frontiers | Frontiers in Future Transportation

TYPE Original Research
PUBLISHED 22 September 2025
DOI 10.3389/ffutr.2025.1601538

OPEN ACCESS

EDITED BY
Rui Esteves Araújo,
University of Porto, Portugal

REVIEWED BY
Jafar A. Alzubi,
Al-Balqa Applied University, Jordan
Alberto Cardoso,
University of Coimbra, Portugal
André Gonçalves,
Universidade do Porto, Portugal

*CORRESPONDENCE
Tong Xing,
xingtong2001@126.com

RECEIVED 28 March 2025
ACCEPTED 19 August 2025
PUBLISHED 22 September 2025

CITATION
Pan L, Xing T, Zhang H, Zhao Y, Yuan Y, Dai W and Dong Z (2025) Research on text information recognition and mining methods for fault records of traction power supply equipment.
Front. Future Transp. 6:1601538.
doi: 10.3389/ffutr.2025.1601538

Research on text information recognition and mining methods for fault records of traction power supply equipment

Like Pan[1], Tong Xing[1]*, Haibo Zhang[1], Yingxin Zhao[1], Yuan Yuan[1], Wenrui Dai[1] and Zhanhao Dong[2]

[1]Standards and Metrology Research Institute, China Academy of Railway Sciences Corporation Limited, Beijing, China, [2]School of Electrical Engineering, Southwest Jiaotong University, Chengdu, China

The fault records of traction power supply equipment contain rich historical fault processing experience, which is of great significance to the fault handling of traction power supply equipment. However, the fault records of TPSE are unstructured text data, and manual processing of them is time-consuming, labor-intensive, and inefficient. Therefore, the fault records have long been left idle in data systems, lacking exploration and application. In view of this situation, this paper proposes an entity information recognition method for fault records based on the BERT-BiLSTM-CRF algorithm, achieving automated and efficient mining of fault record information. Subsequently, based on the recognized entity information from fault records, a knowledge graph for traction power supply equipment fault handling is constructed. Finally, the retrieval capability of the knowledge graph is improved through an entity similarity-based fast retrieval algorithm, and a decision-making method for fault handling in traction power supply equipment is proposed. This method can quickly associate and recommend similar historical fault handling cases for current equipment faults, thus facilitating knowledge sharing and assisting in enhancing the efficiency and intelligence level of fault handling for maintenance operators.

KEYWORDS

rail transportation, text mining, knowledge Engineering, expert systems, traction power supplies

1 Introduction

From the completion and commissioning of China's first electrified railway in 1961 to the end of 2022, the total operating mileage of China's electrified railways has reached 114,000 km, ranking first in the world. This remarkable expansion underscores the increasing reliance on traction power supply equipment (TPSE), which forms the backbone of railway electrification systems. In recent years, with the continuous construction of electrified railways, especially the rapid development of high-speed rail, the number of TPSE units along the railway network has continued to grow. Over time, the traction power supply system inevitably experiences equipment failures and faults during its continuous operation. Efficient handling of these faults is crucial for ensuring the reliability and safety of railway operations. Consequently, corresponding maintenance and repair measures are undertaken, along with the documentation of these occurrences. The number of TPSE fault records continues to grow with ongoing construction and operations. With the gradual advancement of digitalization, most TPSE fault records, both historical and current, are recorded in natural language by operation and maintenance personnel. These textual records form a large corpus of "textual big data" for fault diagnosis and handling, encompassing the historical

fault conditions and remedial measures for all TPSE. These fault records constitute a valuable data asset, offering critical insights into fault diagnosis, failure patterns, and effective handling measures.

There are many types of TPSE, including substation equipment, overhead lines and telecontrol systems. According to the estimation of a maintenance sector in southern China, the annual accumulated number of TPSE fault records for about 2000 km of railway under its jurisdiction has reached nearly 300 entries. Based on this rough estimate, the accumulated TPSE fault records in 24 maintenance and management sections across the country in 1 year will exceed 17,000 entries. These massive TPSE fault records are all unstructured text data, and currently the on-site information system cannot directly understand and process them. The classification and statistical analysis of these records mainly rely on manual processing. Manual handling not only imposes a significant labor burden but also limits the efficiency and precision of fault data analysis. According to on-site estimates, the equipment fault records accumulated in a single maintenance sector over a year require a professional technical management personnel to spend 8–10 days to complete statistical classification. At the same time, these historical equipment fault records are real samples generated during actual operations, containing rich information that cannot be simulated in the laboratory. They provide insights into various aspects such as types, phenomena, causes, and corresponding remedial measures of equipment faults over extended periods. However, due to the lack of information processing methods and the time-consuming nature of manual handling, it is currently challenging to conduct in-depth mining and application of this data at the site.

In response to the aforementioned on-site issues and requirements, numerous studies have been conducted to investigate text information processing and mining. Nouns or phrases with specific meanings in the text are typically identified and extracted using entity information recognition (EIR) methods (Maggini et al., 2020). The traditional method for EIR relies on dictionaries and rules (Wang et al., 2021; Wang et al., 2022). While it has simple principles and is easy to implement, it requires experts to develop rule templates. However, when dealing with complex and diverse text formats, it becomes challenging to exhaust all possible rules, resulting in difficulties in rule compilation. Similarly, entity information methods based on hidden Markov models (HMM) (Morwal et al., 2012), conditional random fields (CRF) (Lei et al., 2014), and other machine learning algorithms depend on the precision of feature templates. These templates need to be written and formulated by domain experts, limiting their generalizability.

As a result, these methods struggle to achieve high precision and efficiency in practical applications.

In contrast, the EIR method based on deep learning can automatically model and extract text features without manual feature selection, making it highly relevant in practical applications. Recent advances in deep learning, such as the bidirectional encoder representations from transformers (BERT) (Jacob et al., 2018) model, offer promising solutions by capturing contextual semantic information more effectively. Yang et al. (2025) proposes a primary equipment fault diagnosis method based on language models such as BERT. Dai et al. (2021) introduced a BiLSTM-CRF model by combining the bidirectional long short-term memory (BiLSTM) network with CRF to recognize specified entity information. The inclusion of BiLSTM enhances the extraction of text features, leading to significant improvements in recognition performance compared to the CRF model. Wang et al. (2024) proposes a novel data-driven dynamic predictive maintenance strategy using a CNN-BiLSTM ensemble for RUL prediction, which uniquely incorporates uncertain system mission cycles into maintenance, order, and stock decisions. Jin et al. (2025) employed a CRF layer combined with a BiLSTM network to model label dependencies and achieve high-accuracy entity recognition (92.49% F1 score) for constructing the composite insulator knowledge graph. Scholars have also utilized BERT pre-trained language models (Liu et al., 2022), which integrate contextual semantic information and train word vectors capable of expressing semantic features more accurately. Despite these advancements, the application of deep learning methods to TPSE fault records remains underexplored, and the potential for leveraging this data to improve fault handling decision-making is yet to be fully realized.

Regarding the application of text information mining, Meng et al. (2023), Stephen et al. (2020) have utilized deep learning or topic modeling methods to mine text information and achieve power text classification. Qiu et al. (2016) have classified circuit breaker fault texts using the one-hot model and the K-Nearest Neighbors (kNN) algorithm. Rudin et al. (2012) have mined information from tens of thousands of cable fault records in New York City, predicting the fault risk of components and systems, thus providing assistance for cable maintenance well inspections. Sun et al. (2016) have proposed a probability framework for identifying power outage-related social data in the social software Twitter, offering insights for power grid outage management. However, these applications primarily focus on the overall features of fault texts and do not fully explore the specific entity information present in these texts. When various types of entity information extracted from text are stored in conventional relational databases such as Structured Query Language (SQL) Server (Menasce and Gomaa, 2000) or Oracle (Jahangirova et al., 2021), the expansion of the database size can significantly reduce data operation and query speed. In contrast, a knowledge graph enables the comprehensive organization and utilization of text information. It visually represents entity information and their relationships in the form of a graph and allows for rapid data operations and queries (Ji et al., 2022; Sawant et al., 2019; Shang et al., 2021). Gao et al. (2020) have developed a logic scheme for line tripping fault processing by constructing a knowledge graph of power system dispatching, which assists dispatchers in on-site handling. Wang et al. (2021) have built a knowledge graph of power grid fault disposal based on the fault plan, applying it to support decision-making in fault disposal. However, this method only matches historical fault cases that are identical to the current

Abbreviations: Q, Query vector in self-attention mechanism; $S(x,y)$, Score of a label sequence for a given input in CRF; K, Key vector in self-attention mechanism; $\bar{y}$, The true label sequence of the input; V, Value vector in self-attention mechanism; Y, Set of all possible label sequences for the input; d_k, Dimension of input vector (columns of Q and K) y^*, Optimal label sequence predicted by CRF (via Viterbi algorithm); W^O, Weight matrix used to linearly map concatenated heads to output vector; P, Precision: ratio of correctly identified entities to all identified entities; $head_i$, Output of the attention operation in the i-th head; R, Recall: ratio of correctly identified entities to all true entities; x_t, Input vector at time t to the BiLSTM; F_1, F_1 score: harmonic mean of precision and recall; h_t, Hidden state vector at time t output by BiLSTM; v_i, Word vector of the i-th word in an entity phrase; A, Transition score matrix for adjacent labels in CRF; ent, Entity vector obtained by summing word vectors of an entity phrase; P, Label score for a specific character and entity type in CRF; $sim(a,b)$, Cosine similarity between vectors a and b.

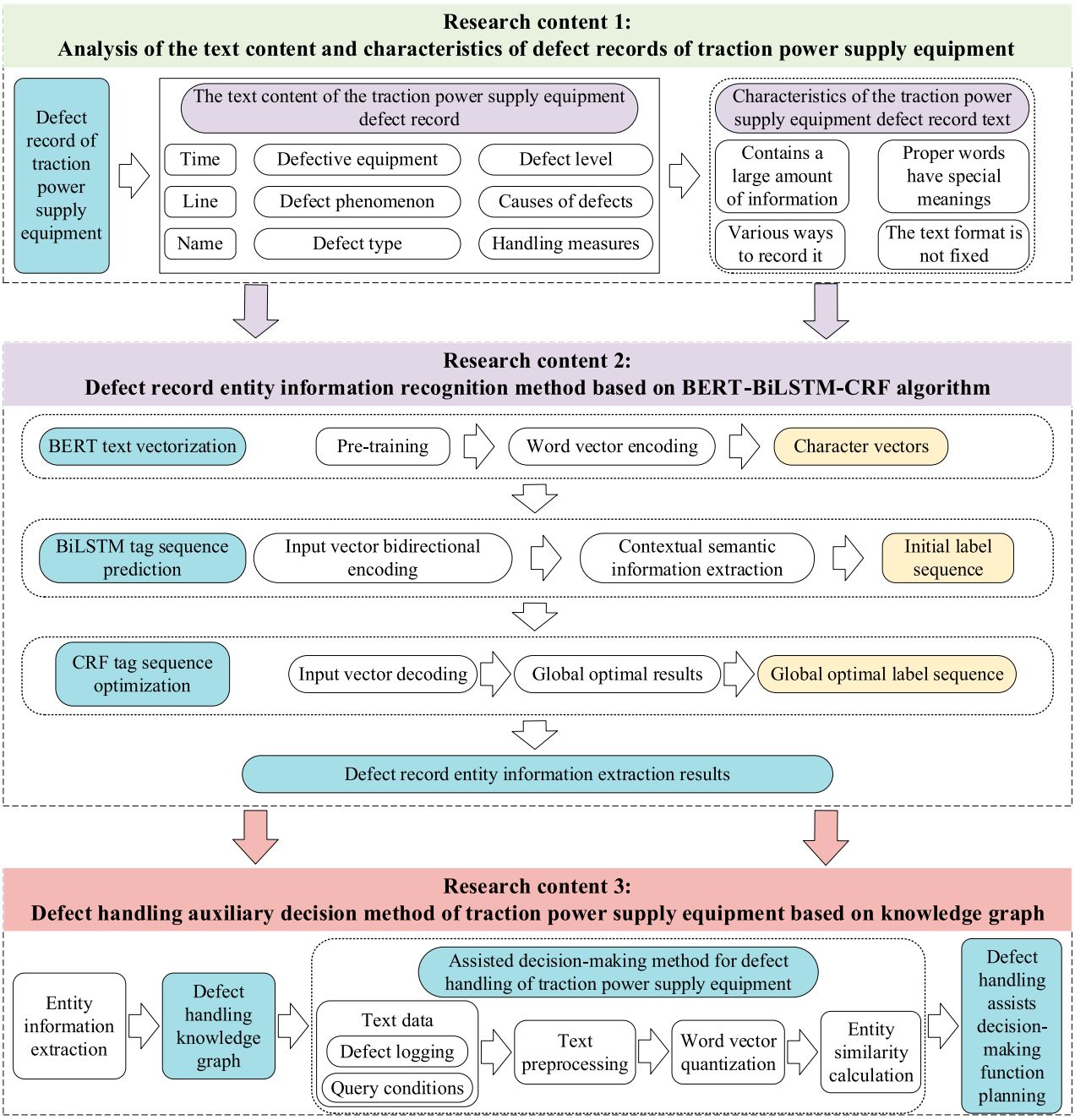

FIGURE 1
The overall framework diagram of the paper.

faults, disregarding parts with high similarity. Consequently, the recommendation information may be incomplete, and the decision-making function is limited.To address these challenges, this paper proposes an entity information recognition method for fault records based on the BERT-BiLSTM-CRF algorithm, achieving automated and efficient mining of fault record information. By automating the extraction and analysis of fault information, this method aims to alleviate the burden of manual processing and enhance the overall intelligence of fault handling. Subsequently, based on the recognized entity information from fault records, a knowledge graph for TPSE fault handling is constructed. This knowledge graph serves as a structured repository of fault knowledge, facilitating efficient retrieval of historical fault cases and supporting informed decision-making. Finally, the retrieval capability of the knowledge graph is improved through an entity similarity-based fast retrieval algorithm, and a decision-making method for fault handling in TPSE is proposed. This comprehensive approach not only streamlines fault handling processes but also promotes knowledge sharing among maintenance operators, ultimately enhancing the efficiency and reliability of railway operations. The specific organizational framework of this paper is illustrated in Figure 1 below.

TABLE 1 Examples of typical TPSE fault records and their characteristics analysis.

Number	Historical TPSE fault records
1	On 20 December 2020, the communication of a 214 protection device in Kunshan Substation of Jinghu Line was interrupted. On 20 December 2020, it was found that the communication plug-in of 214 protection device was damaged, belonging to the class B fault of integrated automation equipment, and it recovered to normal after replacement
2	On 24 December 2020, a 2153 GK temperature measurement in Chunshen Substation of Hukun Line found that the temperature was too high, belonging to other faults of class B. On 28 December 2020, strengthened the temperature measurement, contacted the overhead lines work area, and recovered to normal after handling in combination with the skylight
3	On 1 January 2021, the 2312 GK optocoupler terminal block of the Chang'an Town Section Post on the Hukun Line was damaged, belonging to class C fault of component damage. On 12 January 2021, replaced the 2312 GK optocoupler terminal
4	On 10 January 2021, the 211DL control circuit of Chang'anji Substation of Hewu Line was broken, which belongs to other faults of Class A. On 10 January 2021, it recovered automatically
5	On 19 March 2021, the 219 circuit breaker of Chang'anji Substation of Ningxi Line reported a control circuit broken, and the 219DL panel and cabinet on-off light did not light up. On 19 March 2021, it was found through inspection that 2 conductors with distinctive thicknesses were linked to the identical end, causing a false joining and being classified as a B level fault due to poor insulation. Separated the two wires and shorted them with short connectors to restore normal operation
6	On 5 July 2022, the telecontrol channel of Qiaosi Substation, Genshanmen Switching Post, Nanxingqiao Subsection Post and Switching Post of Hukun Line was interrupted, which is a Class B fault of poor telecontrol channel. From 5 July 2022 to 18 July 2022, it interrupted every night, and would resume in the early morning. After the interruption, personnel would be arranged to be on duty. Several devices are replaced in the substations including routers and cables. After the 18th, no fault occurred, and the main cause is considered as the performance degradation with the accumulation of operation time

1.1 Analysis and organization of TPSE fault records

This study analyzes and categorizes the characteristics of fault records pertaining to TPSE. By examining historical fault records collected from real-world on-site scenarios, this study identifies common issues they contain. This analysis forms the basis for selecting appropriate methods for recognizing entity information.

1.2 Proposed method for EIR

The study introduces a method for recognizing entity information in TPSE fault records, utilizing the BERT-BiLSTM-CRF algorithm. This approach enables automatic and accurate recognition and extraction of fault record information.

1.3 Construction of a knowledge graph for fault handling

The research includes the development of a knowledge graph specifically designed for managing and handling TPSE faults. Additionally, the paper presents a decision-making method for fault handling based on an entity similarity fast retrieval algorithm. This innovative approach facilitates efficient mining of historical fault record information and promotes the sharing of handling experiences. By leveraging this knowledge graph and decision method, the proposed framework enhances the efficiency and intelligence of fault management processes.

The remainder of this paper is organized as follows. Section 2 analyzes the characteristics of TPSE fault records and identifies key challenges for entity recognition. Section 3 presents the proposed BERT-BiLSTM-CRF-based method for entity information recognition, including its structure and evaluation metrics. Section 4 introduces the construction of the TPSE fault handling knowledge graph and details the entity similarity-based fast retrieval algorithm for decision-making. Section 5 demonstrates case studies and experimental results to validate the effectiveness of the proposed approach. Finally, Section 6 concludes the study and outlines potential future work.

2 Analysis of characteristics of TPSE fault records

The TPSE fault record corpus employed in this study was collected from three representative railway maintenance and management sections across different regions, encompassing more than 4,500 km of railway lines and over 20 types of traction power supply equipment. This corpus contains 2,412 historical fault records accumulated over multiple years, which reflect the actual working conditions and maintenance scenarios of electrified railways in China. Each fault record typically includes multi-level information such as time, line, substation, equipment, fault phenomenon, cause, type, fault class, and handling measures. These nine entity types vary greatly in their length and expression, ranging from simple terms to complex phrases spanning multiple words. For instance, as shown in Table 1, some handling measures may be as short as a single action word like "replace" or extend to a complete procedural description involving multiple steps. Additionally, the records frequently contain professional vocabulary specific to the traction power supply domain—for example, "curfew maintenance" refers to scheduled maintenance windows outside train operation hours, and equipment identifiers such as "214DL" may indicate circuit breakers but differ in notation from standard equipment names. This results in semantic variations and inconsistent naming, which present significant challenges for accurate entity recognition. Moreover, the corpus includes examples where certain entity information, such as fault causes, may be absent or implicitly described, adding further complexity to automated text mining. Overall, the corpus is rich, diverse, and highly representative of real-world scenarios, providing a robust basis for analyzing the

difficulties of named entity recognition in TPSE fault records and for developing an efficient, generalizable extraction method.

To achieve automatic and effective extraction and analysis of TPSE fault records, its characteristics need to be analyzed first. After several years of on-site inspection and research, a total of 2,412 fault record samples were collected from three maintenance and management sections, covering over 4,500 km of lines and more than 20 types of TPSE. Due to space limitations, here are several typical historical TPSE fault records, as shown in Table 1.

It can be seen from Table 1 that the fault record of TPSE is a description of the equipment fault situation and handling measures during the previous fault process, which is similar to the universality involved in the power system field and rich in strong characteristics of the traction power supply field. Specifically, the TPSE fault records exhibit several features:

2.1 TPSE fault records contain a large amount of information

A fault record usually contains nine key entity types, including time, line, substation, equipment, phenomenon, cause, type, class, and handling measures. Compared to other types of text, it contains more types of information. In addition, these nine entity types have diverse formats and varying lengths. For example, fault records 1 and 6 in Table 1 contain handling measures that range from a single word to 58 words, respectively. Therefore, to achieve high-precision recognition of all entity types, more robust and generalizable entity recognition methods are required.

2.2 The fault record contains some professional vocabulary in the area of railway power supply, which differs from their commonly used meanings

In the second record of Table 1, the term "skylight," which is also commonly referred to as "curfew maintenance," (Lin et al., 2023) refers to the time period from midnight to 4 a.m. when the catenary system is powered off to allow on-site inspection and maintenance of railway traction power supply equipment. This time period is used to ensure the safe and normal operation of the railway system the following day.

Similarly, in everyday language, "up" and "down" generally mean upward and downward movement, whereas in railway power supply they refer to two distinct train running directions as well as the corresponding power supply sections. In addition, there are other specialized terms. For example, "cross zone" refers to "cross zone power supply," which is a standby mode where an adjacent traction substation temporarily supplies power when a traction substation cannot operate normally due to unexpected faults or maintenance.

2.3 There are numerous types of TPSE and diverse recording methods

In TPSE fault records, the equipment name is often not recorded using its standard name, such as "circuit breaker," "isolation switch," or "transformer." Instead, operators add specific identification numbers and directional information. For example, in Table 1, "211DL," "219DL," and "219 circuit breaker" all refer to circuit breakers, but their numbering and representations vary. The term "circuit breaker" may be used directly, or the code "DL" may be used instead. In addition, the ways of recording transformer names are even more diverse, such as "02B," "03B," "4#B," and "1# main transformer." This variation may lead to identical equipment being recognized as different devices, resulting in deviations or even errors in interpreting fault records. This places higher demands on the semantic understanding capability of the entity recognition model.

2.4 The format of fault record text is not fixed, and some key entity information may be missing

In Table 1, records 1 and 5 contain the cause, while fault records of the others lack the cause entity information. Moreover, while fault causes are similarly absent in these records, the context differs: For Log Entries 2 and 3, the described phenomena inherently indicate the root cause, hence maintenance personnel did not document causal details separately; Fault record 4 is due to the fact that the equipment has automatically recovered to normal before the fault can be processed, so the cause of the fault has not been explored; However, fault record 6 lacks both the fault equipment and cause, which is because the cause has not been found after inspection. From the above analysis, it can be seen that the types of information contained in fault records are different from each other, and the format is not completely the same, which poses certain difficulties in semantic understanding.

3 A method to fault record entity information recognition based on BERT-BILSTM-CRF algorithm

To extract structured information from unstructured TPSE fault records, we formulate entity recognition as a sequence labeling task. In this formulation, each word in the fault record is assigned a label that identifies whether it belongs to a named entity and, if so, its position within that entity. We adopt the widely used Begin-Inside-Outside (BIO labeling scheme), where:

- B (Begin) indicates the first word of an entity,
- I (Inside) denotes a word that is part of an entity but not at the beginning,
- O (Outside) represents words that do not belong to any entity.

This labeling format allows the model to identify both the boundary and type of each entity within fault texts, such as equipment names, causes, or handling measures. An example of BIO labeling is illustrated in Table 2.

Based on this labeling framework, we propose a deep learning-based method combining BERT, BiLSTM, and CRF to recognize multiple entity types in fault records automatically. The overall architecture is shown in Figure 2, and the workflow consists of three main stages: text vectorization with BERT, label sequence prediction with BiLSTM, and global optimization with CRF.

TABLE 2 Demonstration of entity labeling.

Word	Label	Word	Label
The	O	a	O
272	B-Equipment	class	O
circuit	I-Equipment	A	B-Class
breaker	I-Equipment	fault	O
refused	B-Phenomenon	of	O
to	I-Phenomenon	poor	B-Type
operate	I-Phenomenon	remote	I-Type
	O	control	I-Type
which	O	channel	I-Type
is	O		O

As depicted in Figure 2, the TPSE fault records are first fed into the BERT model at the sentence level, and each word is vectorized as input embeddings. Subsequently, the BiLSTM model is used to extract bidirectional textual features and to generate an initial label sequence for the fault records. Finally, the CRF model determines the globally optimal label sequence by applying the Viterbi algorithm for optimal sequence decoding. The recognized entities are then extracted by mapping the final labels back to each word.

3.1 BERT text vectorization

As a deep bidirectional language representation model, BERT consists of four layers: the first layer is the input layer; the second and third layers are Transformer encoder layers; and the fourth layer generates the output. The structure is illustrated in Figure 3, where N denotes the total number of words in the input fault record.

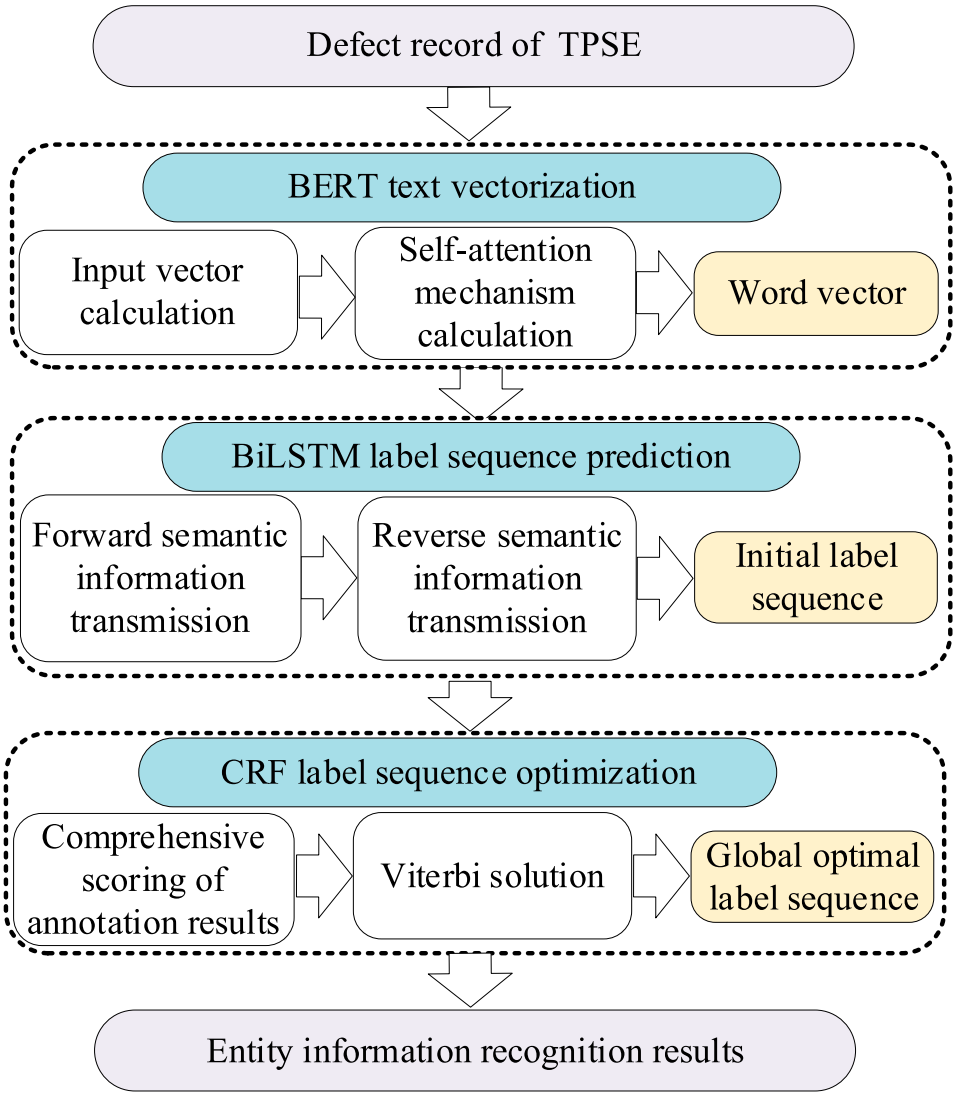

FIGURE 2
EIR process of BERT-BiLSTM-CRF mode.

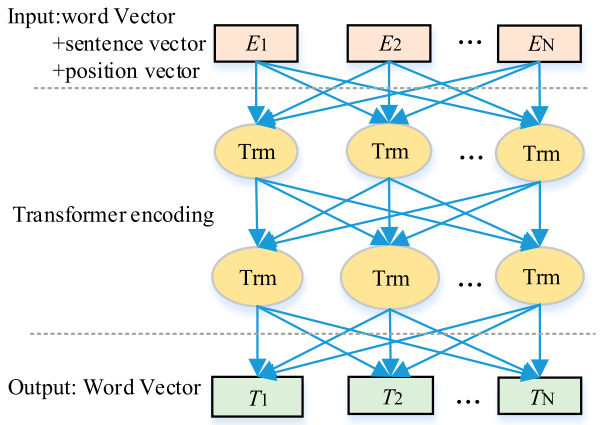

FIGURE 3
Schematic diagram of BERT model.

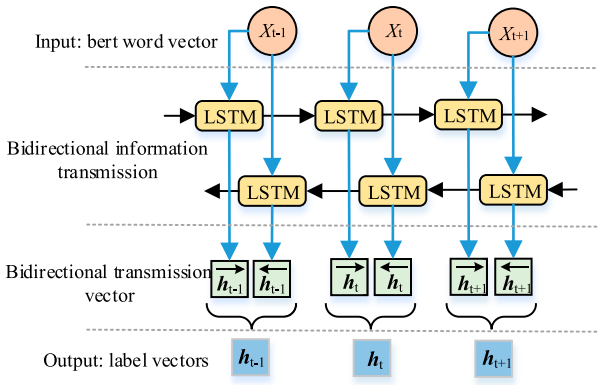

FIGURE 4
Schematic diagram of BERT model.

In the first layer, the input layer of the BERT model consists of three parts: word vector, sentence vector, and position vector.

There is a total of 12 transformer encoding modules in the second and third layers. Its core is the self-attention mechanism, which maps each word with its context, then assign weights to each word in the context, and update the current word vector based on the weights. The word vector obtained through this training consists of both the meaning of the word and the contextual semantic information. The calculation method for the output of the self-attention mechanism is shown in Formula 1. Among them, $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ represent the query vector, key vector, and value vector, respectively, while d_K represents the input vector dimension, which is the number of columns of the query vector $\boldsymbol{Q}$ and key vector $\boldsymbol{K}$.

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

Due to the fact that the self-attention mechanism can only capture one-dimensional information, transformer adopts a multi-head attention mechanism to obtain multi-dimensional information of fault records. It first performs h different linear map of $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$, then calculates the Attention matrix and splices them according to the mapping results of each time, and finally maps the spliced word vector to the same dimension as the input word vector through matrix $\boldsymbol{W}^O$. The calculation formula is shown in Formulas 2, 3.

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Contact(\boldsymbol{head}_1,\ldots,\boldsymbol{head}_h)\boldsymbol{W}^O \tag{2}$$

$$\boldsymbol{head}_i = Attention\left(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^K,\boldsymbol{V}\boldsymbol{W}_i^V\right) \tag{3}$$

Finally, the fourth layer is the output word vector, which integrates the meaning and contextual information of the words and can more appropriately express the semantic features of fault records.

3.2 BILSTM label sequence prediction

To improve the precision of final EIR, this paper connects BiLSTM after the BERT model to further extract contextual semantic information from word vectors, thereby obtaining more comprehensive global semantic features.

As shown in Figure 4, BiLSTM model can achieve the transmission of textural information towards both ways. At the moment of t, inputting vector $\boldsymbol{x}_t$ generates vector $\overrightarrow{\boldsymbol{h}_t}$ at the forward LSTM while $\overleftarrow{\boldsymbol{h}_t}$ at the backward LSTM. Combining the above two yields the label vector $\boldsymbol{h}_t$. Each value of $\boldsymbol{h}_t$ represents the probability of the word classified into the particular type of entity. Therefore, the dimension with the largest value in $\boldsymbol{h}_t$ is considered as the label for the specific word.

3.3 CRF label sequence optimization

After BiLSTM predicts the label sequence, CRF is used to accommodate the constraints and interconnections among various entity labels and optimize the original label sequence.

Each word in the known fault record obtains the result $\boldsymbol{h}_t$ after undergoing BiLSTM operation. Firstly, CRF give scores to the annotation results of fault records by interating the scores of various entity labels in a single word $\boldsymbol{h}_t$ and the transfer scores between adjacent word labels, which is calculated as follows

$$S(x,y) = \sum_{i=1}^{n}\left(A_{y_{i-1},y_i} + P_{i,y_i}\right) \tag{4}$$

In Formula 4, A_{y_{i-1},y_i} indicates the transfer score of nearby labels of two words within a fault record, P_{i,y_i} represents the score of the y_i label of the *ith* word in the fault record, and n represents the number of words.

Secondly, for a given fault record X, the conditional probability formula of any label sequence y is:

$$P(y \mid X) = \frac{e^{S(X,y)}}{\sum_{\tilde{y}\in Y} e^{S(X,\tilde{y})}} \tag{5}$$

In Formula 5, $\tilde{y}$ represents the real label sequence of X, and Y represents all possible label sequences of X.

Finally, when CRF is used to predict the final fault record label sequence, Viterbi algorithm is used to obtain the global optimal solution, and the solution formula is:

$$y^* = \text{argmax}S(X,\tilde{y}) \tag{6}$$

In Formula 6, y^* represents the fault record label sequence with the highest score. Each word in each fault record is labeled with an entity label. By identifying the corresponding label for each word, entity information can be extracted.

3.4 Labeling methods and evaluation indicators

To distinguish entities from non-entities. The BIO labeling scheme, as introduced above, is applied for the annotation of fault records.

In order to observe the effectiveness of the EIR model, three indicators, P, R, and F_1 value, are used to evaluate from the entity level. The calculation formula is shown in (Equations 7–9).

$$P = \frac{\text{Entities correctly identified}}{\text{Entities identified}} \tag{7}$$

$$R = \frac{\text{Entities correctly identified}}{\text{Entities annotated}} \tag{8}$$

$$F_1 = \frac{2PR}{P+R} \tag{9}$$

According to (Equations 7–9), the precision P is aimed at all entities identified using the proposed method, while the recall R is aimed at all entities annotated in the sample. The two assess the performance of the proposed method from varying perspectives. The harmonic mean of the two, denoted as F_1, can show the effect of the EIR. A larger F_1 depicts a better result of the EIR model.

4 Method to fault record information mining rested on entity similarity fast retrieval algorithm

After completing the automatic recognition and extraction of entity information from the TPSE fault records, further management and retrieval of this information are required for direct application in on-site practice. Therefore, this paper constructs a TPSE fault handling knowledge graph to store and manage fault record information and proposes a corresponding decision-making method. The proposed method recommends historical fault cases most similar to the current fault as references for operation and maintenance personnel, thereby enhancing the efficiency and intelligence of fault handling.

4.1 Construction method of knowledge graph of TPSE fault handling

Since the entity types and relationships in TPSE fault records are relatively stable, this paper uses a top-down approach to construct the TPSE fault handling knowledge graph. First, schema links are defined and the schema layer is constructed. Then, relevant entities are extracted from the text based on the schema layer to build the data layer. The specific construction process is illustrated in Figure 5.

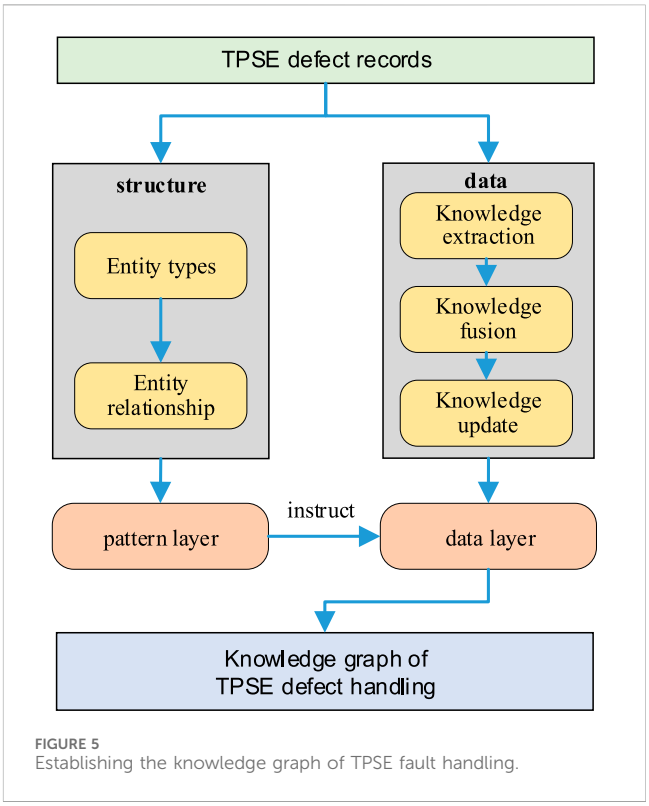

FIGURE 5
Establishing the knowledge graph of TPSE fault handling.

4.1.1 Construction of pattern layer

The Pattern Layer is used to describe the entity types and the relationships between entities, serving as the organizational framework of the knowledge graph. In this study, by summarizing the entity types and relationships, a Pattern Layer for the TPSE fault handling knowledge graph has been constructed, as shown in Figure 6.

First, Figure 6a illustrates the overall structure of the knowledge graph. As depicted, the knowledge graph for fault handling in traction power supply equipment centers around the "Traction Power Supply Equipment" node, with "Faultive Equipment" and "Fault Phenomenon" as key nodes radiating outward, forming a centralized graph structure that gradually diverges from the core.

Second, Figure 6b presents a local subgraph centered on a single traction power supply equipment node. For each equipment, eight categories of entity information were selected: time, line, substation, phenomenon, cause, type, level, and handling measures. These constitute the basic entity types of the knowledge graph nodes.

4.1.2 Construction of data layer

There are three steps to construct the data layer: extracting knowledge, fusing knowledge, and updating knowledge.

Firstly, knowledge extraction involves extracting 9 types of entity information from fault records, including fault equipment, fault phenomena, fault causes, and handling measures through the BERT-BiLSTM-CRF EIR model.

Secondly, for the extracted entity information, knowledge fusion is necessary to improve precision and reduce redundancy. It involves entity disambiguation and coreference resolution. Entity disambiguation refers to distinguishing entities that may be ambiguous or have multiple meanings, while coreference

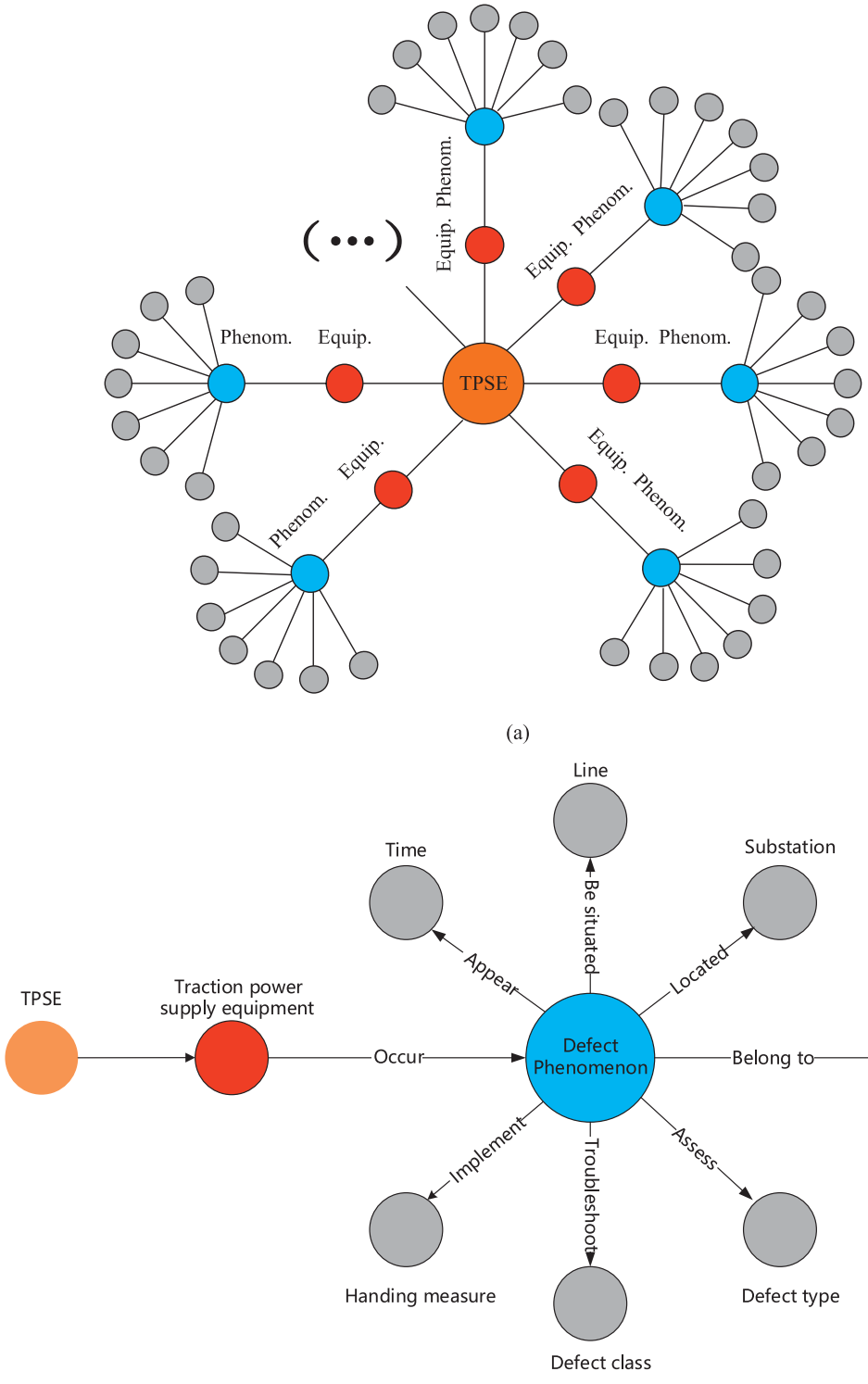

FIGURE 6
(a) Overall Structural Diagram of the Traction Power Supply Equipment Fault Handling Knowledge Graph. **(b)** The pattern layer of knowledge graph of TPSE fault handling.

resolution involves merging entities that refer to the same concept in the knowledge graph. Due to clear terminology specifications in the traction power supply domain, ambiguity is rare in fault records; therefore, entity disambiguation is unnecessary. However, because there are often multiple ways to record a device in fault logs (e.g., "214DL", "217DL", "214 circuit breaker", "217 circuit breaker"),

all of these refer to circuit breakers. Therefore, coreference issues are common in fault records, and it is necessary to perform coreference resolution on equipment names so that all variations referring to the same device type are replaced with a unified name.

Finally, with the ongoing intelligent development of traction power supply systems, both the types of TPSE and the complexity of fault conditions continue to increase. Therefore, the TPSE fault handling knowledge graph must be continuously updated to ensure its timeliness and comprehensiveness. A detailed technical description of the dynamic updating mechanism is provided below.

New fault records are first preprocessed using entity recognition methods (e.g., BERT-BiLSTM-CRF) to extract key entity information and structure them in a format consistent with the existing data in the knowledge graph. The extracted entities are then categorized into predefined categories (e.g., "Equipment," "Fault Phenomenon," "Cause") and their relationships are identified. For entities matching existing ones, a fusion process updates the knowledge graph while resolving ambiguities using techniques like coreference resolution. If new entities or relationships are introduced, they are dynamically added to the knowledge graph, expanding its content. After updates, the system validates the integrity and consistency of the data, ensuring no redundancy or conflicts. As the knowledge graph evolves, a feedback loop—via expert reviews or automated detection—continuously improves entity recognition and relationship extraction, enhancing precision and adaptability over time.

4.2 Decision-making method for TPSE fault handling

Once the TPSE fault handling knowledge graph has been constructed, if it relies solely on its basic retrieval function, only historical faults that are exactly identical to the current fault can be retrieved, which limits its practical usefulness. Therefore, this study enhances the retrieval capability of the knowledge graph by introducing an entity similarity-based fast retrieval algorithm and proposes a decision-making method for TPSE fault handling. The main steps of the entity similarity-based fast retrieval algorithm are as follows.

4.2.1 Text preprocessing

Both the fault phenomenon nodes in the knowledge graph and the descriptions of the current fault are unstructured textual data. First, these texts need to be preprocessed. They are then segmented into separate word sequences. At the same time, irrelevant or redundant words, such as auxiliary words and modal particles, are removed using a stop-word list.

Notably, traction power supply equipment fault records contain a large number of technical terms, and ambiguities or variations in these terms may interfere with information extraction. The system ensures the precision and consistency of term processing through several methods. First, it applies term standardization to map terms with different forms or spellings into a unified standard, such as consolidating "switch," "breaker," and "isolator" into a standardized term. Second, it handles term variants by recognizing synonyms, ensuring consistent identification of similar terms, such as treating "circuit breaker" and "electric circuit breaker" as synonyms. For ambiguous terms, the system determines their specific meaning through contextual analysis; for example, it distinguishes whether "bus" refers to an "electric bus" or a "communication bus" based on the context. Additionally, as equipment and fault types evolve, the system regularly updates its dictionary by extracting new terms from equipment documentation, fault records, and industry reports to maintain its recognition capability for emerging terms. Finally, the system eliminates redundancy and optimizes information processing, preventing duplication caused by different formats of equipment IDs (e.g., "214DL" and "214-DL"), thus ensuring both efficiency and precision in the recognition process.

4.2.2 Entity vectorization

After preprocessing, as for the current faultive equipment, match the same fault equipment node in the knowledge graph, and vectorize the corresponding fault phenomenon node and the input fault phenomenon, so as to calculate the entity similarity. Both the fault phenomenon node and the input fault phenomenon are fault phenomenon entities, belonging to short text and composed of single or multiple words. Therefore, by adding the word vectors corresponding to all the words it contains, the vectorized representation of the fault phenomenon entity is obtained.

Due to the fact that fault phenomenon entity is a part of the TPSE fault record text, the words that constitute the fault phenomenon entity must also be included in the fault record. Therefore, vectorization tools can be used to train the word vectors corresponding to all words based on the fault record. This paper uses word2vec Word embedding model as a fault record vectorization tool. It can learn the relationships between various words from a large amount of corpus, achieve distributed representation of word features, avoid one-to-one mapping between words and vectors, and represent the actual meaning of words in abstract vector form (Yu et al., 2018).

After preprocessing the fault record corpus set, the preprocessed corpus set is trained based on the word2vec word embedding model. The training algorithm selects Skip-Gram model, the word vector dimension is set to 100 dimensions, and the minimum word frequency is set to 1. After training, 2141 words and their corresponding distributed word vectors were ultimately obtained. Some words and their corresponding word vectors are shown in Table 3.

After implementing word vectorization, the word vectors corresponding to the words contained in the fault phenomenon entity are added to form the fault phenomenon entity vector. For example, for a fault phenomenon entity *ent*, if it is composed of p words and $\boldsymbol{v}_i$ ($i = 1, \ldots, p$) is the word vector corresponding to each word, then its corresponding fault phenomenon entity vector is $\boldsymbol{ent} = (\boldsymbol{v}_1 + \cdots + \boldsymbol{v}_p)$.

4.2.3 Entity similarity calculation

The cosine similarity algorithm is a commonly used similarity calculation method that evaluates the similarity between two vectors by calculating the cosine value of the angle between them. Based on the word2vec word vectors obtained through training, the cosine similarity algorithm can calculate the similarity between two word vectors. Similarly, the cosine similarity algorithm can also be used to calculate the similarity between two fault phenomenon entities. Here, the cosine similarity between any two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is defined as $sim\ (\boldsymbol{a},\boldsymbol{b})$.

TABLE 3 Example of entity labeling of fault record.

Word	Word vector
transformer	(0.090765, -0.023804, … … , 0.203654, 0.108283)
circuit breaker	(0.091566, -0.027322, … … , 0.220106, 0.110428)
isolating switch	(0.083842, -0.027324, … … , 0.228602, 0.107575)
relay	(0.139087, 0.056287, … … , 0.220191, 0.067056)
fault	(0.234055, 0.319761, … … , 0.366489, 0.079654)
replace	(0.243218, 0.161711, … … , 0.205377, 0.044075)
alarm	(0.139772, 0.116520, … … , 0.258859, 0.048609)
trip	(0.100268, 0.006582, … … , 0.225062, 0.063269)

Therefore, for the two preprocessed fault phenomenon entities ent_1 and ent_2, which are respectively composed of m and n words, $\boldsymbol{a}_i$ and $\boldsymbol{e}_j$ are the word vectors corresponding to each individual word, and $\boldsymbol{ent_1}$ and $\boldsymbol{ent}_2$ are the entity vectors corresponding to the two fault phenomenon entities. Therefore, $\boldsymbol{ent}_1 = (\boldsymbol{a}_1 + \cdots + \boldsymbol{a}_m)$, $\boldsymbol{ent}_2 = (\boldsymbol{e}_1 + \cdots + \boldsymbol{e}_n)$, and the calculation formula for the entity similarity between the two is shown in (Equation 10):

$$sim(\boldsymbol{ent}_1, \boldsymbol{ent}_2) = \cos\theta = \frac{\boldsymbol{ent}_1 \bullet \boldsymbol{ent}_2}{\|\boldsymbol{ent}_1\|\|\boldsymbol{ent}_2\|} = \frac{(\boldsymbol{a}_1 + \cdots + \boldsymbol{a}_m) \bullet (\boldsymbol{e}_1 + \cdots + \boldsymbol{e}_n)}{\|\boldsymbol{a}_1 + \cdots + \boldsymbol{a}_m\|\|\boldsymbol{e}_1 + \cdots + \boldsymbol{e}_n\|} \tag{10}$$

In Equation 10, θ is the angle between vectors $\boldsymbol{ent}_1$ and $\boldsymbol{ent}_2$.

Then, based on the calculated entity similarity between the input fault phenomenon and each fault phenomenon node, the top-ranked historical fault cases in the knowledge graph are recommended in descending order of similarity. This provides assistance and guidance for the diagnosis and handling of the current fault.

4.2.4 Process of decision-making of TPSE fault handling

In summary, the process of the decision-making method for TPSE fault handling based on entity similarity fast retrieval algorithm is shown in Figure 7.

As shown in Figure 6, the specific processes are described as follows:

a. When operation and maintenance personnel discover a new fault of TPSE, input the current fault equipment and fault phenomenon;
b. According to the input fault equipment, match the same fault equipment node in the knowledge graph, and return all fault phenomenon nodes connected by this node;
c. Preprocess the content of all fault phenomenon nodes and obtain the entity vectors of each fault phenomenon node based on the word2vec word vector;
d. Preprocess the input fault phenomenon and obtain the entity vector of the input fault phenomenon based on the word2vec word vector;
e. Calculate the entity similarity between each fault phenomenon node and the input fault phenomenon separately, and return the historical fault cases corresponding to the top n entity similarity rankings;
f. Output fault information and handling measures of the top n historical fault cases, assist and guide the operation and maintenance personnel in handling current fault.
g. To improve the completeness and adaptability of the knowledge graph, a user feedback loop is introduced, allowing newly handled fault cases to be incorporated into the graph when no prior matching cases exist. This mechanism enables continuous updates with real-world data, gradually enhancing retrieval precision and system robustness.

Among them, word2vec word vectors refer to the word vectors trained based on fault records; n is the number of final output cases, and this paper sets $n = 5$, which is the fault information and processing measures corresponding to the top 5 fault cases in the similarity ranking of the final output fault phenomenon.

5 Case study

5.1 EIR of TPSE fault record

A total of 912 TPSE fault records were selected to verify the effectiveness of the fault record NER method and the fault processing decision-making method. These data were collected from a railway power maintenance administration department in South China, covering the period from 2019 to 2022. The dataset was randomly divided into a training set and a test set in a 4:1 ratio. The training set contains 730 fault records with 63,226 words, and the test set contains 182 fault records with 17,560 words. Table 4 summarizes the entity types and their counts in the dataset.

5.1.1 Effect of EIR of BERT-BILSTM-CRF model

Using the BERT-BiLSTM-CRF model to recognize the entity information of the dataset, the results in terms of different types of entity information are presented in Table 5.

From Table 5, it shows that, several entities can achieve a 100% harmonic mean, e.g., Time, Line, *etc.* It can be attributed to the relatively standard format and it is easy to extract features. Therefore, it achieves better precision. However, several entities experience low precision: e.g., Equipment, Phenomenon, Cause, and Measure. That is because the recorded information is not unified. Different forms and lengths of information makes it difficult for feature extraction, resulting in relatively low recognition precision.

5.1.2 Comparison of EIR performance with other models

Compare the proposed method with manual recognition, dictionary plus regular matching, and word2vec-BiLSTM-CRF model in terms of recognition precision and recognition speed. Perform EIR on 182 historical fault records in the test set, and the results are shown in Table 6.

It shows that in terms of F_1 value, the method proposed in this paper is the highest, reaching 94.66%, while the dictionary plus regular matching method is the lowest, only about 65%. This is because the format of fault records is complex and diverse, making it difficult to fully summarize the text format of fault records by

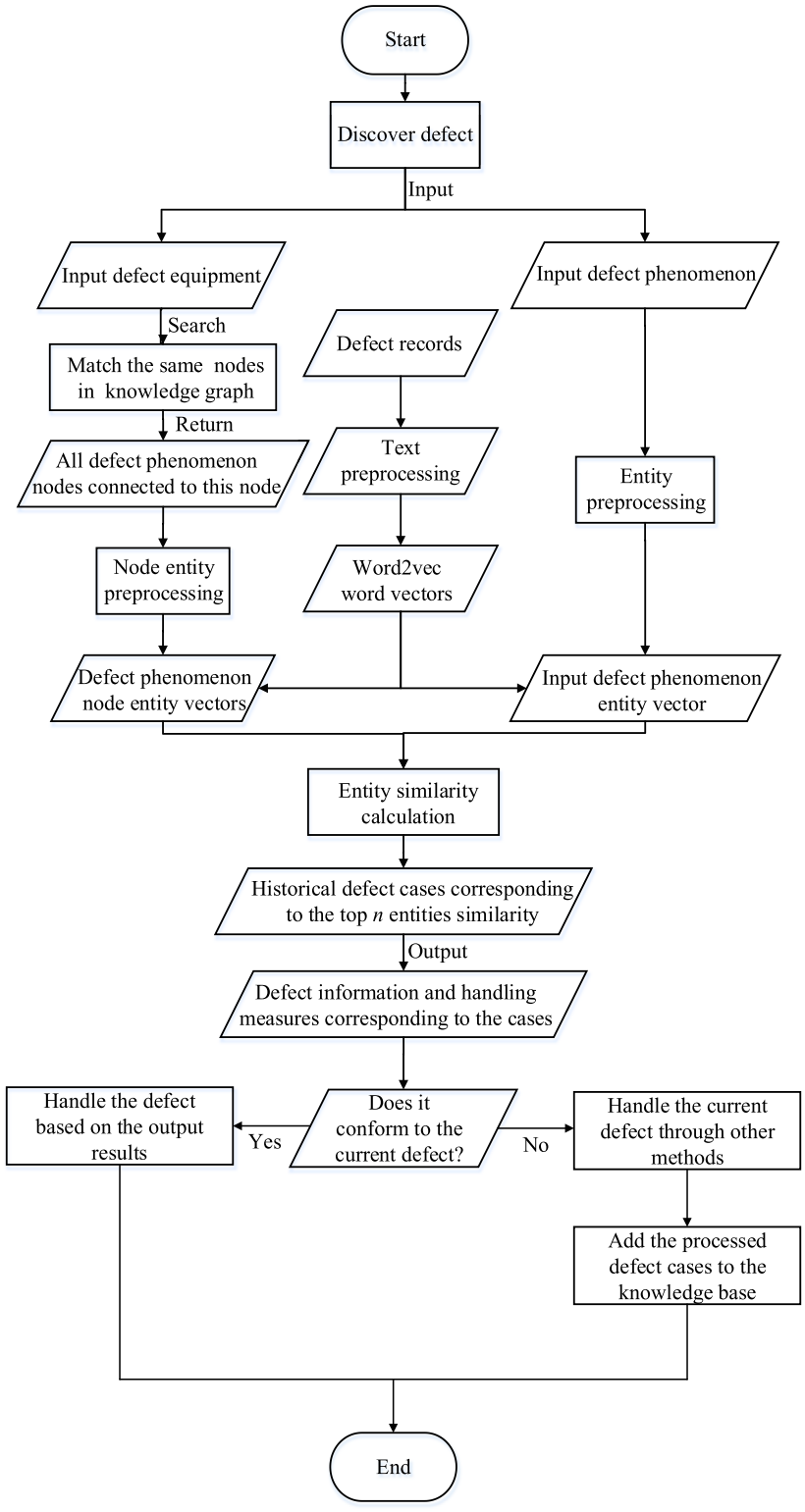

FIGURE 7
Process of decision-making method for TPSE fault handling.

exhaustion. Meanwhile, compared to the word2vec-BiLSTM-CRF model with the F_1 value of 91.92%, the method proposed in this paper achieves better EIR performance. That can be attributed to that the BERT model has trained a dynamic word vector including both the information *per se* and context semantic information, which avoids the problem of word semantic loss caused by the inability to consider the specific context when using word2vec word embedding model.

TABLE 4 Distribution of entity types and their frequency in the dataset.

Entity types	Examples	No. of entities in training set	No. of entities in test set
Time	7 January 2020, 17 February 2021, etc	1451	362
Line	XX Line	730	182
Substation	XX substation, XX section post, XX switching post, etc	730	182
Equipment	101DL, 214 protection device, Measurement and control device, etc	729	182
Phenomenon	Communication interruption, High control bus voltage, Remote operation rejection, etc	730	182
Cause	Loose bolts, Damaged insulation monitoring unit, Air switch tripping, etc	320	74
Type	Parts damage, Poor insulation, Poor contact, etc	730	182
Class	A (urgent), B (major), C (general)	730	182
Measure	Turn the air switch on, Uncover the switch, Turn the fuse on, etc	720	179
Overall	6870		1707

TABLE 5 Effect of EIR for different entity information.

Entity type	P/%	R/%	F_1 value/%
Time	100	100	100
Line	100	100	100
Substation	100	100	100
Equipment	84.24	85.16	84.70
Phenomenon	83.42	85.71	84.55
Cause	83.54	89.19	86.27
Type	100.00	100.00	100.00
Class	100.00	100.00	100.00
Measure	84.32	88.64	86.43
Overall	94.09	95.25	94.66

In terms of recognition speed, although the method proposed in this article is slightly slower than the two traditional methods, it saves 98.5% of time compared to manual recognition and can achieve automated and efficient processing of TPSE fault records. Due to the highest EIR precision of the proposed method, the EIR method based on the BERT-BiLSTM-CRF model is still the optimal choice.

To enhance clarity, the space complexity terms in Table 6 are further explained as follows. For the regular matching method, dictionary storage requires O(N × L), where N is the number of dictionary entries and L is the average entry length; matching operations require O(T), where T is the input text length. For the word2vec-BiLSTM-CRF model, word vector storage requires O(V × D), BiLSTM parameters O(H × (H + D)), and CRF components O(F × C + T × C), where V, D, H, F, and C represent vocabulary size, vector dimension, hidden layer size, number of features, and number of labels, respectively. The BERT-BiLSTM-CRF model includes BERT parameters O (L × H2+V × D), embedding storage O (T × D), BiLSTM and CRF components as above, leading to a higher but manageable complexity that balances performance with resource consumption.

5.2 Decision-making for TPSE fault handling

5.2.1 Knowledge graph of TPSE fault handling

Based on the presented BERT-BiLSTM-CRF EIR model, after extracting the entity information of 912 TPSE fault records, entities and relationships are inputted into Neo4j graph database for storage and representation, and the knowledge graph for TPSE fault handling is constructed, which includes 8646 nodes and 8645 relationships.

TABLE 6 Comparison of EIR effects of different methods.

Method	F_1 value/%	Average recognition time/s	Space complexity
Manual recognition	90	10	——
Dictionary plus regular matching	64.57	0.031	O(N × L) + O(T)
word2vec-BiLSTM-CRF	91.92	0.046	O(V × D) + O(H × (H × D)) + O (T × H) + O(F × C) + O (T × C)
BERT-BiLSTM-CRF	94.66	0.151	O (L × H^2 + V × D) + O (T × D)+ O(H × (H × D)) + O (T × H) + O(F × C) + O (T × C)

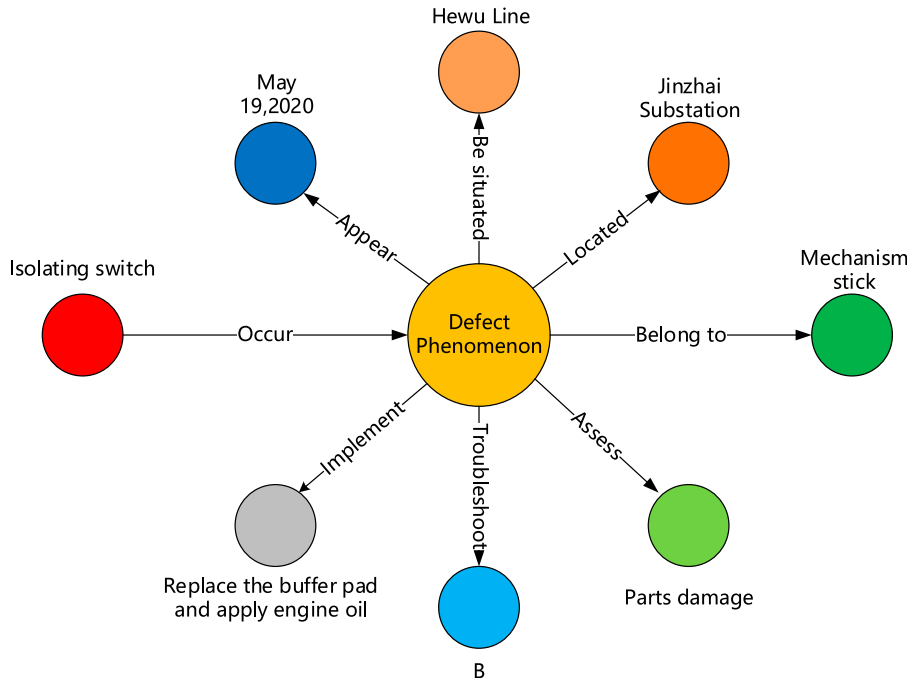

FIGURE 8
Knowledge graph of a historical fault case.

The fault case knowledge graph of isolating switch is displayed as shown in Figure 8. It can be seen that this knowledge graph contains a total of 9 nodes and 8 relationships, which corresponds to all fault information of this fault and the relationships between various entities. According to the knowledge graph of the fault case, the following information can be learned: the fault of the isolating switch occurred on *19 May 2020* in Jinzhai substation of *Hewu line*. The fault phenomenon is "refused to open". Through inspection, it is found that the cause of the fault is "mechanism stick", which belongs to the fault of *parts damage* type. The severity of the fault is class *B*. The handling measure is to "replace the buffer pad and apply engine oil".

5.2.2 Case analysis of decision-making process of isolation switch fault handling

Taking an isolation switch as an example, the feasibility of the decision-making method for fault handling is verified. As shown in Figure 9, when an isolation switch has a fault, the alarm signal is analyzed, and two key pieces of information—fault equipment and fault phenomenon—are extracted. These are matched in the TPSE fault handling knowledge graph as search conditions to output the top five most similar historical fault cases as references to assist operation and maintenance personnel in completing fault handling quickly and accurately. Finally, the details of the current fault are added to the knowledge graph after handling is completed, ensuring that the fault handling knowledge remains up to date.

Then, operation and maintenance personnel should refer to the fault causes in historical cases and inspect the current isolation switch in the following order: whether the air switch has tripped; whether the mechanism is stuck; whether the control power air switch inside the mechanism box is damaged; whether the wire on the back of the motor has fallen, preventing it from opening; and whether the opening contactor is damaged. If any of these conditions are confirmed, as shown in Figure 10, the current fault should be addressed using the corresponding historical handling measures.

Conversely, if inspection reveals that the fault cause of the isolation switch does not match any of the above conditions, maintenance personnel should refer to the maintenance regulations or consult experts to address the current fault.

5.2.3 Case analysis of decision-making process of circuit breaker fault handling

In addition, taking a circuit breaker as an example, when the fault equipment is identified as a circuit breaker and the fault phenomenon is "control circuit was broken," the top five most similar historical fault cases recommended by the knowledge graph are shown in Figure 11. It can be seen that the fault phenomenon corresponding to these five cases is identical to the input fault phenomenon after preprocessing, with both being "control circuit was broken." Thus, the entity similarity between the input fault phenomenon and each fault phenomenon node is 1.0. This indicates that for circuit breakers, "control circuit was broken" is a common fault phenomenon. However, examining the fault causes reveals that, although the fault phenomenon is identical, the underlying causes vary.

Then, operation and maintenance personnel should refer to the fault causes in historical cases and inspect the current circuit breaker in the following order: whether the opening coil is damaged with infinite resistance; whether the closing and locking electromagnet is

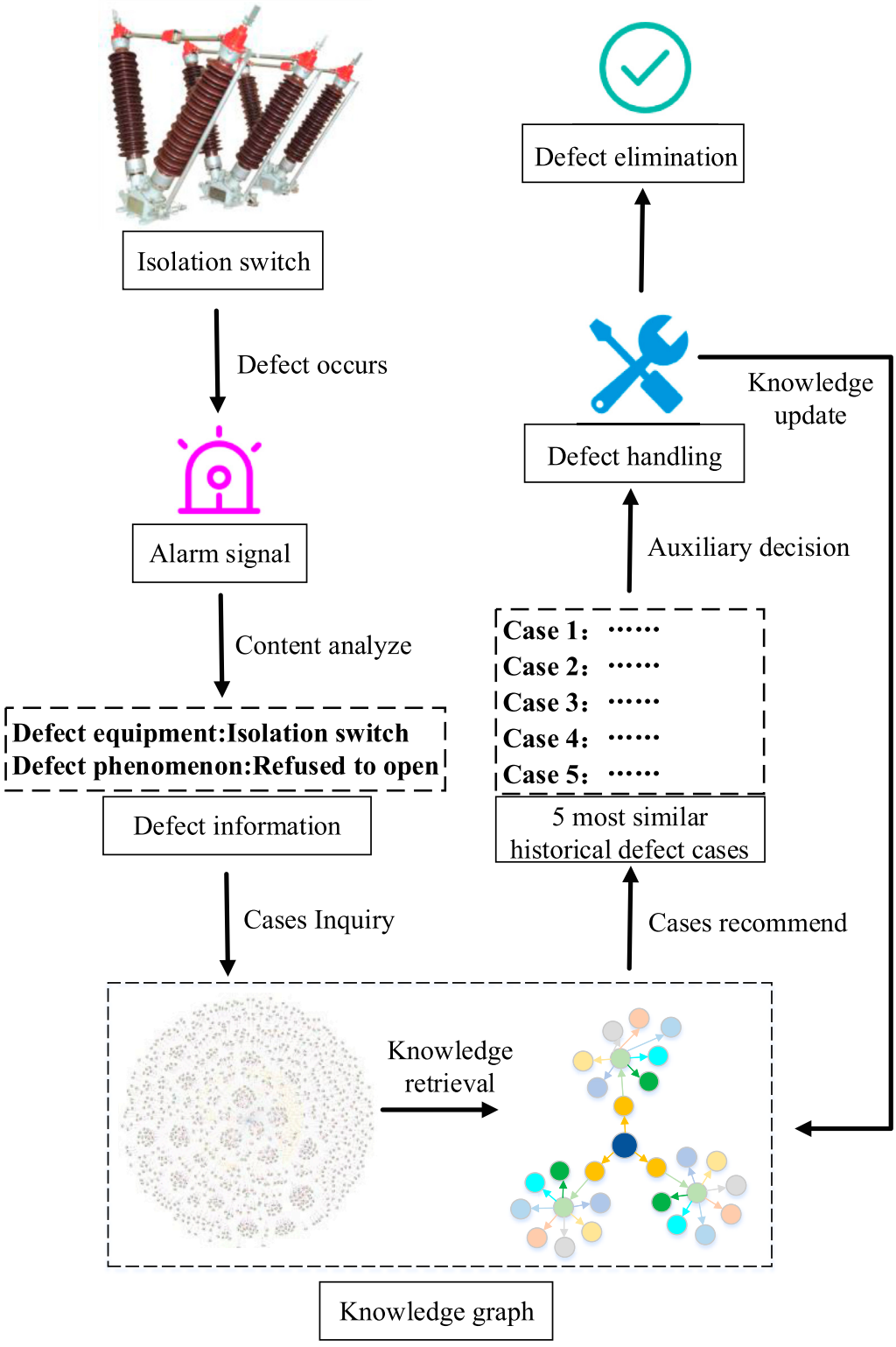

FIGURE 9
Example of decision-making process of isolation switch fault handling.

damaged; whether the opening coil is damaged; and whether the closing coil is damaged. If any of these causes are confirmed, the current fault should be addressed using the corresponding historical handling measures.

Conversely, if inspection reveals that the fault cause of the circuit breaker does not match any of the above causes, maintenance personnel should refer to the maintenance regulations or consult experts to address the current fault.

Auxiliary decision for traction power supply equipment defect handling

Query criteria:

Defect equipment: Isolation switch | Defect phenomenon: Refused to open | Query

Historical defect cases:

Entity similarity	Defect phenomenon	Defect cause	Handling measure
1.0	Refused to open	Air switch tripped	Closed the air switch
1.0	Refused to open	Mechanism stick	Replaced the buffer pad and apply engine oil
1.0	Refused to open	Control power air switch inside mechanism box is damaged	Replaced the air switch
0.9913	2901GK refused to open	A wire on the back of the motor fell off, causing it unable to open	Connected it
0.9898	2132GK refused to open	Opening contactor damaged	Replaced it

FIGURE 10
Historical fault cases of isolation switch recommended by Knowledge graph.

Auxiliary decision for traction power supply equipment defect handling

Query criteria:

Defect equipment: Circuit breaker | Defect phenomenon: Control circuit was broken | Query

Historical defect cases:

Entity similarity	Defect phenomenon	Defect cause	Handling measure
1.0	Control circuit was broken	/	Automatically recovered
1.0	Control circuit was broken	The opening coil of 211DL was damaged with infinite resistance	Replaced the coil and conduct characteristic test
1.0	Control circuit was broken	The closing and locking electromagnet of 214DL was damaged	Replaced it
1.0	Control circuit was broken	The opening coil of 212DL was damaged	Replaced it
1.0	Control circuit was broken	The closing coil of 216DL was damaged	Replaced it

FIGURE 11
Historical fault cases of circuit breaker recommended by knowledge graph.

6 Conclusion

This study proposes a data-driven framework for extracting and organizing fault information from traction power supply equipment (TPSE) records, aiming to enhance the efficiency and intelligence of fault diagnosis in railway systems. Based on the insights gained from extensive historical data, the following key conclusions can be drawn:

1. The integration of deep learning and structured knowledge modeling enables accurate and scalable fault information extraction. By combining BERT-BiLSTM-CRF for named entity recognition with a domain-specific BIO labeling scheme, the proposed method effectively captures complex fault attributes in unstructured text records. Empirical results show that the model achieves high precision and recall across multiple entity types, laying a solid foundation for the downstream construction of a TPSE knowledge graph.
2. The use of a knowledge graph coupled with entity similarity retrieval supports intelligent decision-making in fault handling. The structured fault knowledge graph allows for efficient storage and query of historical cases, while the similarity-based retrieval mechanism enables the recommendation of relevant prior solutions based on current fault features. This approach reduces reliance on manual searches, promotes knowledge reuse, and improves the consistency and timeliness of field-level maintenance actions.

7 Future work and outlook

Looking ahead, the proposed framework can be extended to support more intelligent and proactive maintenance applications within traction power supply systems. For real-time fault monitoring, on-site deployment of sensors such as surveillance cameras, infrared temperature detectors, and position sensors at key substations and equipment nodes may enable continuous state tracking of critical infrastructure. By evolving the current static knowledge graph into a dynamic, spatiotemporal matching mechanism, real-time equipment anomalies can be automatically correlated with historical fault patterns, allowing timely maintenance alerts and response suggestions based on past cases.

Furthermore, the original fault records used in this study are in Chinese. While the processing language is Chinese, the core unit of our algorithm is the "word," and Chinese words have a one-to-one correspondence with English words in the modeling framework. Therefore, the proposed method is equally applicable to English text, assuming proper segmentation and vectorization are performed. In future research, exploring the applicability of this framework to other languages—particularly those with different morphological or syntactic structures—will be an important direction for broadening its generalizability and robustness across multilingual environments.

In terms of predictive maintenance, future work may explore the application of temporal graph neural networks (TGNN) and frequency-based entity trend modeling to anticipate potential equipment failures. For example, faults in catenary components such as the messenger wire or cantilever may statistically precede specific pantograph issues. Mining such latent correlations will enable data-driven maintenance scheduling and risk prevention. Furthermore, with ongoing advancements in large language models (LLMs) and multimodal AI, future systems are expected to autonomously generate fault reports based on voice, imagery, or sensor data, thereby minimizing reliance on manual annotations. This trajectory points toward a gradual shift from AI-assisted tools to semi-autonomous maintenance agents, fundamentally reshaping field operations in railway infrastructure management.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

LP: Methodology, Software, Writing – review and editing, Writing – original draft. TX: Writing – original draft, Writing – review and editing, Investigation, Methodology. HZ:

Writing – original draft, Methodology, Conceptualization, Writing – review and editing. YZ: Writing – review and editing, Writing – original draft, Data curation, Investigation. YY: Investigation, Writing – review and editing, Writing – original draft, Methodology. WD: Conceptualization, Writing – review and editing, Methodology, Writing – original draft. ZD: Supervision, Investigation, Writing – original draft, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the Research Project of China Academy of Railway Sciences Corporation Limited under Grant 2023YJ048. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

Conflict of interest

Authors LP, TX, HZ, YZ, YY, and WD were employed by China Academy of Railway Sciences Corporation Limited.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Bo, W., Jian, N., Mei, D. Z., Yong, Z., Shan, X., and Changming, J. (2020). "Auxiliary decision technology and application of power grid fault disposal based on knowledge understanding of fault preplan," in *2020 5th International Conference on power and Renewable Energy (ICPRE)* (Shanghai, China), 246–251. doi:10.1109/icpre51194.2020.9233245

Dai, S., Ding, Y., Zhang, Z., Zuo, W., Huang, X., and Zhu, S. (2021). GrantExtractor: accurate grant support information extraction from Biomedical Fulltext based on Bi-LSTM-CRF. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 18 (1), 205–215. doi:10.1109/tcbb.2019.2939128

Gao, H., Miao, L., Liu, J., Dong, K., and Lin, X. (2020). "Construction and application of knowledge graph for power system dispatching," in *2020 7th International Forum on electrical engineering and automation (IFEEA)* (Hefei, China), 690–695. doi:10.1109/ifeea51475.2020.00147

Jacob, D., Chang, M., Kenton, L., and Kristina, T. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *ArXiv*, 4171–4186. abs/1810.04805. doi:10.18653/v1/n19-1423

Jahangirova, G., Clark, D., Harman, M., and Tonella, P. (2021). An Empirical validation of Oracle improvement. *IEEE Trans. Softw. Eng.* 47 (8), 1708–1728. doi:10.1109/tse.2019.2934409

Ji, S., Pan, S., Cambria, E., Marttinen, P., and Yu, P. S. (2022). A Survey on knowledge graphs: representation, Acquisition, and applications. *IEEE Trans. Neural Netw. Learn. Syst.* 33 (2), 494–514. doi:10.1109/tnnls.2021.3070843

Jin, H., Zhang, Y. F., Jia, Y. X., Yuan, Z. K., and Tu, Y. (2025). Condition diagnosis of composite insulator based on knowledge graph. *IEEE Trans. Dielectr. Electr. insulation* 32 (1), 28–35. doi:10.1109/tdei.2024.3510219

Lei, J., Tang, B., Lu, X., Gao, K., Jiang, M., and Xu, H. (2014). A comprehensive study of named entity recognition in Chinese clinical text. *J. Am. Med. Inf. Assoc.* 21 (5), 808–814. doi:10.1136/amiajnl-2013-002381

Lin, S., Shang, C., Li, N., Sun, X., Feng, D., and He, Z. (2023). An optimization method for maintenance resource Allocation in electrified railway catenary systems. *IEEE Trans. Industry Appl.* 59 (1), 641–651. doi:10.1109/tia.2022.3217107

Liu, N., Hu, Q., Xu, H., Xu, X., and Chen, M. (2022). Med-BERT: a Pretraining framework for Medical records named entity recognition. *IEEE Trans. Industrial Inf.* 18 (8), 5600–5608. doi:10.1109/tii.2021.3131180

Maggini, M., Marra, G., Melacci, S., and Zugarini, A. (2020). Learning in text Streams: Discovery and disambiguation of entity and relation instances. *IEEE Trans. Neural Netw. Learn. Syst.* 31 (11), 4475–4486. doi:10.1109/tnnls.2019.2955597

Menasce, D. A., and Gomaa, H. (2000). A method for design and performance modeling of client/server systems. *IEEE Trans. Softw. Eng.* 26 (11), 1066–1085. doi:10.1109/32.881718

Meng, Q., Song, Y., Mu, J., Lv, Y., Yang, J., Xu, L., et al. (2023). Electric power Audit text classification with multi-Grained pre-trained language model. *IEEE Access* 11, 13510–13518. doi:10.1109/access.2023.3240162

Morwal, S., Jahan, N., and Chopra, D.(2012). "Named entity recognition using hidden Markov model (HMM)", in *Int. J. Nat. Lang. Comput. (IJNLC)*, vol. 1, no. 4, pp. 15–23. doi:10.5121/ijnlc.2012.1402

Qiu, J., Wang, H., Ying, G., Zhang, B., Zou, G., and He, B. (2016). Text mining technique and application of lifecycle condition assessment for circuit breaker. *Automation Electr. Power Syst.* 40 (6), 107–112. doi:10.7500/aeps20150812003

Rudin, C., Waltz, D., Anderson, R. N., Boulanger, A., Salleb-Aouissi, A., Chow, M., et al. (2012). Machine learning for the New York city power grid. *IEEE Trans. Pattern Analysis Mach. Intell.* 34 (2), 328–345. doi:10.1109/tpami.2011.108

Sawant, U., Garg, S., Chakrabarti, S., and Ramakrishnan, G. (2019). Neural architecture for question answering using a knowledge graph and web corpus. *Inf. Retr. J.* 22 (3-4), 324–349. doi:10.1007/s10791-018-9348-8

Shang, Y., Tian, Y., Zhou, M., Zhou, T., Lyu, K., Wang, Z., et al. (2021). EHR-oriented knowledge graph system: toward efficient utilization of Non-used information Buried in Routine clinical practice. *IEEE J. Biomed. Health Inf.* 25 (7), 2463–2475. doi:10.1109/jbhi.2021.3085003

Stephen, B., Jiang, X., and McArthur, S. D. J. (2020). Extracting distribution network fault semantic labels from free text Incident Tickets. *IEEE Trans. Power Deliv.* 35 (3), 1610–1613. doi:10.1109/tpwrd.2019.2947784

Sun, H., Wang, Z., Wang, J., Huang, Z., Carrington, N., and Liao, J. (2016). Data-driven power outage detection by social sensors. *IEEE Trans. Smart Grid* 7 (5), 2516–2524. doi:10.1109/tsg.2016.2546181

Wang, H., Liu, Z., Xu, Y., Wei, X., and Wang, L. (2021). Short text mining framework with specific design for operation and maintenance of power equipment. *CSEE J. Power Energy Syst.* 7 (6), 1267–1277. doi:10.17775/cseejpes.2019.01120

Wang, H., Cao, J., and Lin, D. (2022). Deep analysis of power equipment faults based on semantic framework text mining Technology. *CSEE J. Power Energy Syst.* 8 (4), 1157–1164. doi:10.17775/cseejpes.2019.00210

Wang, L. B., Zhu, Z. B., and Zhao, X. F. (2024). Dynamic predictive maintenance strategy for system remaining useful life prediction via deep learning ensemble method. *Reliab. Eng. Syst. Saf.* 245, 110012. doi:10.1016/j.ress.2024.110012

Yang, H., Meng, X. K., Yu, H., Bai, Y., Han, Y., and Liu, Y. X. (2025). Research on primary equipment fault diagnosis method based on the BERT model. *Power Syst. Prot. Control* 53 (7), 155–164. doi:10.19783/j.cnki.pspc.240485

Yu, L. C., Wang, J., Lai, K. R., and Zhang, X. (2018). Refining word embeddings using Intensity scores for Sentiment analysis. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 26 (3), 671–681. doi:10.1109/taslp.2017.2788182