In the original Transformer, position encoding is achieved through linear combinations of sine and cosine functions. While this position encoding approach enhances the model's capacity to perceive distances, it exhibits limitations in discriminating the directional relationships among tokens[29]. Additionally, the conventional self-attention mechanism exclusively focuses on correlations between different positional features within individual samples, neglecting potential latent associations across different samples[30]. To address these limitations, we propose position-aware global attention(PAGA) mechanism. The PAGA mechanism implements novel positional embedding and attention score computation methodologies, incorporating two learnable global units to capture inter-sample correlational information. The process initiates with the computation of query ($\boldsymbol{Q}$), key ($\boldsymbol{K}$), and value ($\boldsymbol{V}$) representations:

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_q, \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_k, \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_v, \tag{4}$$

where the matrices Q, K, and V all belong to $\mathbb{R}^{l\times d_k}$, with W_q, W_k, $\mathrm{W}_v \in \mathbb{R}^{d\times d_k}$ serving as their respective learnable projection parameters. Subsequently, we compute the relative positional embedding vector representation between two tokens:

$$\boldsymbol{P}_{t-j} = \left[\ldots \sin\left(\frac{t-j}{10000^{2i/d_k}}\right)\cos\left(\frac{t-j}{10000^{2i/d_k}}\right)\ldots\right]^T, \tag{5}$$

where *t* represents the current position (position of query), *j* represents the position to be attended (position of key), and $\boldsymbol{P}_{t-j}$ depends only on the relative distance (*t*-*j*), not on the specific positions *t* or *j*. To fully utilize the positional information contained in $\boldsymbol{P}_{t-j}$ and enhance the expressive capacity of the attention mechanism, we introduce an improved position-aware attention score computation method. Specifically, for the position-aware attention score $\boldsymbol{M}_{t,j}$ between positions *t* and *j*, we decompose it into four components:

$$\boldsymbol{M}_{t,j} = \boldsymbol{Q}_t\boldsymbol{K}_j^T + \boldsymbol{Q}_t\boldsymbol{P}_{t-j}^T + \boldsymbol{u}\boldsymbol{K}_j^T + \boldsymbol{v}\boldsymbol{P}_{t-j}^T, \tag{6}$$

where $\boldsymbol{Q}_t\boldsymbol{K}_j^T$ represents the basic content attention term, computing the content correlation between query position *t* and key position *j*, reflecting the semantic associations between tokens in the sequence; $\boldsymbol{Q}_t\boldsymbol{P}_{t-j}^T$ is the interaction term between query and relative positional embedding, enabling the query to perceive relative positional information and allowing the model to learn position-based attention patterns; $\boldsymbol{u}$ and $\boldsymbol{v}$ are two learnable bias vectors which, through their interactions with relative positional embeddings and content information, enhance the model's capability to capture both positional and content relationships within the sequence.

In calculating the relative positional embedding of the tokens at position *t* (*j*=0) and -*t* (*j*=2*t*), it is obtained according to (5):

$$\begin{aligned}\boldsymbol{P}_t &= \left[\begin{array}{cccccc}\sin(z_0 t) & \cos(z_0 t) & \cdots & \sin(z_{\frac{d}{2}-1}t) & \cos(z_{\frac{d}{2}-1}t)\end{array}\right],\\ \boldsymbol{P}_{-t} &= \left[\begin{array}{cccccc}-\sin(z_0 t) & \cos(z_0 t) & \cdots & -\sin(z_{\frac{d}{2}-1}t) & \cos(z_{\frac{d}{2}-1}t)\end{array}\right],\end{aligned} \tag{7}$$

where $z_i = \frac{1}{10000^{2i/d_k}}$. From the results, it can be seen that $\boldsymbol{P}_t \neq \boldsymbol{P}_{-t}$, which indicates that the positional embedding is different in the two directions. So by calculating the attention score in the above way the positional information between different tokens can be perceived.

In standard self-attention mechanisms, dependency modeling is limited to local context within a single input sequence. This prevents the model from directly using information from the entire dataset. To address this limitation, we introduce two global shared learnable memory units, denoted as U_k and U_v. These global memory units are shared across all attention heads and all training samples. They serve as a global latent dictionary that captures high-level semantic prototypes specific to the power maintenance domain. Specifically, U_k acts as a set of semantic keys representing different entity patterns or clusters observed across the entire corpus, while U_v stores the corresponding feature representations.

This interaction is designed as a global addressing and retrieval process. For an input token query Q, the model computes similarity scores with the global U_k to obtain global context:

$$\boldsymbol{I}_{\mathrm{Global}} = \text{L-DNorm}(\boldsymbol{Q}\boldsymbol{U}_k^T)\boldsymbol{U}_v, \tag{8}$$

where $U_k, U_v \in \mathbb{R}^{S\times d_k}$, and *S* is the dimension of the global memory units, representing the number of memory slots. This operation addresses the global memory and allows the model to match current local tokens with learned dataset-level prototypes. Based on these addressing weights, the model then retrieves relevant global context from U_v.

During the computation of $\boldsymbol{Q}\boldsymbol{U}_k^T$, if a token's feature vector q_1 in $\boldsymbol{Q}$ contains abnormally large values, it results in large dot products between q_1 and all memory vectors in $\boldsymbol{U}$. When using single softmax normalization, this leads to extreme distributions in the corresponding row of the attention matrix, even when the token should not have strong correlations with certain memory vectors. Applying normalization in both row and column directions prevents anomalous feature values of individual tokens from dominating the entire attention distribution.

However, this introduces another challenge: overly balanced attention distributions might impair the model's feature recognition capabilities. To address this, we introduced two learnable parameters, α and β, establishing a learnable double normalization (L-DNorm) method. This enables the model to dynamically adjust normalization intensity in different directions during training. The computation process is as follows:

$$A'_{i,j} = \boldsymbol{Q}\boldsymbol{U}_k^T, \tag{9}$$

$$\boldsymbol{A}''_{i,j} = \frac{\exp(A'_{i,j}/\alpha)}{\sum_n \exp(\boldsymbol{A}'_{n,j}/\alpha)}, \tag{10}$$

$$\boldsymbol{A}_{i,j} = \frac{\boldsymbol{A}''_{i,j}}{(\sum_n \boldsymbol{A}''_{i,n})^\beta}. \tag{11}$$

Ultimately, our design of position-aware global attention mechanism can be expressed as:

$$Attn = \text{softmax}(\boldsymbol{M})\boldsymbol{V} + \boldsymbol{I}_{Global} \tag{12}$$

Fine-grained information enhancement module

PAGA mechanism is adept at processing global contextual information; however, it may lack precision in capturing finer details. Especially in the named entity recognition task, the model needs to accurately recognize a specific entity with a relatively small number of words in a long sentence. To address this issue, we propose a fine-grained information enhancement module (FIEM) aimed at augmenting the model's capability to capture local information, thereby improving its accuracy in entity recognition within text. The architecture of this module is depicted in Fig. 3 and is primarily composed of CNNs. CNNs excel at extracting features from localized windows, effectively capturing phrase-level characteristics that are especially beneficial for named entity recognition[31].

The FIEM employs two parallel convolutional layers with distinct kernel sizes. The smaller kernel is designed to capture local fine-grained features, while the larger kernel is tasked with identifying broader contextual dependencies. An interactive design is implemented between the two convolutional layers, such that the output of each layer modulates the feature extraction of the other, facilitating feature interaction through element-wise multiplication. The module is calculated as follows:

$$Out_1 = \text{Norm}(\text{Conv}_1(X)) \odot \text{GELU}(\text{Norm}(\text{Conv}_2(X))), \tag{13}$$

$$Out_2 = \text{Norm}(\text{Conv}_2(X)) \odot \text{GELU}(\text{Norm}(\text{Conv}_1(X))), \tag{14}$$

$$Out = \text{Conv}_3(Out_1 + Out_2) + X, \tag{15}$$

where Conv_1 and Conv_2 represent 1-dimensional convolutional layers with small and large convolutional kernels, respectively. Norm denotes the application of layer normalization, which enhances training stability, while GELU refers to the activation function employed in the module. Conv_3 maps the processed features back to the original feature dimensions. Finally, a residual connection is employed to mitigate the vanishing gradient problem. In the experimental section, we conducted a detailed analysis of the impact of convolutional kernel sizes in the convolutional layers on model performance.

Conditional random fields

Conditional Random Fields (CRF) represents a probabilistic framework designed to address sequence labeling challenges in machine learning[32]. This architecture excels at capturing inter-label dependencies, making it particularly advantageous for tasks where traditional independent classification approaches prove insufficient[7].

Given an input sequence $\mathrm{X} = [x_1, \ldots, x_n]$ and its associated output sequence $\mathrm{Y} = [y_1, \ldots, y_n]$, the conditional random field defines their relationship via:

$$P(\mathrm{Y}|\mathrm{X}) = \frac{\exp\left(\sum_{i=1}^{n}(A_{y_i,y_{i+1}} + P_{i,y_i})\right)}{\sum_{\mathrm{Y}'}\exp\left(\sum_{i=1}^{n}(A_{y'_i,y'_{i+1}} + P_{i,y'_i})\right)}, \tag{16}$$

$A_{y_i,y_{i+1}}$ quantify the compatibility between adjacent labels, P_{i,y_i} evaluate the appropriateness of assigning label y_i at position i. The denominator serves as a normalization term, summing over all possible label sequences (Y') to ensure proper probability distribution properties. The optimization objective focuses on maximizing the conditional probability $P(\mathrm{Y}|\mathrm{X})$ during the training phase.

Experiments

Datasets and experimental configurations

Datasets

We conducted a series of analyses and experiments on model performance using our own constructed PEM-NER dataset. Subsequently, to validate the model's effectiveness across diverse languages and domains, we conducted comparative experiments on three public benchmark datasets. These included one English dataset, CoNLL-2003[33], and two Chinese datasets, namely Resume[34] and the China People's Daily Corpus[35]. Table 2 provides the specifics of each public benchmark dataset. The detailed description of all the datasets used is provided below.

- *PEM-NER* is derived from the power equipment maintenance work order, which adopts the BMES annotation system, including EN, VL, Line, Sub, DP, MS, and Time with 7 categories of entities, 6,371 sentences, and 235,796 characters. It is a customized NER dataset for the power sector.
- *CoNLL-2003* is a classic NER dataset with data from the Reuters News Corpus. It contains four entity types: PER (person's name), ORG (organization's name), LOC (place's name), and MISC (other proper name). Since its release, CoNLL-2003 has become a standard test set for evaluating the performance of NER models, and numerous research efforts have used the F1 score on this dataset as an important measure of model performance.
- *Resume* is tailored for Named Entity Recognition applications, comprising professional profiles across diverse sectors and job roles. It annotates eight distinct entity types: personal names, nationalities, geographic locations, academic qualifications, occupational fields, institutional affiliations, job titles, and ethnic groups.
- *China People's Daily Corpus*, sourced from China's official news publications, exhibits rigorous linguistic standardization and formal discourse patterns. This resource spans multiple thematic areas such as govern-

Dataset	Type	Train	Dev	Test	Entity type
CoNLL-2003	Sentences	15.0k	3.4k	3.6k	4
	Chars	203.6k	51.4k	46.4k	
	Entities	23.5k	5.9k	5.6k	
Resume	Sentences	3.8k	0.46k	0.48k	8
	Chars	124.1k	13.9k	15.1k	
	Entities	13.4k	1.5k	1.6k	
People's Daily	Sentences	20.86k	2.32k	4.64k	3
	Chars	979.2k	109.9k	219.2k	
	Entities	33.9k	3.8k	7.7k	

Table 2. Statistical summary of multiple datasets.

ment affairs, financial trends, cultural phenomena, and social dynamics, providing rich domain-specific lexical resources that reflect real-world NER challenges.

Evaluation metrics and hyperparameters setting

In our experiments, we used precision (P), recall (R) and F1 score to evaluate model performance. These metrics are defined as follows:

$$\begin{aligned} \mathrm{P} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \times 100\%, \\ \mathrm{R} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100\%, \\ \mathrm{F1} &= 2 \times \frac{\mathrm{P} \times \mathrm{R}}{\mathrm{P}+\mathrm{R}} \times 100\% \end{aligned} \tag{17}$$

where TP denotes correctly predicted positive samples, FP represents samples incorrectly predicted as positive, and FN indicates positive samples incorrectly predicted as negative. Precision measures the accuracy of positive predictions, recall indicates the proportion of actual positives correctly identified, and F1 score balances these metrics by calculating their harmonic mean.

During the model training process, the improved encoder architecture consists of two layers, each comprising six attention heads. The Adam optimizer is employed for training, with the learning rate and batch size set to 0.0009 and 128, respectively. To mitigate overfitting, a dropout rate of 0.4 is applied.

Baseline models

To comprehensively evaluate the effectiveness of our proposed method and ensure a rigorous comparison within the power equipment maintenance domain, we selected a diverse set of baseline models ranging from classical deep learning architectures to strong pre-trained language models. For traditional sequence labeling models, we selected BiLSTM-CRF[7], which captures contextual features through bidirectional LSTM and performs sequence decoding via a CRF layer; BiLSTM-Attention-CRF[36], which enhances recognition capabilities for key information through attention mechanisms; and Lattice LSTM[34], which effectively integrates lexical information tailored to the characteristics of Chinese text. In terms of convolutional network models, CNN-BiLSTM[8] combines the local feature extraction capability of CNN with the sequence modeling capability of BiLSTM; IDCNN-CRF[37] uses an inflated convolutional network to improve the efficiency and sense field of feature extraction.

In the category of pre-trained language models, we selected a diverse set of benchmarks to evaluate model performance across different architectural strategies. We employed standard baselines including BERT-CRF[38], RoBERTa[39], and SpanBERT[40]. To assess the model's capability against varied mechanisms, we also included Longformer[41] for long-sequence processing, DeBERTa[42] for its disentangled attention design, and ModernBERT[43] as a representative of recent efficient architectures. Additionally, TadNER-CRF[44] was included to represent hybrid frameworks combining sequence labeling with contrastive learning.

Experimental validation of hierarchical knowledge-driven data completion

We first evaluated the impact of our proposed Hierarchical Knowledge-Driven Data Completion method on model training. Table 3 presents the performance metrics of all employed models on the PEM-NER dataset before and after the application of the data completion strategy.

Results demonstrate that the data completion strategy consistently enhanced performance across all tested models, underscoring the efficacy and suggesting the generalizability of our proposed method. Following data completion, all models achieved notable performance improvements. On average, F1 scores increased by 3.2 percentage points, indicating that the HKDC method effectively mitigates the negative impact of data deficiencies on model training. Notably, benefits were observed across architectural complexity, from simpler CNN-BiLSTM models (3.0% F1 improvement) to more sophisticated ModernBERT models (3.6% F1 improvement), thus highlighting the model-agnostic characteristics of our approach.

When examining changes in precision (P) and recall (R), we observed that most models demonstrated more substantial improvements in recall following data completion. This phenomenon aligns with the conceptual

Model	Before			After		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
CNN-BiLSTM	68.5	66.2	67.3	69.2	71.5	70.3
IDCNN-CRF	70.1	68.9	69.5	71.8	73.1	72.4
BiLSTM-CRF	71.4	73.8	72.6	75.9	74.5	75.2
Lattice LSTM	74	73.5	73.7	76.3	77.6	76.9
BiLSTM-Attention-CRF	75.9	75.1	75.5	78.1	78.9	78.5
BERT-CRF	77.5	74.3	75.9	79.3	77.1	78.2
RoBERTa	79.8	79.2	79.5	81.4	82.7	82.0
Longformer	81.2	80.5	80.8	83.1	83.9	83.5
SpanBERT	82.1	80.6	81.3	85.2	83.9	84.5
DeBERTa	85.4	83.7	84.5	89.6	88.5	89.0
TadNER-CRF	85.3	84	84.6	88.6	89.3	88.9
ModernBERT	87.7	86.5	87.1	91.1	90.3	90.7
Ours	**89.1**	**88.2**	**88.6**	**92.8**	**91.9**	**92.3**

Table 3. Comparison of model performance before and after data completion. Bold values indicate the best results.

design of the HKDC method: through knowledge-driven completion strategies, models can identify more potential entities, particularly those previously overlooked due to incomplete information, thereby enhancing recall. We attribute the comparatively limited improvement in precision to the following factor: although the completed data was constructed based on domain knowledge, differences still exist between the entity instances introduced in specific contexts and the actual test distribution. Consequently, while models expanded their recognition scope (improving recall), their precision judgment capabilities faced certain constraints.

The experimental results strongly suggest that our HKDC method effectively mitigates entity omission issues in power equipment maintenance texts, notably enhancing the performance of various NER models. By integrating domain knowledge with hierarchical constraints, the HKDC method provides richer training samples while ensuring the rationality and consistency of the completed data, thereby strengthening model generalization capabilities.

Performance comparison on PEM-NER dataset

This section presents a comparison between the proposed framework and baseline models on the completed PEM-NER dataset (see the "After" column in Table 3). Our model achieves an F1 score of 92.3%, which outperforms all baseline methods. It is worth noting that although large-scale pre-trained models such as DeBERTa and ModernBERT provide strong semantic representation capabilities, our framework still achieves an F1 score that is 1.6 percentage points higher than ModernBERT. This advantage comes from specific architectural improvements that address the inherent limitations of general pre-trained models when applied to power equipment maintenance texts.

The main limitation of standard pre-trained models is that their self-attention mechanism can only model dependencies within a single input sequence. Although this approach works well for general language understanding, it often fails to explicitly capture the cross-sample statistical patterns that are typical in standardized maintenance logs. In contrast, our PAGA mechanism introduces a global memory unit to model dataset-level features, which allows the model to use global context that cannot be obtained through single-sequence processing. In addition, general pre-trained models typically use subword tokenization, which may lack the necessary granularity for precisely identifying entity boundaries in domain-specific texts with dense terminology and similar structures. Our framework addresses this problem by integrating FIEM, which uses multi-scale convolution to extract character-level local dependencies. The model's superior precision compared to ModernBERT demonstrates the effectiveness of this fine-grained feature extraction. This significant improvement in precision indicates that our method effectively reduces false positive predictions.

Ablation study

To investigate the effectiveness of different components in our proposed method, we conduct comprehensive ablation studies on the PEM-NER dataset. Table 4 shows the precision, recall, and F1 scores of different model variants on the test set.

The baseline model is set as Transformer+CRF. The experimental results provide insights into the specific contributions of each module. First, when the baseline attention mechanism is replaced with PAGA (without global units), the F1 score increases to 79.4%. This shows that the relative positional embedding in PAGA captures directional relationships between tokens more effectively than the absolute positional encoding in standard Transformers. More importantly, adding global units further increases the F1 score to 85.3%. This large improvement indicates that standard self-attention mechanisms are limited by processing each sample in isolation. In contrast, global units work as a dataset-level memory bank that captures cross-sample entity co-occurrence patterns (such as common pairs of specific substations and lines), which helps address the sparse context and fragmented text that often appear in maintenance logs.

Model	P(%)	R(%)	F1(%)
Baseline	76.8	73.9	75.3
+PAGA without global units	80.5	78.3	79.4
+PAGA	85.7	84.9	85.3
+FIEM	85.4	83.1	84.2
+PAGA+FIEM	88.3	88.9	88.6
+All	**92.8**	**91.9**	**92.3**

Table 4. Ablation results of the proposed model. Bold values indicate the best results.

Model	Params (M)	Training speed (s)	Inference speed (ms)	FLOPs (G)	F1 (%)	Efficiency ratio
DeBERTa	183	25.8	11.7	22.5	89.0	3.96
ModernBERT	149	21.7	**11.2**	17.8	90.7	5.10
Ours	**112**	**17.6**	13.4	**13.2**	**92.3**	**6.99**

Table 5. Comparison of computational efficiency and model performance. Bold values indicate the best results.

Second, the model with only FIEM achieves an F1 score of 84.2%. This confirms its effectiveness in extracting character-level local features. Because Chinese power equipment texts lack clear word boundaries and contain many complex equipment codes that mix numbers, letters, and Chinese characters, FIEM's multi-scale convolution kernels can effectively capture these local morphological features and n-gram patterns, which allows it to identify precise entity boundaries more accurately than pure Transformer architectures.

Finally, the combination of PAGA and FIEM achieves an F1 score of 88.6%. This demonstrates clear complementarity between the two: PAGA handles long-distance semantic dependencies and global consistency, while FIEM focuses on precise local boundaries. This combination effectively reduces common semantic classification errors and boundary segmentation errors in single models, which leads to balanced improvements in both precision and recall. The complete model (+All) achieves the highest F1 score of 92.3% after introducing ELECTRA, which confirms that strong pre-trained semantic representations are an important foundation for our proposed feature enhancement modules to work effectively.

Computational efficiency analysis

To evaluate the deployment potential of the model in industrial scenarios, this section compares the computational costs of the proposed model with two best-performing baseline models, DeBERTa and ModernBERT, on the PEM-NER dataset. The comparison covers multiple dimensions, including parameter size, training and inference time, floating-point operations (FLOPs), and efficiency ratio. The experimental results are shown in Table 5, where the efficiency ratio is defined as the F1 score achieved per unit of FLOPs. Overall, the proposed model significantly reduces parameter size and computational cost while maintaining high recognition accuracy.

The proposed model has only 112M trainable parameters because it uses the more compact ELECTRA as the underlying character embedding representation and employs a carefully designed feature enhancement module instead of stacking multiple Transformer layers. Compared to DeBERTa with 183M parameters and ModernBERT with 149M parameters, the proposed model reduces the model size by 38.8% and 24.8%, respectively. This structural simplification directly improves training efficiency. Under the same hardware and batch size settings, the proposed model requires 17.6s per training epoch, which is significantly lower than the 25.8s for DeBERTa and 21.7s for ModernBERT. In other words, within a given training budget, the proposed model can complete more parameter updates or process more data samples. This is particularly important for power system maintenance scenarios that require regular retraining and continuous integration of new equipment and operational data.

In terms of inference latency, the proposed model requires an average of 13.4 ms per sample, which is slightly higher than the 11.2 ms for ModernBERT and 11.7 ms for DeBERTa. This slight increase in latency mainly comes from the parallel computation overhead introduced by the FIEM module and the sequential dependency inherent in the CRF layer during decoding. The additional overhead is mainly concentrated in the feature interaction of a few layers rather than large-scale repetitive computations across the entire network. However, the proposed model achieves the highest efficiency ratio of 6.99, which is significantly better than the 5.10 for ModernBERT and 3.96 for DeBERTa. Although the inference time increases slightly, this investment of computational resources is highly valuable from the perspective of performance conversion. Therefore, this trade-off is acceptable in power system applications where real-time requirements are not extremely strict but overall throughput and energy efficiency are more important.

Evaluation of generalization ability

In real-world power equipment maintenance, new devices, new components, and new processes appear continuously. During training, the model cannot be exposed to all possible entity forms in advance. As a result, its performance on seen entities alone cannot fully reflect its usefulness in practical applications. To further

examine the model's generalization ability on unseen entities, that is, whether it truly learns contextual semantics and character-level structural features instead of only memorizing entity strings in the training set, we conduct a dedicated analysis of its recognition performance on unseen entities in this section.

Based on the original PEM-NER test set, we extract entities that do not appear in the training set at the entity level and construct an unseen entities subset. Using this subset, we compare three models: BiLSTM-CRF, ModernBERT, and our proposed model. To make it easy to compare model performance on unseen entities, we report the F1 scores of each model on the unseen entities subset. We also show their overall F1 scores on the full PEM-NER test set and the corresponding relative drop, as illustrated in Fig. 4.

The experimental results show that our model has the smallest performance drop. This result indicates that, when it encounters entities that never appear in the training set, the model can still make correct predictions by relying on contextual semantics and character-level structural information. In particular, FIEM uses multi-scale one-dimensional convolutions to explicitly model local character-level dependencies and morphological patterns. This design helps the model recognize device and component names that are new in surface form but similar in morphological structure. PAGA introduces relative positional encoding and global memory units. In this way, it captures dependencies within each sentence and also learns global co-occurrence patterns at the dataset level. As a result, it maintains relatively stable discrimination ability even when the context changes slightly or the entity text is completely unseen. The joint effect of FIEM and PAGA reduces boundary shifts and label confusion in unseen entities scenarios and thus clearly alleviates the drop in F1.

Parameter sensitivity analysis

Global memory unit dimension and normalization strategies

To investigate the impact of different parameter settings on model performance, we conducted comprehensive experiments on the dimension of global memory units and normalization strategies in PAGA. Parameter S represents the dimension of global memory units, while the two normalization methods are: the conventional Softmax and the L-DNorm approach proposed in this work.

As shown in Figs. 5 and 6, both parameter S and the choice of normalization method significantly influence model performance. We observed two key findings:

Regardless of the normalization method employed, the model achieves optimal performance when S=64. Performance deterioration is observed when S deviates from this optimal value, with particularly notable declines at S=8 and S=256. This indicates that when S is too small, the model's feature representation capacity is insufficient to fully capture complex semantic characteristics of entities. Conversely, when S is too large, it introduces excessive potentially irrelevant feature dimensions that interfere with the model's ability to learn truly important features. Additionally, a larger parameter space makes the optimization process more challenging.

L-DNorm demonstrates consistent advantages over Softmax across all S values. The box plots reveal that results obtained using L-DNorm exhibit higher means and medians compared to those obtained using Softmax. Furthermore, the F1 score distribution with L-DNorm is more concentrated, indicating more stable training dynamics. This improvement can be attributed to normalization in both row and column directions and the dynamic adjustment of two learnable parameters.

Different convolutional kernel size combinations

To further analyze the effect of the size of the convolutional kernel in FIEM on the overall performance of the model, we conducted six sets of experiments on the PEM-NER dataset. As shown in Fig. 7, we first conducted experiments with different combinations of convolutional kernel sizes for two convolutional layers, and then we also investigated the impact of more complex configurations on the model performance by parallelizing three convolutional layers.

The experimental results show that the combination of convolution kernel sizes has a significant effect on the prediction performance of the model. When both convolution layers use a kernel size of 1 (1,1), the F1 value of the model is as low as 87.9%, and it is clear that this combination of modules does not enable the model to obtain sufficient fine-grained feature extraction. The performance of the model is improved by increasing the kernel size of the second convolutional layer, the optimal performance is achieved with the kernel size combination of

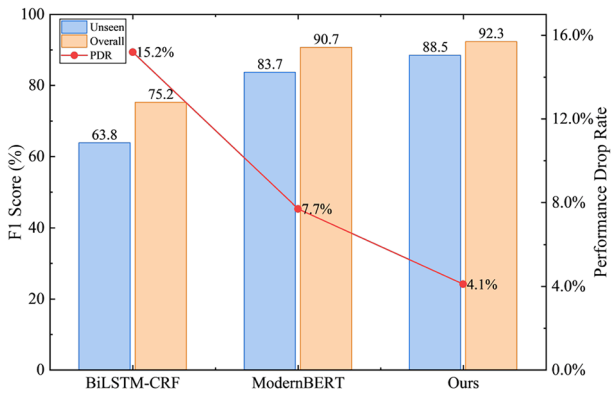

Fig. 4. Generalization analysis on unseen entities.

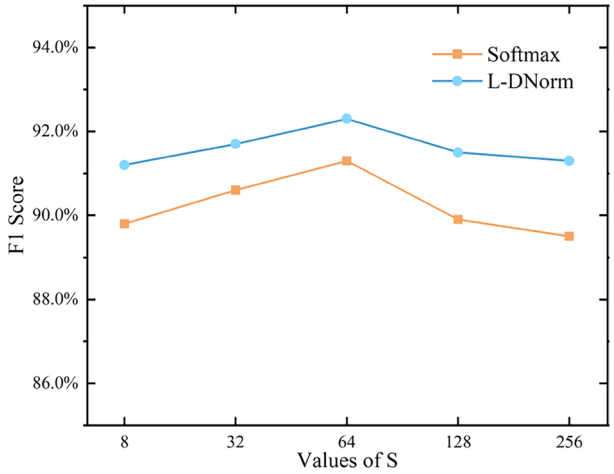

Fig. 5. Effect of parameter *S* on model performance.

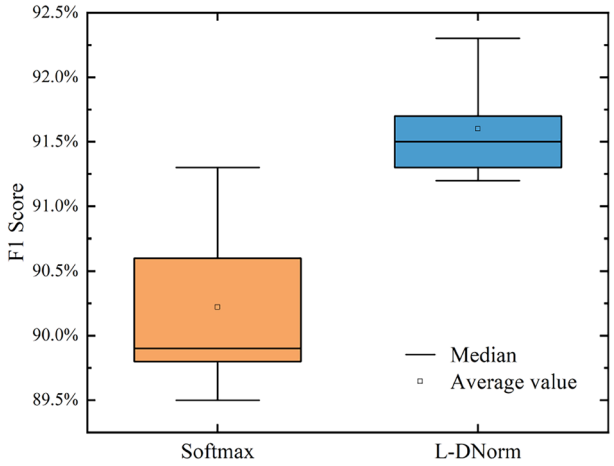

Fig. 6. Performance distribution of different normalization strategies.

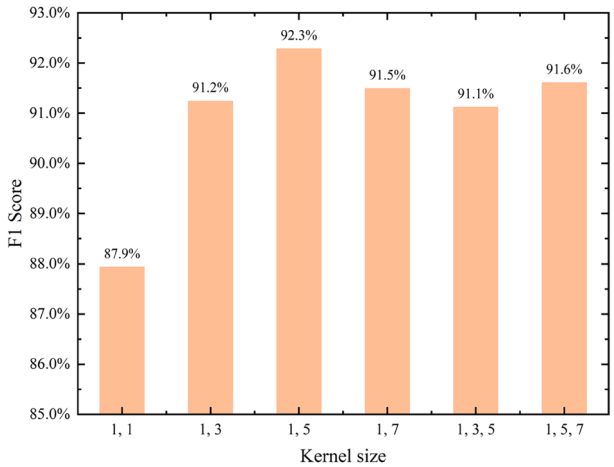

Fig. 7. Impact of different kernel size combinations on model performance.

(1,5), which yields an F1 score of 92.3%. However, further increasing the kernel size to 7 (1,7) leads to a slight performance degradation, with the F1 score dropping to 91.5%. This degradation can be attributed to two main factors: (1) larger kernels tend to introduce more noise from distant context that may not be relevant to entity boundary detection, and (2) the increased number of parameters makes the model more prone to overfitting, potentially compromising its ability to generalize well on unseen data.

We also investigated more complex configurations by incorporating three parallel convolution layers. The combination (1,3,5) achieves an F1 score of 91.1%, while (1,5,7) performs slightly better at 91.6%. However, neither of these triple-layer configurations outperforms the optimal dual-layer setup of (1,5), indicating that simply adding more parallel convolution layers does not necessarily lead to better performance. This observation aligns with our analysis that excessive context information and model complexity may hinder rather than help in the fine-grained entity boundary detection task.

These results suggest that the proposed FIEM works best with a moderate kernel size combination, where the first layer focuses on local features (kernel size 1) and the second layer captures broader contextual information (kernel size 5). This configuration strikes an effective balance between local and global feature extraction, enabling the model to better identify entity boundaries while avoiding the negative effects of over-extensive receptive fields and excessive model complexity.

Visual analysis

Confusion matrix heatmaps

To assess the effectiveness of our model, we performed an extensive comparison with the baseline model, analyzing results for each entity type. Figure 8 illustrates the confusion matrix heatmaps for both models, where the vertical axis represents the ground truth entity categories and the horizontal axis indicates the model-predicted categories. The color gradient transitions from deep purple to yellow, corresponding to the progression from low to high prediction ratios, with each cell containing the corresponding prediction percentage. The diagonal elements quantify per-class prediction accuracy, while non-diagonal elements indicate cross-category misclassification proportions.

Analysis of the confusion matrices reveals that the proposed model demonstrates demonstrably higher recognition accuracy (values on the main diagonal) across all entity categories compared to the baseline model. In addition, it can be observed from the off-diagonal elements that the false recognition rate of the improved model is notably reduced.

Specifically, MS entities usually contain variable-length unstructured descriptions with high vocabulary diversity, making them the most difficult category to identify. By comparing Fig. 8a and b, we can see that our model significantly reduces the error rate for MS entities. This is mainly due to the global memory units in the PAGA mechanism, which can capture dataset-level patterns of recurring operational terms and provide global semantic support when the context of a single sentence is insufficient.

In addition, our model shows stronger discrimination for easily confused entity pairs, such as Sub and Line, as well as VL and Time. Sub and Line often appear together and have specific topological hierarchies. The relative positional embedding in PAGA effectively models the directional dependency between them, which reduces hierarchy inversion errors to some extent. For VL and Time, which both contain numeric characters, the FIEM module plays a key role: its multi-scale convolution kernels can capture character-level morphological features and distinguish between voltage level and time formats, which allows it to differentiate these two types of numeric entities at a fine-grained level.

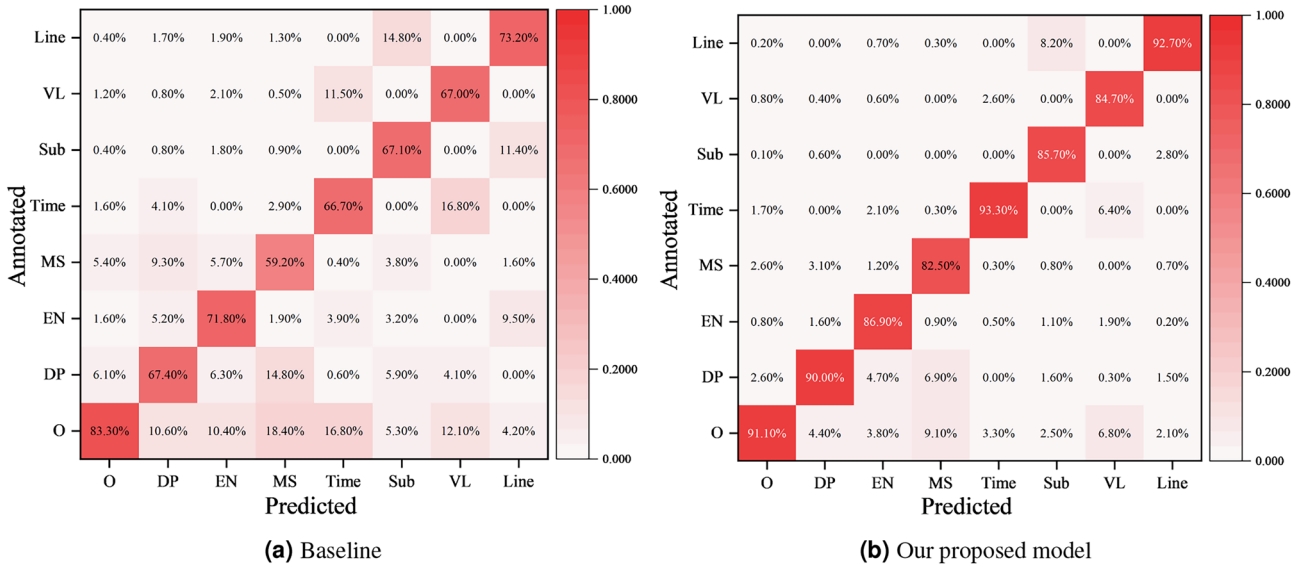

Fig. 8. Heatmaps of confusion matrices for model predictions.

Comparative analysis of prediction instances

To visually validate our proposed model's entity recognition capabilities in power equipment maintenance work order texts, particularly its performance in handling domain-specific terminology and complex entity structures, Table 6 presents a comparative analysis of three representative model prediction cases against the baseline model to evaluate the practical effectiveness of our proposed approach. The samples differ in sentence structure and composition style, with each conveying separate semantic meanings.

In Instance 1, the baseline model incorrectly identifies DP as part of EN and fails to accurately determine the boundaries of MS. This error shows the limitations of standard Transformer architectures in capturing fine-grained local dependencies, which often causes the boundaries between adjacent entities to become unclear in continuous Chinese text. In contrast, our model uses FIEM to explicitly capture character-level morphological features and local boundaries, which allows the model to clearly recognize that DP is a subcomponent attached to MS.

Instance 2 highlights the challenge of distinguishing structurally similar entities. Due to similar lexical patterns, the baseline model confuses Line with Sub. The relative positional embedding in our proposed PAGA allows the model to perceive the directional and distance relationships between key anchor points (such as the voltage level '110kV') and surrounding entities. By modeling these relative positions, the framework can understand that in such contexts, entities appearing before the voltage level usually refer to Line, while entities immediately after usually indicate Sub, which resolves the semantic ambiguity that causes confusion in standard attention mechanisms.

In Instance 3, the baseline model incorrectly labels the task code '23008' as a Time entity, likely misled by the surface feature that dates usually consist of numbers. This indicates that the baseline model fails to understand the global semantic distribution at the dataset level. Our framework captures dataset-scale features and statistical patterns through the global memory units integrated in the PAGA mechanism. Specifically, the model learns that the context in which task codes appear is clearly different from that of timestamps. By interacting with these global memory representations, our model suppresses false positive predictions based solely on numeric features and correctly identifies it as non-temporal information.

These results demonstrate our model's capability to precisely localize entities, contextual information, and boundary information, while understanding the dependencies between entities.

Overall performance comparison

To comprehensively evaluate the effectiveness and generalization capabilities of the proposed model across diverse languages, domains, and dataset scales, we conducted comparative experiments on three representative public datasets: CoNLL-2003, Resume, and the China People's Daily Corpus. We compare our model with other excellent contrast models on three datasets to verify its performance.

Table 7 presents the comparative performance between our model and other excellent contrast models across the three public datasets. On all datasets, our model achieved competitive results, performing on par with or exceeding the other models in most evaluation scenarios.

The experimental results demonstrate several notable performance characteristics:

Cross-lingual performance Our model achieves highly competitive F1 scores on both English (CoNLL-2003) and Chinese datasets, which confirms the generality of the PAGA mechanism when handling different language structures. English focuses on word order and syntactic structure, while Chinese focuses on semantic composition between characters. PAGA does not rely on language-specific grammar rules, but instead captures deep structural dependencies by modeling the relative distance and direction between tokens. This allows the

Instance 1	**Tang 110kV #1 main transformer measurement and control device communication failure inspection.**
Correct Label	[Tang]$_{Sub}$ [110kV]$_{VL}$ [#1 main transformer]$_{EN}$ [measurement and control device]$_{DP}$ [communication failure inspection]$_{MS}$.
Baseline	[Tang]$_{Sub}$ [110kV]$_{VL}$ [*#1 main transformer measurement and control device*]$_{EN}$*communication* [*failure inspection*]$_{MS}$.
Our Model	[Tang]$_{Sub}$ [110kV]$_{VL}$ [#1 main transformer]$_{EN}$ [measurement and control device]$_{DP}$ [communication failure inspection]$_{MS}$.
Instance 2	**Yunzhang second circuit 110kV Chang 04 isolation switch Phase C closing stop contact damage defect handling.**
Correct Label	[Yunzhang second circuit]$_{Line}$ [110kV]$_{VL}$ [Chang]$_{Sub}$ [04 isolation switch]$_{EN}$ Phase C [closing stop contact]$_{DP}$ [damage defect handling]$_{MS}$.
Baseline	[Yunzhang second circuit]$_{Sub}$ [110kV]$_{VL}$*Chang* [04 isolation switch]$_{EN}$ Phase C [*closing*]$_{DP}$*stop* [*contact damage defect handling*]$_{MS}$.
Our Model	[Yunzhang second circuit]$_{Line}$ [110kV]$_{VL}$ [Chang]$_{Sub}$ [04 isolation switch]$_{EN}$ Phase C [closing stop contact]$_{DP}$ [damage defect handling]$_{MS}$.
Instance 3	**Defect elimination task: Eliminate QX23008 Huashan substation 110kV bus coupler Hua 23 circuit breaker operating mechanism energy storage motor damage treatment.**
Correct Label	Defect elimination task: Eliminate QX23008 [Huashan substation]$_{Sub}$ [110kV]$_{VL}$ bus coupler [Hua]$_{Sub}$ [23 circuit breaker]$_{EN}$ [operating mechanism energy storage motor]$_{DP}$ [damage treatment]$_{MS}$.
Baseline	Defect elimination task: Eliminate QX[*23008*]$_{Time}$ [Huashan substation]$_{Sub}$ [110kV]$_{VL}$ [*bus coupler Hua*]$_{Sub}$ [23 circuit breaker]$_{EN}$*operating mechanism* [*energy storage motor*]$_{DP}$ [damage treatment]$_{MS}$.
Our Model	Defect elimination task: Eliminate QX23008 [Huashan substation]$_{Sub}$ [110kV]$_{VL}$ bus coupler [Hua]$_{Sub}$ [23 circuit breaker]$_{EN}$ [operating mechanism energy storage motor]$_{DP}$ [damage treatment]$_{MS}$.

Table 6. Assessing the prediction outcomes of the baseline model versus the proposed model. (Incorrect predictions are shown in italics).

CoNLL-2003				Resume				China People's Daily Corpus			
Model	**P(%)**	**R(%)**	**F1(%)**	**Model**	**P(%)**	**R(%)**	**F1(%)**	**Model**	**P(%)**	**R(%)**	**F1(%)**
Zhang et al.[45]	94.71	92.42	93.64	Guo et al.[46]	94.9	94.56	94.62	Zhang et al.[47]	93.23	92.42	92.82
Shah et al.[48]	93.13	**95.67**	95.3	Wang et al.[49]	–	–	96.53	Zhang et al.[45]	94.71	92.42	93.64
Chen et al.[50]	–	–	93.09	Mai et al.[51]	**96.91**	96.26	96.58	Wang et al.[52]	–	–	95.32
Fei et al.[53]	92.96	93.85	93.4	Chen et al.[54]	96.14	96.52	96.33	Liu et al.[55]	92.17	90.63	91.4
Chang et al.[56]	–	–	93.46	Zhang et al.[57]	96.09	96.44	96.26	Lv et al.[58]	97.12	96.11	96.61
Yu et al.[59]	–	–	93.42	Chen et al.[60]	96.54	96.41	96.58	Wang et al.[61]	–	–	95.62
Gan et al.[62]	93.84	93.6	93.72	Pan et al.[63]	–	–	96.35	Ke et al.[64]	95.93	96.45	96.19
Ours	**94.65**	94.44	**94.55**	Ours	96.54	**97.25**	**96.89**	Ours	**97.45**	**96.93**	**97.19**

Table 7. Performance comparison of different methods on multiple datasets. Bold values indicate the best results.

model to maintain sharp perception of entity boundaries when handling both English SVO structures and Chinese compact phrase structures.

Adaptation to entity density & local features The model's excellent performance on the Resume dataset is particularly noteworthy, as this dataset contains many short technical terms and abbreviations with very high entity density. This is highly similar to the characteristics of power equipment maintenance texts. This result strongly demonstrates the effectiveness of our designed FIEM. FIEM has local receptive field properties that can very efficiently capture local morphological features (such as abbreviation patterns) in these texts, which addresses the limitations of pure Transformer architectures when handling such high-density, short-span entities.

Robustness to dataset scale Although the data scales differ greatly (Resume $\approx$ 3.8k sentences, People's Daily $\approx$ 21k sentences), the model still maintains stable high performance. This robustness is largely due to the global memory units introduced in PAGA. On smaller datasets (such as Resume), these learnable memory units serve as anchor points for global prior knowledge, helping the model quickly identify dataset-specific distribution patterns, which reduces the risk of overfitting with small samples. On larger datasets (such as People's Daily), they effectively aggregate broad contextual information, which enhances the model's ability to distinguish long-tail entities.

These results indicate that our model not only performs well on power industry data but also achieves good results on public datasets, supporting our proposed model's generalization ability and robustness in handling various NER scenarios.

Conclusion

We propose a multi-level feature enhancement framework for automated parsing of critical operational data from power maintenance texts. By leveraging the operation and maintenance data of the State Grid, we established a comprehensive PEM-NER dataset using the proposed HKDC method and conducted a detailed analysis of its linguistic features. Based on these analyses, we propose a novel NER model specifically optimized for power equipment maintenance documentation.

Our model introduces several innovative architectural elements built upon the transformer framework. A key contribution is the Position-Aware Global Attention mechanism, which incorporates relative positional embedding between tokens during attention score computation and utilizes dual global memory units to capture dataset-scale feature representations. We further enhanced the model's capabilities through a Fine-Grained Information Enhancement Module, implementing an interactive convolutional neural network architecture that captures character-level local dependencies, complementing the attention mechanism where the latter captures global context while the former focuses on local features.

The model architecture integrates ELECTRA for character embedding to strengthen contextual representations and employs CRF for optimal label sequence determination. Empirical evaluation demonstrated strong performance, achieving a competitive F1 score of 92.3% on the PEM-NER dataset. The model's effectiveness was further validated through comparative analyses across three diverse public benchmark datasets, encompassing multiple languages and domains, thereby demonstrating its strong generalization capabilities.

Limitations

Despite the promising results, we acknowledge several limitations that provide directions for future research. First, since the PEM-NER dataset originates from a single organization, the model's robustness against varied terminologies and recording styles from other power utilities requires further validation. Second, our work is currently confined to NER; a crucial next step is to extend the framework to perform relation and event extraction to construct more comprehensive knowledge graphs of maintenance activities. Additionally, the model's ability to adapt to novel or unseen entity types is an open question, suggesting a need to explore few-shot or zero-shot learning techniques. Despite these limitations, this work holds promising implications for industrial applications by enhancing equipment management efficiency. The developed methodology offers an effective framework for information extraction tasks, with potential applicability across various engineering domains, thus contributing to the broader field of industrial text analytics.

Data availability

The datasets used and analysed during the current study available from the corresponding author on reasonable request.

Received: 9 September 2025; Accepted: 12 December 2025
Published online: 08 January 2026

References

1. Dileep, G. A survey on smart grid technologies and applications. *Renew. Energy* **146**, 2589–2625 (2020).
2. Lv, L., Wu, Z., Zhang, L., Gupta, B. B. & Tian, Z. An edge-ai based forecasting approach for improving smart microgrid efficiency. *IEEE Trans. Industr. Inf.* **18**, 7946–7954 (2022).
3. Li, J., Sun, A., Han, J. & Li, C. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* **34**, 50–70 (2020).
4. Yu, Z. et al. Construction of knowledge graph for gas polyethylene pipelines based on ALBERT-BiGRU-CRF. *Sci. Rep.* **15**, 25002 (2025).
5. Pan, J., Zhang, C., Wang, H. & Wu, Z. A comparative study of Chinese named entity recognition with different segment representations. *Appl. Intell.* **52**, 12457–12469 (2022).
6. Nasar, Z., Jaffry, S. W. & Malik, M. K. Named entity recognition and relation extraction: State-of-the-art. *ACM Comput. Surv. (CSUR)* **54**, 1–39 (2021).
7. Huang, Z., Xu, W. & Yu, K. Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015).
8. Chiu, J. P. & Nichols, E. Named entity recognition with bidirectional LSTM-CNNS. *Trans. Assoc. Comput. Linguist.* **4**, 357–370 (2016).
9. Yan, H., Deng, B., Li, X. & Qiu, X. Tener: adapting transformer encoder for named entity recognition. arXiv preprint arXiv:1911.04474 (2019).
10. Liu, Y., Huang, S., Li, R., Yan, N. & Du, Z. USAF: Multimodal Chinese named entity recognition using synthesized acoustic features. *Inf. Process. & Manag.* **60**, 103290 (2023).
11. Yin, S., Zhong, H., Huang, W. & Zhang, W. Deep learning enabled design of terahertz high-Q metamaterials. *Opt. & Laser Technol.* **181**, 111684 (2025).
12. Huang, W., Wei, Z., Tan, B., Yin, S. & Zhang, W. Inverse engineering of electromagnetically induced transparency in terahertz metamaterial via deep learning. *J. Phys. D Appl. Phys.* **54**, 135102 (2021).
13. Huang, R., Li, M., Zheng, H. & Zhao, Z. Chinese crop diseases and pests named entity recognition based on variational information bottleneck and feature enhancement: R. Huang, M. Li et al. *Sci. Rep.* **15**, 31573 (2025).
14. Ahmed, S. F. et al. Unveiling the frontiers of deep learning: Innovations shaping diverse domains. *Appl. Intell.* **55**, 1–55 (2025).
15. Yu, Y. et al. Chinese mineral named entity recognition based on Bert model. *Expert Syst. Appl.* **206**, 117727 (2022).
16. Liu, C. & Yang, S. A text mining-based approach for understanding Chinese railway incidents caused by electromagnetic interference. *Eng. Appl. Artif. Intell.* **117**, 105598 (2023).
17. Nath, N., Lee, S.-H. & Lee, I. Near: Named entity and attribute recognition of clinical concepts. *J. Biomed. Inform.* **130**, 104092 (2022).
18. Chu, J., Liu, Y., Yue, Q., Zheng, Z. & Han, X. Named entity recognition in aerospace based on multi-feature fusion transformer. *Sci. Rep.* **14**, 827 (2024).
19. Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **76**, 378 (1971).
20. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977).
21. Clark, K., Luong, M.-T., Le, Q. V. & Manning, C. D. Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020).
22. Vaswani, A. Attention is all you need. *Adv. Neural. Inf. Process. Syst.* **30**, 1 (2017).
23. Li, J., Tang, T., Zhao, W. X., Nie, J.-Y. & Wen, J.-R. Pre-trained language models for text generation: A survey. *ACM Comput. Surv.* **56**, 1–39 (2024).
24. Su, J. et al. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* **568**, 127063 (2024).
25. Geng, R., Chen, Y., Huang, R., Qin, Y. & Zheng, Q. Planarized sentence representation for nested named entity recognition. *Inf. Process. & Manag.* **60**, 103352 (2023).
26. Church, K. W. Word2vec. *Nat. Lang. Eng.* **23**, 155–162 (2017).
27. Goyal, A., Gupta, V. & Kumar, M. A deep learning-based bilingual Hindi and Punjabi named entity recognition system using enhanced word embeddings. *Knowl.-Based Syst.* **234**, 107601 (2021).
28. Rogers, A., Kovaleva, O. & Rumshisky, A. A primer in Bertology: What we know about how Bert works. *Trans. Assoc. Comput. Linguist.* **8**, 842–866 (2021).
29. Ke, G., He, D. & Liu, T.-Y. Rethinking positional encoding in language pre-training. arXiv preprint arXiv:2006.15595 (2020).
30. Niu, Z., Zhong, G. & Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **452**, 48–62 (2021).
31. Shang, F. & Ran, C. An entity recognition model based on deep learning fusion of text feature. *Inf. Process. & Manag.* **59**, 102841 (2022).
32. Lafferty, J., McCallum, A., Pereira, F. *et al.* Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, vol. 1, 3 (Williamstown, MA, 2001).
33. Sang, E. F. & De Meulder, F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050 (2003).
34. Zhang, Y. & Yang, J. Chinese NER using lattice LSTM. arXiv preprint arXiv:1805.02023 (2018).
35. Cui, H., Zhang, L., Wu, W. & Peng, Y. A two-layer bilstm model with linear gating for Chinese named entity recognition. In *2023 International Joint Conference on Neural Networks (IJCNN)*, 1–8 (IEEE, 2023).
36. Luo, L. et al. An attention-based BILSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics* **34**, 1381–1388 (2018).
37. Strubell, E., Verga, P., Belanger, D. & McCallum, A. Fast and accurate entity recognition with iterated dilated convolutions. arXiv preprint arXiv:1702.02098 (2017).
38. Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
39. Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019).
40. Joshi, M. et al. Spanbert: Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguist.* **8**, 64–77 (2020).
41. Beltagy, I., Peters, M. E. & Cohan, A. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150 (2020).
42. He, P., Liu, X., Gao, J. & Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654 (2020).

43. Warner, B. *et al.* Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2526–2547 (2025).
44. Li, Y., Yu, Y. & Qian, T. Type-aware decomposed framework for few-shot named entity recognition. arXiv preprint arXiv:2302.06397 (2023).
45. Runmei, Z. et al. Chinese named entity recognition method combining albert and a local adversarial training and adding attention mechanism. *Int. J. Semant. Web Inf. Syst. (IJSWIS)* **18**, 1–20 (2022).
46. Guo, X. et al. Chinese agricultural diseases and pests named entity recognition with multi-scale local context features and self-attention mechanism. *Comput. Electron. Agric.* **179**, 105830 (2020).
47. Zhang, L. et al. Lexicon and attention-based named entity recognition for kiwifruit diseases and pests: A deep learning approach. *Front. Plant Sci.* **13**, 1053449 (2022).
48. Shah, S. A. A., Masood, M. A. & Yasin, A. Dark web: E-commerce information extraction based on name entity recognition using bidirectional-lstm. *IEEE Access* **10**, 99633–99645 (2022).
49. Wang, Y., Lu, L., Wu, Y. & Chen, Y. Polymorphic graph attention network for Chinese NER. *Expert Syst. Appl.* **203**, 117467 (2022).
50. Chen, Y. et al. Semi-supervised named entity recognition in multi-level contexts. *Neurocomputing* **520**, 194–204 (2023).
51. Mai, C. et al. Pronounce differently, mean differently: A multi-tagging-scheme learning method for Chinese NER integrated with lexicon and phonetic features. *Inf. Process. & Manag.* **59**, 103041 (2022).
52. Wang, Z., Liu, H., Liu, F. & Gao, D. Why KDAC? A general activation function for knowledge discovery. *Neurocomputing* **501**, 343–358 (2022).
53. Fei, Y. & Xu, X. GFMRC: A machine reading comprehension model for named entity recognition. *Pattern Recogn. Lett.* **172**, 97–105 (2023).
54. Chen, J., Xi, X., Sheng, V. S. & Cui, Z. Randomly wired graph neural network for Chinese NER. *Expert Syst. Appl.* **227**, 120245 (2023).
55. Liu, J. et al. DAE-NER: Dual-channel attention enhancement for Chinese named entity recognition. *Comput. Speech & Lang.* **85**, 101581 (2023).
56. Chang, J. & Han, X. Multi-level context features extraction for named entity recognition. *Comput. Speech & Lang.* **77**, 101412 (2023).
57. Zhang, B., Cai, J., Zhang, H. & Shang, J. Visphone: Chinese named entity recognition model enhanced by visual and phonetic features. *Inf. Process. & Manag.* **60**, 103314 (2023).
58. Lv, Y., Qin, X., Du, X. & Qiu, S. Deep adaptation of CNN in Chinese named entity recognition. *Eng. Rep.* **5**, e12614 (2023).
59. Yu, Y. et al. Exploiting global contextual information for document-level named entity recognition. *Knowl.-Based Syst.* **284**, 111266 (2024).
60. Chen, Y., Fan, Q., Yuan, X., Zhang, Q. & Dong, Y. Pgd-gp: A Chinese named entity recognition model for constructing food safety standard knowledge graph. *IEEE Trans. Multimed.* (2024).
61. Wang, C.-H. et al. Multisource accident datasets-driven deep learning-based traffic accident portrait for accident reasoning. *J. Adv. Transp.* **2024**, 8831914 (2024).
62. Gan, Y. et al. Optimizing boundary dynamics for nested named entity recognition via semantic refinement and trimming. *Neural Netw.* **196**, 108218 (2025).
63. Pan, X.-Q., Feng, Z.-Q., Lu, Y. & Zhao, L.-F. LGENER: A lattice-and GAN-based method for Chinese ethnic NER. *Alex. Eng. J.* **115**, 297–307 (2025).
64. Ke, X., Wu, X., Ou, Z. & Li, B. Chinese named entity recognition method based on multi-feature fusion and biaffine. *Complex & Intell. Syst.* **10**(5), 6305–6318 (2024).

Author contributions

Conceptualization, Z.W.; methodology, Z.W.; software, Z.W. and H.G.; validation, S.Q.; formal analysis, Z.W. and H.G.; investigation, H.G., L.Z., Q.S. and C.Z.; data curation, L.Z., Q.S. and C.Z.; writing–original draft, Z.W. and S.Q.; supervision, S.Q.; project administration, S.Q. All authors reviewed the manuscript.

Funding

This research was funded by the key research and development program of Hubei Province (Grant No. 2023BAB049) and self-determined research funds of CCNU from the colleges' basic research and operation of MOE (Grant No. CCNU25ai025 & 30106250004).

Declarations

Competing interests

The authors declare that they have no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.Q.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.