energies

MDPI

*Article*

# A Named Entity Recognition Method for GIS Defect Texts Incorporating an Engineering Format-Aware Masking Strategy

**Ya Wu [1,*], Cuiru Yang [1], Yao Yao [1] and Jian Lu [2]**

[1] Guangdong Power Grid Electric Power Research Institute, Guangzhou 510080, China
[2] China Southern Power Grid Internet Services Co., Ltd., Guangzhou 510699, China
* Correspondence: 18592025013@163.com

**Abstract**

Named entity recognition (NER) is a key technique for extracting entities such as equipment, components and defect types from GIS defect texts, providing a basis for subsequent knowledge graph construction. However, GIS defect texts contain many engineering structures, including engineering abbreviations, equipment numbers and phase identifiers, making it difficult for general-purpose models to stably recognize their semantic associations and entity boundaries. To address this problem, this paper proposes an engineering format-aware masking strategy. The strategy identifies candidate fragments using format rules for phase identifiers, measurement value-unit patterns and engineering abbreviations and preferentially selects them as perturbation targets to strengthen the model's understanding of engineering structures and their contextual relationships. Bidirectional long short-term memory is used to extract bidirectional sequence features, and a conditional random field is used to model transition constraints between labels and obtain the globally optimal label sequence. The results show that the proposed model achieves precision, recall and $F_1$ scores of 0.89, 0.92 and 0.90, respectively. The analysis indicates that the proposed method improves entity recognition for phase-related structures, engineering abbreviations and equipment hierarchy fragments.

**Keywords:** GIS; named entity recognition; defect text; natural language processing; knowledge graph; engineering format-aware masking

## 1. Introduction

Gas-insulated switchgear (GIS) has a compact footprint, high operational reliability and strong environmental adaptability. It has therefore been widely used in power grids at different voltage levels [1–4]. With the continuous expansion of power grid assets and the increasing service age of equipment, a large volume of GIS defect records has accumulated during inspection, maintenance, defect handling and fault elimination. These records usually contain information on voltage level, defective equipment, defective components, defect manifestations and treatment processes. They are an important data source for analyzing GIS operating conditions, mining typical defect patterns and supporting quality evaluation. However, compared with structured data, defect descriptions are mostly recorded manually by field personnel. They often show inconsistent wording, dense technical terminology, frequent abbreviations and informal expressions. As a result, the key information embedded in these records is difficult to count and use directly [5–9]. Therefore, text mining for GIS defect records is important for improving defect analysis efficiency and supporting subsequent equipment quality evaluation.

Check for updates

Academic Editor: Krzysztof Górecki

Received: 10 July 2026
Revised: 17 August 2026
Accepted: 24 August 2026
Published: 26 August 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Beyond these text-level considerations, the increasing penetration of power-electronic interfaces in smart distribution systems has made reliable equipment operation and maintenance increasingly important for grid stability [10]. Although power-electronic converters facilitate renewable-energy integration and improve system controllability, they also increase the complexity of power-quality management. In particular, harmonic distortion and uncertain grid impedance can degrade the performance of grid-connected converters, creating a need for robust control strategies [11]. In this context, timely extraction of equipment, component and defect information from GIS maintenance records supports condition assessment and maintenance decision-making. GIS NER is, therefore, not merely a text-processing task but an enabling technique for maintaining the reliability and operational performance of increasingly complex power infrastructures.

Knowledge graphs have been widely used in text mining for power equipment. By organizing unstructured information through entities, relations and attributes, they enable structured representation and associative reasoning over domain knowledge such as equipment components, fault causes and maintenance measures. They therefore play an important role in power equipment defect analysis, operation and maintenance knowledge management and decision support [12,13]. Entity extraction identifies specific types of entities from unstructured text, such as equipment names and voltage levels. Named entity recognition (NER) is a key technique for entity extraction, and its performance directly affects the accuracy of subsequent knowledge graph construction [14–17].

Early NER techniques mainly relied on professional dictionaries, rule templates and traditional machine learning methods [18,19]. These methods are highly interpretable, but their recognition performance depends on manually constructed dictionaries and features. They are therefore less adaptable to out-of-vocabulary terms, synonymous expressions and non-standard records. In recent years, pretrained language models represented by Bidirectional Encoder Representations from Transformers (BERT) have provided a new direction for NER [20]. Xie et al. improved BERT and extracted operation and maintenance entities for power transformers based on pretraining with A Lite BERT (ALBERT). This work further supported the construction of a power transformer operation and maintenance knowledge graph [21]. Dai et al. used Masked Language Model as Correction BERT (MacBERT) as the pretrained model and integrated bidirectional long short-term memory (BiLSTM) and conditional random field (CRF) to construct a MacBERT-BiLSTM-CRF model. The model achieved effective entity extraction for knowledge graph construction in relay protection devices [22]. Jia et al. developed PowerBERT using large-scale power equipment text data to improve the semantic representation of power-domain texts [23]. Compared with general-purpose pretrained models such as BERT and MacBERT, PowerBERT focuses on learning domain-specific terminology and semantic information from power equipment texts.

However, existing entity recognition methods mainly perform pretraining around general characters, words or entity fragments. They lack targeted modeling of the engineering format structures specific to GIS defect texts. GIS defect texts contain many fragments that mix digits, letters and Chinese characters, such as 110 kV, 5012 switch and ABC three-phase labels. During random masking, whole-word masking or general N-gram masking, the units within the same engineering fragment are difficult to select jointly and stably as perturbation targets. This limits the model's ability to learn their internal compositional relations and entity boundaries.

To address this problem, this paper proposes a named entity recognition model incorporating an engineering format-aware masking strategy. Based on the MacBERT pretraining framework, the proposed strategy increases the probability that engineering-format fragments in GIS defect texts are selected as perturbation targets. This allows the model

to learn engineering structures and their contextual relations while recovering the original text. The pretrained encoder is then combined with BiLSTM and CRF to achieve more accurate GIS entity recognition. The effectiveness of the proposed engineering format-aware masking strategy is verified through comparative experiments on GIS defect records.

## 2. GIS Defect Text Preprocessing

### *2.1. Text Cleaning*

This study focuses on GIS defect description texts, with representative examples shown in Table 1. Because field defect records are often informal, they may contain redundant spaces, inconsistent letter case and abnormal line breaks. Text cleaning was therefore performed before model training.

**Table 1.** Examples of GIS fault texts.

| Defect Level | Manufacture | Defect Type | Defect Description |
|---|---|---|---|
| General | A | $SF_6$ Gas Leakage | #2 main transformer high-voltage 102 circuit breaker has insufficient $SF_6$ gas pressure. The on-site pressure is 0.605 MPa, while the rated value is 0.6 MPa. |
| General | B | Leakage | The mechanical box current switching circuit is leaking oil, and oil leakage is observed along the lower edge of the box. |
| Emergency | C | Component Failure | Internal failure of the $SF_6$ density relay (pressure gauge), causing the pointer to remain in the alarm range. |

First, stop words were removed. In this study, stop words refer to words and symbols that do not provide useful information for defect analysis, including punctuation, modal particles, personal names and place names. The HIT stop-word list was adopted [24], and the Jieba word segmentation tool was used to segment the defect texts and remove the stop words.

Second, abnormal values and redundant marks were removed. Line breaks, tabs and extra spaces were cleaned from the original defect descriptions. In addition, some records used square brackets to mark the responsible maintenance team before the complete defect description. Because this information was not directly relevant to defect analysis, the bracketed text was removed. The specific processing criteria are shown in Table 2.

**Table 2.** Text preprocessing rules for GIS defect records.

| Processing Object | Criterion | Processing Operation |
|---|---|---|
| Line breaks, tabs, and whitespace | \r, \n, \t, and other whitespace characters | Removed |
| Full-width symbols | Predefined character mappings | Normalized to standard forms |
| Abnormal decimal separators | 。 or ． occurring between digits | Replaced with. |
| Domain-specific abbreviations | Variants in letter case or character width | Convert to a unified form |

### *2.2. Text Annotation*

After text cleaning, the defect records were annotated to construct the training set for the NER model. This study used the BIO annotation scheme. In this scheme, O denotes a non-entity token, DEV denotes defective equipment, COMP denotes a defective component, PHEN denotes a defect phenomenon, and VOLT denotes a voltage level.

Actual defect records may contain multiple defective components and defect phenomena. These cases were annotated separately to reduce the difficulty of entity extraction and to avoid overfitting caused by excessively long entities. The dataset was annotated by two annotators following the same annotation criteria. Cohen's Kappa calculated based on the BIO annotation labels was 0.86, indicating a high level of annotation agreement. A total of 2854 GIS defect records were manually annotated for subsequent model training. The annotated texts had an average sentence length of 41.2 characters. The resulting corpus contained 1730 DEV, 6919 COMP, 5113 PHEN, and 1379 VOLT instances.

## 3. Model Pretraining

The pretraining architecture of the proposed model is shown in Figure 1. First, in the vector representation layer, the input text is converted into a combined representation of character embeddings, position embeddings and segment embeddings. The resulting matrix is then passed to the Transformer encoding layer. The encoding layer is formed by stacking multiple encoder modules from the standard Transformer architecture. In each module, the multi-head attention layer captures the internal semantics of the input text, while the residual connections, normalization layers and fully connected layers integrate and normalize feature weights within the modules. After processing by multiple Transformer layers, the model outputs an encoding matrix for downstream training.

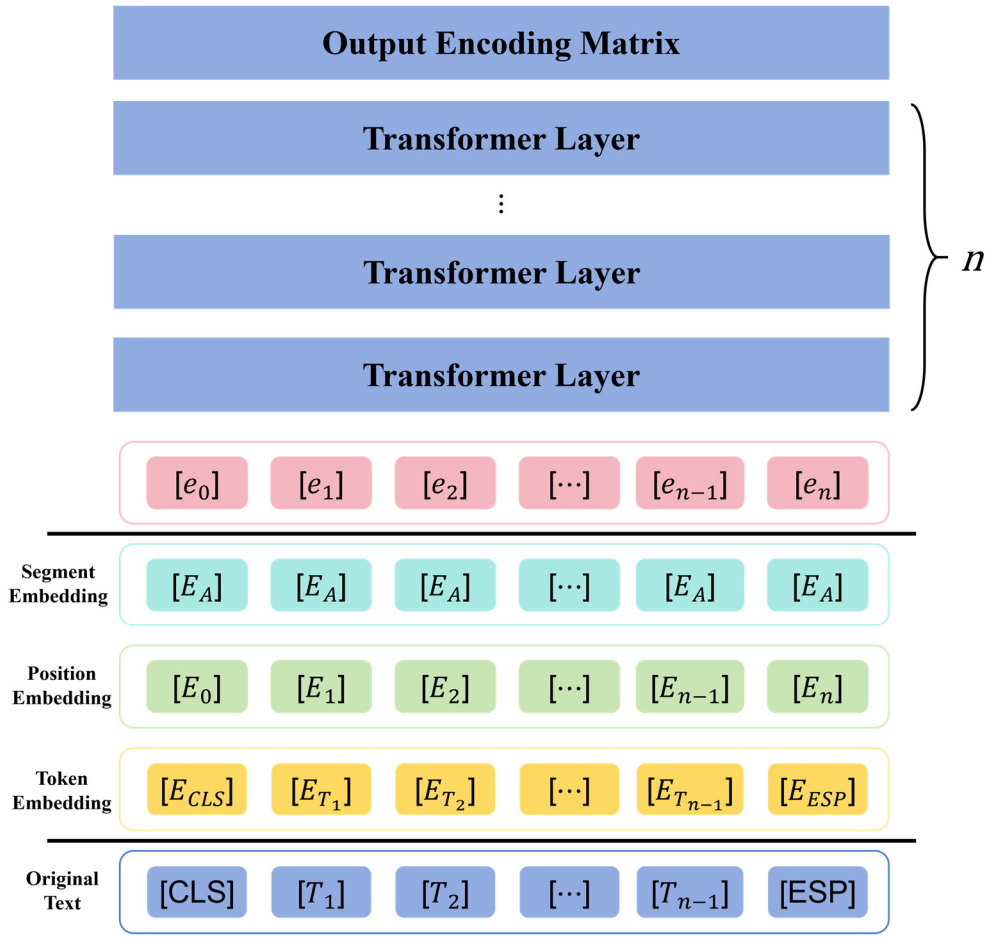



**Figure 1.** Model pretraining architecture.

### 3.1. Vector Representation Layer

The vector representation layer converts the input GIS defect text into an input vector matrix for the Transformer encoding layer. The input GIS defect text is represented as vector $\boldsymbol{X}$:

$$\boldsymbol{X} = (x_1, x_2, \cdots, x_n) \tag{1}$$

where $x_i$ denotes the $i$th character or token unit in the input sequence, and $n$ denotes the length of the input sequence.

Each input unit is mapped to a continuous vector representation composed of a character embedding, a position embedding and a segment embedding:

$$\boldsymbol{e}_i = \boldsymbol{E}_{\text{token}}(x_i) + \boldsymbol{E}_{\text{pos}}(i) + \boldsymbol{E}_{\text{seg}}(s_i) \tag{2}$$

where $\boldsymbol{e}_i$ denotes the initial vector of the $i$th input unit. $\boldsymbol{E}_{\text{token}}(x_i)$ denotes the embedding vector of the character or token unit, which represents its basic semantic information. $\boldsymbol{E}_{\text{pos}}(i)$ denotes the position embedding, which preserves the order information of the character in the sentence. $\boldsymbol{E}_{\text{seg}}(s_i)$ denotes the segment embedding, which distinguishes different sentences or text fragments.

Given $N$ input texts, the vector representations of all input texts are combined to obtain the final input matrix $\boldsymbol{E}$.

$$\boldsymbol{E} = (\boldsymbol{e_1}, \boldsymbol{e_2}, \cdots, \boldsymbol{e_N})^{\mathrm{T}} \tag{3}$$

### 3.2. Transformer Encoder Layer

The architecture of the multi-head attention mechanism in the Transformer encoding layer is shown in Figure 2. After linear transformation, the input matrix $\boldsymbol{E}$ is converted into a query matrix $\boldsymbol{Q}$, a key matrix $\boldsymbol{K}$ and a value matrix $\boldsymbol{V}$. The attention weights are then obtained through scaled dot-product attention. In the multi-head attention mechanism, each attention head generates its own attention weights. These weights are concatenated in the concatenation layer and then passed through a linear transformation layer. The concatenated result is finally mapped back to the dimensionality required by the model for output.

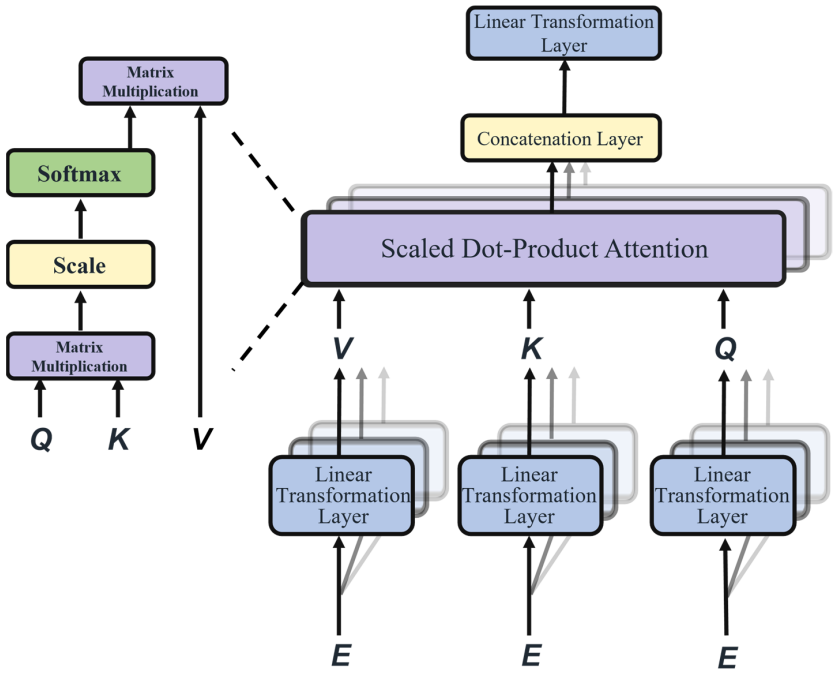



**Figure 2.** Architecture of the multi-head attention mechanism.

## 4. Pretraining Procedure

The advantage of BERT-based pretrained language models lies in their ability to learn word meanings, syntactic structures and contextual relations from large-scale unlabeled corpora before being fine-tuned on downstream tasks with a small number of labeled samples. In practice, the model learns linguistic patterns and domain-specific expressions through pretraining tasks. Therefore, the pretraining strategy has a substantial influence on the final model performance.

This paper proposes an engineering format-aware masking strategy for GIS defect texts. Instead of selecting masking positions randomly, the strategy incorporates engineering description fragments commonly found in GIS defect records, such as equipment and component expressions. Text units with domain semantics and structural features are preferentially selected for masking, allowing the model to learn the expression patterns of GIS defect texts. This strengthens the model's representation of key engineering fragments and defect entities, thereby improving its prediction accuracy for the corresponding entities.

### *4.1. Engineering Format-Aware Masking Strategy*

Based on the conventional BERT framework, MacBERT replaces random masking with a correction-style masking strategy to reduce the input distribution gap between pretraining and downstream tasks. It also introduces whole-word masking and N-gram masking, thereby improving the model's ability to understand Chinese vocabulary.

The existing masking strategies in MacBERT perform well on general Chinese texts, but they are less effective at recognizing engineering structures in GIS defect texts, such as the 5042 disconnector, 500 kV and SF6. These text fragments are difficult for the tokenizer to segment correctly, which makes it difficult to mask them as complete units in the subsequent masking process. To address this problem, this paper proposes an engineering format-aware masking strategy. The strategy aims to solve the problem that general tokenization and random masking cannot preserve complete engineering entity boundaries, causing key engineering fragments to be split. The specific procedure is as follows.

For an input raw defect text, the original text sequence $\boldsymbol{X}$ is obtained after segmentation by the original MacBERT tokenizer:

$$\boldsymbol{X} = (x_1, x_2, \cdots, x_n) \tag{4}$$

During continued pretraining, a subset of positions must be selected from the original text as positions to be masked. In this paper, the text at these positions is referred to as perturbation targets. During training, the perturbation targets are replaced with synonyms or random words, and the model is then forced to predict the original text. This process improves the model's understanding of the perturbation targets and their contexts.

Let the perturbed input sequence be

$$\widetilde{X} = (\widetilde{x}_1, \widetilde{x}_2, \cdots, \widetilde{x}_n) \tag{5}$$

If the $i$th position is not selected as a perturbation target, then it remains unchanged; otherwise, it is replaced. The final sequence fed into the model is

$$X_{in} = [\mathrm{CLS}]x_1, \widetilde{x}_2, \cdots, \widetilde{x}_n[\mathrm{SEP}] \tag{6}$$

The core of the proposed engineering format-aware masking strategy is to determine the set of perturbation targets. Traditional masking strategies randomly select perturba-

tion targets from the entire text sequence, whereas the proposed strategy masks the corresponding engineering-format fragments. Let the engineering-format set be

$$R = \left\{ R_{pha}, R_{mea}, R_{acr}, R_{loc} \right\} \tag{7}$$

The construction and boundary rules for these four format sets are summarized in Table 3. The engineering-format sets were constructed from recurrent patterns observed in the GIS training texts. Representative fragments were first collected and grouped into four categories: phase formats, measurement value–unit formats, technical abbreviation formats, and equipment and location formats. The variable elements in each category were then replaced with reusable slots. For example, "phase A" and "phase B" were generalized as "phase identifier + phase marker", whereas "0.45 MPa" and "59.5 °C" were generalized as "numerical value + engineering unit". These templates were implemented as regular-expression or lexicon-matching rules and checked against the training texts to remove evident false matches and add missed patterns. The finalized rules were fixed before evaluation and used to extract the character spans of candidate engineering-format fragments from each normalized input text. More generally, the proposed strategy treats recurrent semi-structured domain expressions as complete perturbation units, and the corresponding format lexicon and matching templates can be adapted to other technical domains.

**Table 3.** Construction rules for the engineering-format sets.

| Format | Construction Rule | Examples |
|---|---|---|
| $R_{pha}$ | Phase identifiers A, B, and C followed by a phase marker; isolated letters are excluded. | phase A; phases A/B/C |
| $R_{mea}$ | An integer or decimal value combined with a predefined engineering unit and treated as a complete fragment. | 0.45 MPa; 59.5 °C; 220 kV |
| $R_{acr}$ | Complete technical abbreviations and their written variants defined in a domain abbreviation lexicon. | SF6; PT; CT |
| $R_{loc}$ | Combinations of equipment identifiers, equipment terms, phase or side markers, and location expressions, bounded by punctuation. | 5022 disconnector; Phase A CT lower flange |

Based on the set $R$, the set of engineering-format fragments extracted from the input text is

$$S = \{s_1, s_2, \cdots, s_k\} \tag{8}$$

where the $j$th candidate format fragment is

$$s_j = (l_j, r_j, c_j, w_j) \tag{9}$$

Here, $l_j$ and $r_j$ denote the start and end positions of the fragment in the original text, respectively. $c_j$ denotes the engineering-format type to which the fragment belongs, and $w_j$ denotes the sampling weight of fragments of this type.

MacBERT cannot directly process raw strings. The text must first be segmented by the tokenizer and converted into sequence units that can be recognized by the model. In this paper, the basic sequence units output by the tokenizer are referred to as tokens. For Chinese text, a token usually corresponds to one Chinese character. For digits, letters and technical abbreviations, an engineering-format fragment may be segmented into one or more tokens. The tokenizer also records the start and end character positions of each token in the original text, thereby establishing the correspondence between original text fragments and model input positions.

Suppose the input sequence contains $n$ valid tokens. Let the overall perturbation ratio be $\rho$, and let the engineering-format perturbation ratio be $\lambda$. Here, $\rho$ denotes the proportion of tokens in the input sequence selected as perturbation targets, while $\lambda$ denotes the proportion preferentially selected from the engineering-format set. Therefore, the number of engineering-format perturbation targets is defined as

$$N_R = \rho \lambda n \tag{10}$$

For fragment $s_j$, its character span $[l_j, r_j)$ is first determined. Then, combined with the position mapping returned by the tokenizer, $[o_i^{start}, o_i^{end})$, its corresponding set of token positions is obtained as

$$C(s_j) = \left\{ i \middle| [o_i^{start}, o_i^{end}) \cap [l_j, r_j) \neq \varnothing \right\} \tag{11}$$

where $o_i^{start}$ and $o_i^{end}$ denote the start and end positions of the $i$th token in the original text, respectively. $C(s_j)$ denotes all token positions covered by fragment $s_j$.

During masking, once an $s_j$ is selected, its corresponding token set $C(s_j)$ is added to the engineering-format perturbation target set $\mathrm{T}_R$. Sampling stops when $\mathrm{T}_R$ reaches the predefined number of engineering-format perturbation targets, or when no sampleable fragments remain in the engineering-format fragment set $S$. If the number of perturbation targets in $\mathrm{T}_R$ is smaller than the overall number of perturbation targets after engineering-format sampling, text fragments are randomly selected from non-engineering-format fragments using N-gram masking. These fragments form the supplementary perturbation target set $\mathrm{T}_N$. The final perturbation target set is, therefore,

$$\mathrm{T} = \mathrm{T}_R \cup \mathrm{T}_N \tag{12}$$

For the perturbation targets in set T, text replacement is performed at a ratio of 8:1:1, as shown in Equation (13):

$$\widetilde{x}_i = \begin{cases} g(x_i), & p = 0.8 \\ r(x_i), & p = 0.1, \\ x_i, & p = 0.1 \end{cases} \qquad i \in \mathrm{T} \tag{13}$$

where $g(x_i)$ denotes replacement with an engineering fragment of the same type, $r(x_i)$ denotes random replacement, and $x_i$ denotes that the original text remains unchanged. To preserve sequence length, $g(x_i)$ accepts only same-type candidates containing the same number of MacBERT tokens as $x_i$; otherwise, another candidate is sampled, or a random replacement of the same token length is used if none is available. Therefore, no tokens are inserted or deleted, and the batch tensors remain aligned. The model is then forced to recover the original text, allowing it to learn the linguistic features of the replaced text.

### *4.2. Pretraining Output and Loss Calculation*

After the perturbation set T is obtained, the objective of continued pretraining is to recover the original tokens at the positions in T before perturbation, based on their contextual information. The perturbed text passes through the vector representation layer and the multilayer Transformer encoder described in Section 3. Each input position then obtains a feature vector containing contextual semantics. In this paper, this vector is referred to as the hidden vector. Let the encoder with parameters $\theta$ be denoted as $\mathrm{Encoder}_\theta(\cdot)$. Its output is

$$\boldsymbol{H} = \mathrm{Encoder}_\theta(\boldsymbol{X}_{in}) = \{\boldsymbol{h_1}, \boldsymbol{h_2}, \cdots, \boldsymbol{h_M}\} \tag{14}$$

where $M$ denotes the length of the input sequence, and $\boldsymbol{h_i}$ denotes the hidden vector output by the last Transformer layer at the $i$th position. According to set T, the corresponding hidden vectors are extracted from the encoder output $\boldsymbol{H}$:

$$\boldsymbol{H}_{\mathrm{T}} = \{\boldsymbol{h}_{t_1}, \boldsymbol{h}_{t_2}, \cdots, \boldsymbol{h}_{t_m}\} \tag{15}$$

The extracted $\boldsymbol{H}_{\mathrm{T}}$ is input into the prediction layer. The prediction layer first applies a nonlinear transformation to the hidden vectors:

$$\boldsymbol{G}_{\mathrm{T}} = \mathrm{LayerNorm}[\mathrm{GELU}(\boldsymbol{H}_{\mathrm{T}}\boldsymbol{W}_{\mathrm{MLM}} + \boldsymbol{b}_{\mathrm{MLM}})] \tag{16}$$

where $\boldsymbol{W}_{\mathrm{MLM}}$ and $\boldsymbol{b}_{\mathrm{MLM}}$ denote the weight matrix and bias vector of the fully connected layer, respectively. GELU is the activation function, and LayerNorm denotes the normalization operation. The transformed $\boldsymbol{G}_{\mathrm{T}}$ is mapped to the original MacBERT vocabulary space, and Softmax is used to calculate the prediction probability of each candidate token:

$$\boldsymbol{P}_{\mathrm{T}} = \mathrm{Softmax}\left(\boldsymbol{G}_{\mathrm{T}}\boldsymbol{E}_{\mathrm{vocab}}^{\mathrm{T}} + \boldsymbol{b}_{\mathrm{vocab}}\right) \tag{17}$$

where $\boldsymbol{E}_{\mathrm{vocab}}^{\mathrm{T}}$ denotes the word embedding matrix, $\boldsymbol{b}_{\mathrm{vocab}}$ denotes the bias vector of the vocabulary mapping layer, and $\boldsymbol{P}_{\mathrm{T}}$ denotes the vocabulary probability distribution at the perturbation target positions. The $j$th row of matrix $\boldsymbol{P}_{\mathrm{T}}$ represents the probability that the model predicts each token in the vocabulary as the original token at position $t_j$. Let the original token at position $t_j$ before perturbation be $x_{t_j}$, and let its index in the vocabulary be $a_j$. The true outputs for all perturbation target positions are, therefore,

$$a_{\mathrm{T}} = \{a_1, a_2, \cdots, a_m\} \tag{18}$$

Cross-entropy is used to measure the difference between the predicted probabilities and the true tokens. The final loss function is

$$L_{\mathrm{MLM}} = -\frac{1}{m}\sum_{j=1}^{m} \log \boldsymbol{P}_{\mathrm{T}}(j, a_j) \tag{19}$$

where $\boldsymbol{P}_{\mathrm{T}}(j, a_j)$ denotes the probability that the model successfully predicts the original token $a_j$ at the $j$th perturbed position.

After forward propagation is completed for each batch, backpropagation is performed according to the loss function, and the parameters of the encoder and prediction layer are updated. The updated encoder is then used for the downstream named entity recognition task, where it is combined with BiLSTM and CRF to extract entity sequence features and decode labels.

## 5. Named Entity Recognition Task and Validation for GIS Defect Texts

### *5.1. Construction of the Named Entity Recognition Model*

To verify the representation ability of the engineering format-aware masking strategy for GIS defect texts, BiLSTM and CRF are connected after the pretrained encoder to construct a named entity recognition model for GIS defect texts. After encoding, the input text is converted into a semantic vector sequence $Z$. $Z$ is then fed into BiLSTM, which extracts sequence features in the forward and backward directions. The hidden states from the two directions are concatenated. The specific process is shown in Figure 3.

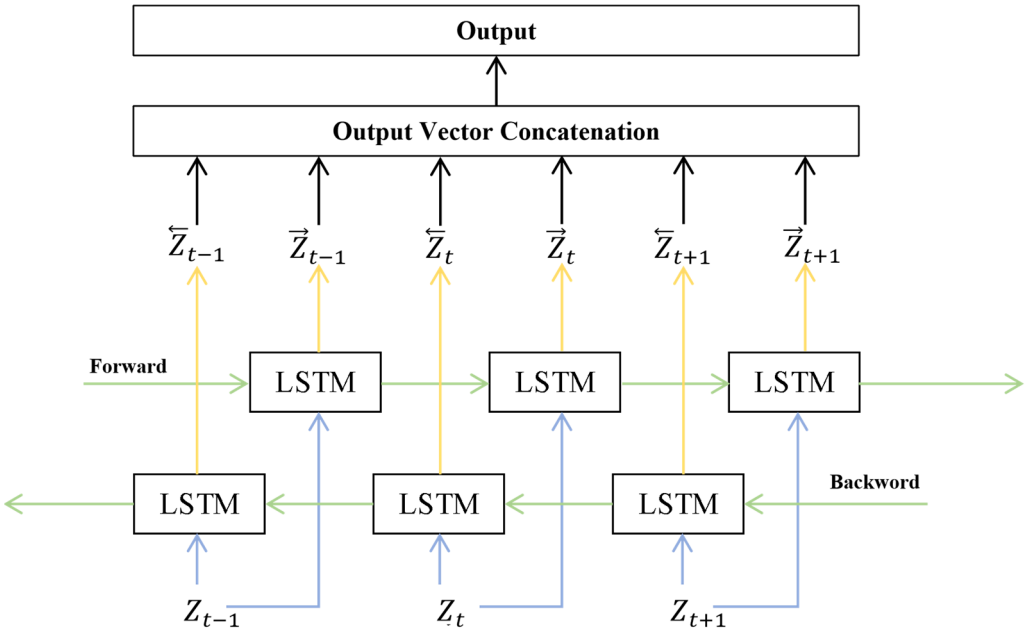



**Figure 3.** Architecture of the BiLSTM network.

The bidirectional sequence feature representation of each character is finally obtained as

$$\boldsymbol{u} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_n) \tag{20}$$

where $\boldsymbol{u}_i$ denotes the feature vector of the $i$th character after integrating contextual information.

The vector $\boldsymbol{u}_i$ is then converted into emission scores for different BIO labels through linear mapping. However, selecting labels at each position only according to emission scores may produce label combinations that do not satisfy BIO constraints. Therefore, the model further uses CRF to learn transition relations between adjacent labels. The structure of the CRF model is shown in Figure 4. After CRF modeling, the optimal label sequence can be obtained.

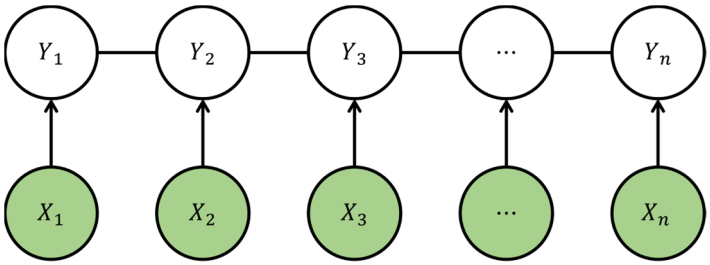



**Figure 4.** Architecture of the CRF model.

### *5.2. Experimental Results and Analysis*

Based on the GIS defect texts provided by China Southern Power Grid, 2854 records were obtained after text preprocessing and used for subsequent model training. The proposed named entity recognition model incorporating the engineering format-aware masking strategy was then used for entity extraction. The annotated defect texts were randomly divided into training, test and validation sets at a ratio of 8:1:1.

To verify the effectiveness of the proposed engineering format-aware masking strategy, the number of completely masked engineering-format text fragments of different types was counted during model pretraining. The proposed masking strategy and the original MacBERT masking strategy were compared. To reduce the influence of randomness,

the experiment was repeated 20 times. The average count was calculated and rounded down, and the results are shown in Table 4. As shown in the table, the proposed masking strategy produces a larger number of completely masked engineering-format fragments than the MacBERT masking strategy for all types of engineering-format texts.

**Table 4.** Number of complete masks for engineering-format text under different masking strategies.

| Text Type | MacBERT Masking Strategy | Engineering-Format-Aware Masking Strategy |
|---|---|---|
| Phase format | 127 | 316 |
| Measurement value-unit format | 40 | 82 |
| Technical abbreviation format | 108 | 459 |
| Equipment and location format | 55 | 120 |

To verify the overall entity extraction performance of the proposed model on GIS defect texts, comparative experiments were conducted using BERT-BiLSTM-CRF, MacBERT-BiLSTM-CRF and PowerBERT-BiLSTM-CRF models. Hyperparameters were selected on the validation set using 30 trials of Tree-structured Parzen Estimator-based search. The overall MLM probability, engineering-format mask ratio, encoder learning rate, and NER batch size were included in the search. To summarize their marginal influence, the completed trials were grouped according to each parameter, and the mean validation $F_1$ score within each group was calculated. The search results and selected settings are presented in Table 5.

**Table 5.** Hyperparameter search results and selected settings.

| Parameter | Search Range | Grouped Mean Validation $F_1$ (%) | Selected Value |
|---|---|---|---|
| Overall MLM probability | 0.08–0.20 | 0.08–0.12: 92.73; 0.12–0.16: 92.51; 0.16–0.20: 92.71 | 0.1733 |
| Engineering-format mask ratio | 0.30–0.80 | 0.30–0.45: 92.61; 0.45–0.60: 92.76; 0.60–0.80: 92.53 | 0.4832 |
| Encoder learning rate | $1 \times 10^{-5}$~$5 \times 10^{-5}$ | 1~1.5 × $10^{-5}$:92.77; 1.5 × $10^{-5}$~3 × $10^{-5}$:92.62; 3 × $10^{-5}$~5 × $10^{-5}$:92.27 | $1.1235 \times 10^{-5}$ |
| Batch size | 4, 8, 16 | 4:92.73; 8:92.49; 16:92.46 | 4 |
| Maximum NER epochs | Fixed | / | 30 |
| Number of LSTM hidden units | Fixed | / | 128 |

Precision $P$, recall $R$ and $F_1$ score were used as evaluation metrics. They are calculated as follows:

$$\begin{aligned} P &= \frac{N_{\text{correct}}}{N_{\text{predicted}}} \\ R &= \frac{N_{\text{correct}}}{N_{\text{true}}} \\ F_1 &= \frac{2PR}{P+R} \end{aligned} \tag{21}$$

where $N_{\text{correct}}$ denotes the number of correctly recognized entities, $N_{\text{predicted}}$ denotes the total number of recognized entities, and $N_{\text{true}}$ denotes the total number of annotated entities.

Precision represents the proportion of true entities among the results predicted as entities by the model. Recall represents the proportion of entities actually present in the dataset that are correctly recognized. The harmonic mean, $F_1$ score, represents the overall performance of the model. A higher $F_1$ score indicates better overall performance. The validation-set $F_1$ curves and test-set experimental results of different models are shown in Figure 5 and Table 6.

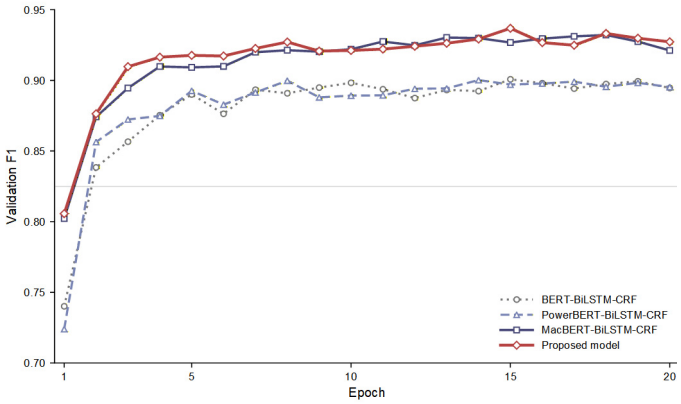



**Figure 5.** $F_1$ score curves of different models.

**Table 6.** Results of different models.

| Model Type | *P* | *R* | $F_1$ |
|---|---|---|---|
| BERT-BiLSTM-CRF | 0.86 | 0.89 | 0.87 |
| PowerBERT-BiLSTM-CRF | 0.87 | 0.89 | 0.88 |
| MacBERT-BiLSTM-CRF | 0.87 | 0.92 | 0.89 |
| The proposed model | 0.89 | 0.92 | 0.90 |

As shown in Figure 5, BERT-BiLSTM-CRF and PowerBERT-BiLSTM-CRF achieved their best validation $F_1$ scores of 0.9007 and 0.9002 at epochs 15 and 14, respectively. MacBERT-BiLSTM-CRF converges faster and reaches its highest value of 0.9245 at epoch 13. The proposed model reaches its highest $F_1$ score of 0.9354 at epoch 15.

As shown in Table 6, the proposed model achieves higher precision, recall and $F_1$ score on the test set than BERT-BiLSTM-CRF and PowerBERT-BiLSTM-CRF [23]. Compared with MacBERT-BiLSTM-CRF, the recall remains unchanged, while precision increases by 2% and the $F_1$ score increases by 1%. To assess the stability of the comparison, the proposed model and MacBERT-BiLSTM-CRF were evaluated using three paired random-seed runs. A paired permutation test showed that the difference in overall $F_1$ score between the two models was not statistically significant ($p = 0.25$).

To further examine the model's performance across entity categories, the entity-level precision, recall, $F_1$ score, and support for each entity type are presented in Table 7.

**Table 7.** Results of different entity types.

| Entity Type | *P* | *R* | $F_1$ | Support |
|---|---|---|---|---|
| COMP | 0.8807 | 0.9191 | 0.8995 | 482 |
| DEV | 0.9167 | 0.9333 | 0.9249 | 165 |
| PHEN | 0.8465 | 0.8977 | 0.8713 | 430 |
| VOLT | 1.0000 | 1.0000 | 1.0000 | 141 |

The proposed model achieved the highest performance for voltage-level entities, with an $F_1$ score of 1.0000, whereas defect-phenomenon entities showed the lowest $F_1$ score of 0.8713. Compared with MacBERT-BiLSTM-CRF, the proposed model increased the $F_1$ score for defective-component entities from 0.8833 to 0.8995, representing an improvement of 0.0162. The $F_1$ score for defective-equipment entities increased from 0.9129 to 0.9249, an improvement of 0.0120. A smaller improvement of 0.0032 was observed for defect-phenomenon entities, whose $F_1$ score increased from 0.8681 to 0.8713. The $F_1$ score for voltage-level entities remained unchanged at 1.0000. These results show that the performance gains under the selected hyperparameter configuration were mainly concentrated

in the recognition of defective components and defective equipment, while the improvement for defect phenomena was modest.

Further analysis identified 162 paired error events. Spurious entity predictions were the most frequent error type (68), followed by boundary errors (67), missed entities (18), type-confusion errors (7), and combined boundary-and-type errors (2). The boundary errors comprised 43 boundary contractions, 23 boundary expansions, and one boundary shift. These results indicate that the remaining limitations are mainly associated with suppressing spurious predictions and determining the complete boundaries of defect phenomena and long engineering expressions.

In summary, the results indicate that incorporating engineering-format information can improve the recognition of certain entity types under the current experimental setting, although the overall $F_1$ score difference from MacBERT-BiLSTM-CRF was not statistically significant.

## 6. Conclusions

This paper proposes a named entity recognition model for GIS defect texts incorporating an engineering format-aware masking strategy. Named entity recognition was then performed on GIS defect texts. The main conclusions are as follows:

1. The engineering format-aware masking strategy can change the target distribution during continued pretraining. The proposed strategy produces 459, 316, 82 and 120 complete masks for engineering abbreviations, phase formats, measurement value and unit formats, and equipment and location formats, respectively. These values are all higher than those obtained using the conventional MacBERT masking strategy.
2. Under the same downstream model and training conditions, the precision, recall and $F_1$ score of the proposed model on the test set are 0.89, 0.92 and 0.90, respectively. Compared with MacBERT-BiLSTM-CRF, precision increases by 2%, recall remains unchanged, and $F_1$ score increases by 1%.
3. Analysis of typical samples shows that the proposed model can more completely recognize engineering fragments such as “sectional 2015 circuit breaker bay”, “PT disconnector” and “phases A and C are in the open position, while phase B cannot be opened”. This reduces incorrect segmentation of equipment numbers, engineering abbreviations and continuous phase structures. However, the model still shows unstable boundary recognition for some value-unit combinations, long engineering fragments and certain symbols.

Because the current corpus was collected from a single power-grid system, the generalizability of the results to other systems and reporting styles remains to be further validated. Future research will further improve the complete-fragment sampling mechanism and evaluate the proposed method using defect records collected from different power-grid systems to improve and verify its generalization capability.

**Author Contributions:** Conceptualization, Y.W. and C.Y.; methodology, Y.W.; software, J.L.; validation, Y.W. and Y.Y.; formal analysis, Y.Y.; investigation, Y.W.; resources, C.Y.; writing—original draft, C.Y.; writing—review and editing, C.Y.; visualization, Y.W. and Y.Y.; supervision, J.L.; funding acquisition: Y.W.; project administration, Y.W. and C.Y.; date curation, J.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Science and Technology Project of China Southern Power Grid Co., Ltd. (GDKJXM20240217).

**Data Availability Statement:** The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

**Conflicts of Interest:** Authors Ya Wu, Cuiru Yang and Yao Yao were employed by the company Guangdong Power Grid Electric Power Research Institute. Author Jian Lu were employed by the company China Southern Power Grid Internet Services Co., Ltd. All the authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The authors declare that this study received funding from Science and Technology Project of China Southern Power Grid Co., Ltd. The funder had the following involvement with the study: design, collection, analysis and the decision to submit it for publication.

## References

1. Li, G.; Yu, H.; Liu, H.; Li, S.; Yan, Y.; Zhang, Z.; Li, J. Comparative analysis and research on vibration characteristics of internal and external fastening loose mechanical defects of GIS equipment. *High Volt. Appar.* **2024**, *60*, 12–19; discussion 30.
2. Li, J.; Han, X.; Wang, H.; Zhou, Y.; Chen, H.; Guo, R.; Si, W. Review of partial discharge detection technology in electrical equipment: 2015–2025. *High Volt. Eng.* **2025**, *51*, 3132–3158.
3. Chen, H.; Qin, Y.; Shi, X.; Wang, H.; Han, X.; Li, J. Built-in gas state sensing technology for gas-insulated switchgear. *High Volt. Technol.* **2026**, *52*, 550–559.
4. Wang, Y.; Yan, J.; Yang, Z.; Jing, Q.; Wang, J.; Geng, Y. GAN and CNN for imbalanced partial discharge pattern recognition in GIS. *High Volt.* **2022**, *7*, 452–460. [CrossRef]
5. Mou, J. Defect Analysis of Secondary Equipment Based on Data Mining Methods. Master's Thesis, Shandong University, Jinan, China, 2023. (In Chinese) [CrossRef]
6. Li, Y. Research on Several Problems in Power Equipment Defect Text Mining and Application Based on Knowledge Graph. Doctoral Dissertation, Zhejiang University, Hangzhou, China, 2022. (In Chinese) [CrossRef]
7. Ji, Z.; Kong, D.; Liu, W.; Dong, W.; Sang, Y. Named entity recognition based on deep learning. *Comput. Integr. Manuf. Syst.* **2022**, *28*, 1603–1615.
8. Wang, X.; Gu, Y.; Zhang, H.; Liu, L.; Liu, H.; Li, Q. Data augmentation and application of defect texts for power equipment based on knowledge integration manifold. *Power Syst. Technol.* **2024**, *48*, 1690–1699.
9. Zou, G.; Wei, J.; Wang, C.; Zhang, Y. Knowledge graph-based portrait depiction technology for substation equipment. *Guangdong Electr. Power* **2024**, *37*, 86–93.
10. Bayoumi, E.H.E. Power electronics in smart grid distribution power systems: A review. *Int. J. Ind. Electron. Drives* **2016**, *3*, 20–48. [CrossRef]
11. Soliman, H.M.; Saleem, A.; Bayoumi, E.H.E.; De Santis, M. Harmonic Distortion Reduction of Transformer-Less Grid-Connected Converters by Ellipsoidal-Based Robust Control. *Energies* **2023**, *16*, 1362. [CrossRef]
12. Li, G.; Li, Y.; Wang, H.; Xie, Q.; Huang, W.Q.; Hou, J.X. Knowledge graph of power equipment health management: Basic concepts, key technologies and research progress. *Autom. Electr. Power Syst.* **2022**, *46*, 1–13.
13. Pu, T.; Tan, Y.; Peng, G.; Xu, H.; Zhang, Z. Construction and application of knowledge graph in the electric power field. *Power Syst. Technol.* **2021**, *45*, 2080–2091.
14. Gu, Y.; Huo, J.; Yang, H.; Lu, Y.; Guo, Y. BERT-based Chinese named entity recognition method in motor field. *Comput. Eng.* **2021**, *47*, 78–83; discussion 92.
15. Li, A.; Song, H.; Tian, J.; Sheng, G.; Jiang, X. Relation-oriented method for joint extraction of text entities and relations of power equipment fault defects. *High Volt. Appar.* **2026**, *62*, 42–49; discussion 70.
16. Chen, P.; Tai, B.; Shi, Y.; Jin, Y.; Kong, L.; Wang, J.F. Text entity extraction of power equipment defects based on BERT-BiLSTM-CRF algorithm. *Power Syst. Technol.* **2023**, *47*, 4367–4376.
17. Huang, J.; Yang, C.; Feng, T.; Sun, N.; Li, J.; Yu, H.; Cui, H. Research and application of knowledge graph construction for wind turbine operation and maintenance. *Power Syst. Prot. Control* **2024**, *52*, 167–177.
18. Liu, Z.; Wang, H. Retrieval method for defect records of power equipment based on knowledge graph technology. *Autom. Electr. Power Syst.* **2018**, *42*, 158–164.
19. Shao, G.; Wang, H.; Wu, X. Precise information identification method of power equipment defect text based on dependency parsing. *Autom. Electr. Power Syst.* **2020**, *44*, 178–185.
20. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.
21. Xie, Q.; Cai, Y.; Xie, J.; Wang, C.; Zhang, Y.; Xu, Z. Research on construction method and application of knowledge graph for power transformer operation and maintenance based on ALBERT. *Trans. China Electrotech. Soc.* **2023**, *38*, 95–106.
22. Dai, Z.; Zhang, F.; Zhang, J.; Han, X. Construction method of a defect knowledge map of a relay protection device based on a MacBERT-BiLSTM-CRF model. *Power Syst. Prot. Control* **2024**, *52*, 131–143.

23. Jia, J.; Yang, Q.; Fu, H.; Yang, J.; He, Y. Research on Pre-training Language Model Construction and Text Semantic Analysis Based on Power Equipment Big Data. *Proc. CSEE* **2023**, *43*, 1027–1036.
24. goto456. HIT Stopword List. Available online: https://github.com/goto456/stopwords/blob/master/hit_stopwords.txt (accessed on 17 August 2026).

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