Research on the Construction of Fault Knowledge Graph for Wind Power Hydraulic Equipment

Aichun Du[1,2], Xiaoyu Ye[1,2], Xincai Chen[3], Zhengqiang Zhao[1,2], Hong Chen[1,2,4,*], Xiaoyun Gong[5]
[1]*Mechanical and Electrical Department, Hami Polytechnic, Hami, Xinjiang, China*
[2]*Hami Wind Power Equipment Intelligent Operation and Maintenance Engineering Technology Research Center, Hami, Xinjiang, China*
[3]*Zhengzhou Machinery Research Institute, Zhengzhou, Henan, China*
[4]*School of Mechanical and Power Engineering, Zhengzhou University, Zhengzhou, Henan, China*
[5]*School of Mechanical and Electrical Engineering, Zhengzhou University of Light Industry, Zhengzhou, Henan, China*
**Corresponding Author*

Abstract: Hydraulic systems are widely used in wind turbines. However, with the extension of the operation time, the failure rate of the hydraulic system gradually increases, which seriously drags down the operation efficiency of the wind turbine. Therefore, it is urgent to find a fast and accurate fault identification method to identify the abnormality of the wind turbine hydraulic system. The hydraulic system failure analysis documents collected over the past 15 years constitute a data set that has been processed by the BIO annotation technique to make it suitable for analysis. Based on the classic Bert-BILSTM-CRF entity recognition model, an optimized version is developed. Firstly, Bert model is used to collect and extract the features of scattered light; then, these features are fused with the output vector of BILSTM model; finally, the CRF model is used to complete the classification of labels. By embedding the strategy of adversarial learning in the BERT architecture, the robustness of the entity recognition model is successfully enhanced. Subsequently, we will analyze the overall architecture of the obtained triad information and save it in the Neo4j graph database to promote its adaptability. Finally, with the help of Python, we have created a system of fault knowledge mapping. The study ultimately revealed that the optimized model achieved a superior performance of 93 on the F1 score. Up by 1.8 percentage points. The proposed model exhibits an enhancement in performance of roughly 56 percentage points over the Bert-BILSTM-CRF model.

Keywords: Knowledge Graph; Troubleshooting; Entity Extraction; Neo4j

1. Introduction

Wind power is one of the new energy industries with the broadest prospects for current and future development, and is also one of the key new energy industries supported in China. China's wind power installed capacity has topped global rankings [1]. But renewable energy from wind turbines site environment is harsh, wind turbine working conditions are complex, wind turbines often because of various types of failure can not work properly. Hydraulic system is widely used in yaw system, pitch system and braking system of large wind turbine because of its advantages of smooth transmission, high power density and easy to realize stepless speed change. With the increase of service time of wind turbine, the wind turbine operation and maintenance is not timely, the environment is poor, the working condition is complex, the aging of components, etc., resulting in a significant increase in the failure rate of the hydraulic system, which seriously affects the normal operation of the wind turbine, and even lead to major safety accidents in serious cases. Pan Shujun analyzed the typical fault characteristics of the hydraulic transmission system of domestic wind turbines and proposed the transformation ideas and methods [1]. Leng et al. proposed a wind turbine hydraulic pitch system fault diagnosis method based on MISG algorithm [2], and Maldonado-Correa et al. reviewed the presence and development of diagnosis technology of fault of hydraulic system by big data analysis method [3]. Ortega-López et al. investigated the fault detection methodology for wind turbine hydraulic pitch systems using an adaptive observer and carried out simulation tests [4].

Although many useful studies have been carried out on the fault diagnosis of wind turbine hydraulic system, due to the hydraulic system failure factors, most of which are not easy to realize online monitoring and intelligent diagnosis, equipment maintenance and overhaul personnel often do not have a solid theoretical foundation and rich practical experience, it is difficult to find problems in time, resulting in hydraulic system failure in many cases can not be found in a timely manner and dealt with, which affects the normal operation of the wind turbine. Operation. The above problems can be effectively solved by data mining the key information recorded in the fault analysis of the hydraulic system and constructing the corresponding fault knowledge map.

In the field of constructing knowledge graph of mechanical failure, deep learning based named entity extraction method can mine semantic features according to the text contextual relationship, which has received wide attention in specific applications. Peng [5] parsed semi-structured data sources through knowledge acquisition, storage and application display to form knowledge information of system structure, which facilitates the completion of auxiliary analysis when system failure occurs. Su et al. [6] proposed a Lattice Transformer-CRF entity extraction method that integrates word and phrase sequence information for the construction of knowledge map of aero-engine faults, and the effect of this model has been improved to some extent. Pan and Wang [7] used BiLSTM-CRF model for entity extraction and relationship extraction of high-speed rail signaling equipment fault text data, which proved the usability of the model in the field of railroad equipment faults. In 2018, Google proposed a pre-training model BERT, which is based on the Transformer module to achieve bi-directional coding, and showed perfect results in the entity recognition task [8]. Hao et al. [9] fused BERT with BiLSTM-CRF model for knowledge extraction for a large amount of unstructured data in power distribution networks, which was visualized and managed by Neo4j graph database, which improved the decision-making efficiency of hydraulic system troubleshooters, and proved the good prospect of the knowledge graph technology in the field of equipment fault diagnosis.

However, the successful applications of the knowledge graph construction technology in the field of wind power hydraulic system fault diagnosis are few, and knowledge extraction accuracy of the existing entity extraction model is not enough.

In order to solve this problem, the strategy of entity extraction in hydraulic system fault diagnosis is discussed in this paper. Based on the integration of BERT and BILSTM-CRF model architecture, this paper proposes a series of improvements, which markedly enhances the model's entity recognition capabilities through the effective fusion of the BERT layer's and BILSTM layer's outputs. After constructing the ternary information model, a network hydraulic fault diagnosis platform is designed and implemented by using Neo4j graph database technology, which greatly enhances the speed and efficiency of fault diagnosis.

2. Fault Knowledge Graph Construction Process for Wind Power Hydraulic Equipment

There are two main ways to implement the knowledge graph technique, a top-down approach and a bottom-up strategy [10]. In the bottom-up strategy, we first identify the key entities and their associations from the text data, and then construct the corresponding knowledge structure at the high level; in contrast, in the top-down strategy, we first define the entities and associations at the high level, and then gradually go down to extract information from the specific text to form a complete knowledge map. In the process of establishing the knowledge structure map of a specific domain, clear knowledge boundaries and high specialization characteristics are observed, which leads to the persistence and stability of the entities involved in the map and their interrelated types [11]. Therefore, the top-down strategy is adopted, which shows the construction process of the fault knowledge map of the hydraulic system, as shown in the Figure 1.

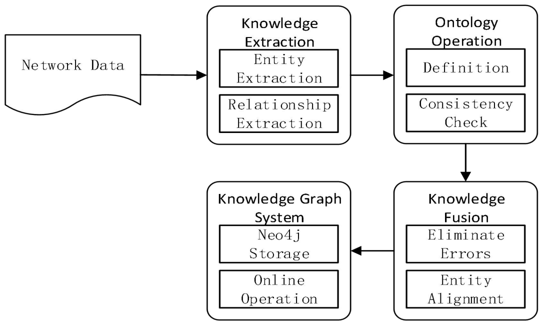

Figure 1. Fault Knowledge Graph Construction Process

3. BERT-BiLSTM-CRF Model

3.1 Overview of the Model

The Bert-BILSTM-CRF model involves three main parts: the first is the pre-training of word embedding based on Bert, the second is the in-depth analysis of sentence semantics by the bidirectional long-short term memory (LSTM) network, and the last is the accurate decoding of entity recognition through the conditional probability field model. As shown in the figure, the extraction model of hydraulic equipment fault text entity is displayed 2.

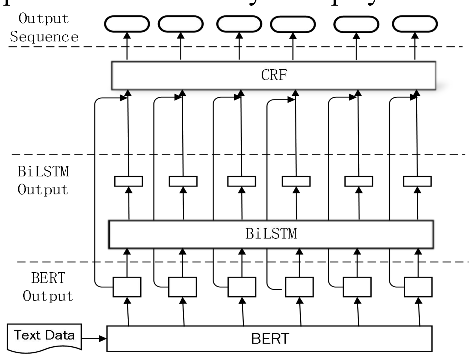

Figure 2. The entity extraction model of hydraulic equipment fault text

As shown in Figure 2, the entity recognition architecture based on the advanced Bert-BILSTM-CRF algorithm first encodes the character sequence in the incoming text with the help of the Transformer unit in the Bert pre-training model, so as to acquire the relative position and vocabulary vector of each character after training. These vectors are then used as the input data of the BILSTM bidirectional semantic extraction layer, and then the framework uses the BILSTM layer to perform bidirectional coding on the input vectors to extract the semantic vectors closely related to the text context; Finally, the series of processed output vectors are fused with the output results of the BILSTM layer, and then sent to the CRF layer for dimension reduction. By estimating the probability of different label sequences, the maximum probability value is used as the criterion to determine the optimal classification result, so as to obtain the category label that the entity should be classified into.

3.2 BERT Model

BERT is a profound bidirectional linguistic representation framework, whose function is to extract semantic and relational features from text, and transform them into word vectors, and then pass them to the next layer [12]. Thanks to the advantage of pre-training, it requires less effort in the downstream tasks of entity naming and recognition tasks, thus improving the recognition effect. In terms of entity extraction, Bert shows obvious advantages, and its structure is shown in the Figure. 3.

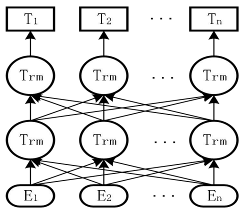

Figure 3. BERT Model

As shown in the Figure 3. The basic building block of the Bert model is the input level, with the Transformer encoding component in the middle of its architecture and the level responsible for producing the final output at the top. When exception information is accepted as input, the Bert architecture deconstructs it into a lexical embedding vector, a sentence vector, and a location identification vector to form input-level data processing, as shown in the Figure 4. The [CLS] tag marks the beginning of a document or sequence, while [Sep] marks the separation of a paragraph or element, or serves as a final identification point.

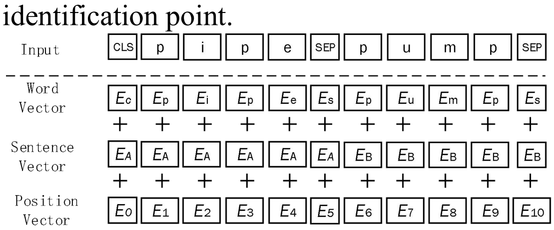

Figure 4. Input Layer of the BERT Model

Bert carries out pre-training through the masking strategy, using a unique marker [mask] to replace words or phrases, so that 15% of the sequence is in a random state: 80%

of them will be presented as [mask]; 10% of them will be replaced by random words; and the other 10% will remain unchanged.

In the Bert architecture, the Transformer encoder in the middle layer uses an algorithm called self-attention mechanism to evaluate the deep semantic connection between each word and the rest of its sentence. In this way, the model assigns a specific importance score to each word in the sentence, which in turn facilitates the refreshing and optimization of the word vector. The expression of the algorithm is as follows:

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

Where: Q, K, V is the word vector matrix, d_k is the Embedding dimension.

The BERT model projects Q, K, V through several different linear transformations by using a multi-attention mechanism which is based on the self-attention mechanism [13], and finally the different $Attention$ results are spliced and passed to the next layer using the $Concat$ function to obtain the final output, with the formulas as in Eq. (2) and Eq. (3):

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (2)$$

$$MultiHead(Q,K,V) = Concat(head_1, \ldots, head_n)W^O \quad (3)$$

Where, W is the weight matrix.

3.3 BiLSTM Model

Through the integration of BERT and Bidirectional Long-Short Time Memory Network (BILSTM), the precise capture of the deep meaning of the context of words is realized in the output layer. The BILSTM model combines forward and backward LSTM structures [14] with the aim of capturing more comprehensive semantic information. LSTM uses input gates, forget gates and output gates to regulate long-term and short-term memory, and then update and utilize information. A schematic of the LSTM architecture is shown in the Figure 5.

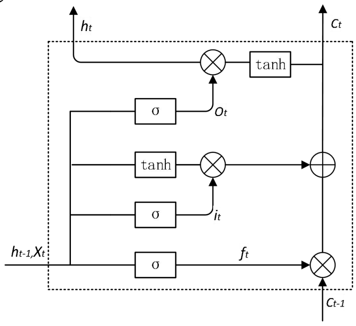

Figure 5. LSTM Model

The overall output of the structure is primarily derived by multiplying the output of the memory cell with the output of the output gate, as depicted in the following mathematical expressions:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (4)$$

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (5)$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (6)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (7)$$

$$h_t = o_t \cdot \tanh(c_t) \quad (8)$$

Where, W_i, W_f, W_o are the weight matrices of the input gate, oblivion gate, and output gate, respectively; b_i, b_f, b_o are the corresponding offset matrices, x, h The inputs and outputs, respectively. σ is Sigmoid activation function, $\tanh$ is the hyperbolic tangent activation function, The final outputs of the BILSTM model are forward and backward outputs h_t put together ,Obtaining contextual information in text enables two-way learning of information.

3.4 CRF Forecasting Model

The sequence labeling problem is a challenge within the domain of computational linguistics, because the traditional Bert and BILSTM models fail to effectively capture the order of the output sequence. To this end, researchers proposed the CRF (Conditional Random Fields) model, which, with its SoftMax classifier, can better characterize the interdependence between labels, especially when dealing with continuous text labels, significantly reduce the prediction probability of misordered labels, thus improving the overall prediction accuracy [15].

Set the input sequence $X = \{x_1, x_2, \ldots, x_n\}$, its prediction sequence $Y = \{y_1, y_2, \ldots, y_n\}$ conditional probability $P(y|x)$, the formula is:

$$P(y|x) = \frac{1}{Z(x)} \exp \sum_{k=1}^{K} W_k f_k(y,x) \quad (9)$$

$$Z(x) = \sum_y \exp \sum_k^K w_k f_k(y,x) \quad (10)$$

Where: f_k is the characteristic function; w_k is the weight of f_k; $Z(x)$ is the normalization term.

The Conditional Random Field (CRF) framework is refined by using the maximum likelihood estimation method, aiming at estimating the probability of the prediction sequence Y of the hydraulic system fault description in the text sequence B.:

$$P(y|B) = \frac{\exp\,(score(B,y))}{\sum_{\tilde{y}\epsilon y_B} \exp\,(B,y)} \quad (11)$$

$$L = log\,(P(y|B)) \quad (12)$$

Where $\tilde{y}$ is the true label; y_B is the set of all labels; score is the score of the degree of correspondence between the faulty text data B and the predicted sequence y; L and is the loss function.

In the CRF model for faulty text data prediction, the viterbi algorithm is used to find the global optimal solution, which is formulated as:

$$y^* = arg \max_{\tilde{y} \epsilon y_B} score\,(B, y) \tag{13}$$

Where: y^* is the sequence of hydraulic system fault text labels with the maximum calculated score.

4 Experimental Results and Analysis

4.1 Labeling of Data Sets

In the past 15 years, a special text database for fault analysis has been constructed by using the data of fault and failure analysis. The information comes from extensive browsing of monographs, case studies and related literature in the field of hydraulic fault diagnosis. The data set is initially processed using a bioinformatics annotation technique in which B identifies the beginning of the entity, I refers to the continuation or end of the entity, and O refers to the non-entity segment. The documents related to hydraulic machinery failure cover eight main contents: hydraulic device, failure manifestation, performance system, failure cause, failure consequence, repair suggestion, repair log and the time point of failure occurrence, totaling about 30,000 words. These data are divided into training set, test set and validation set according to the ratio of 8:1:1, and the labeled instances of various entities are shown in the Figure 6.

Word	Label	Word	Label	Word	Label	Word	Label	Word	Label
M	B-time	W	B-structure	n	I-structure	c	I-system	l	I-system
a	I-time	i	I-structure	e	I-structure	a	I-system		I-system
r	I-time	n	I-structure			l	I-system	s	I-system
c	I-time	d	I-structure	e	B-system		I-system	y	I-system
h	I-time		I-structure	l	I-system	c	I-system	s	I-system
	I-time	t	I-structure	e	I-system	o	I-system	t	I-system
2	I-time	u	I-structure	c	I-system	n	I-system	e	I-system
0	I-time	r	I-structure	t	I-system	t	I-system	m	I-system
2	I-time	b	I-structure	r	I-system	r	I-system		
0	I-time	i	I-structure	i	I-system	o	I-system		

Figure 6. Example of Entity Labeling

4.2 Experimental Conditions

The experimental environment is set as follows: Windows operating system and Python 3 programming language are used. It uses seven programming languages, is equipped with AMD Ryzen 7 4800H high-performance processor, and integrates Keras deep learning architecture.

When dealing with the data structure, we implement the following strategies: (1) Adjust the learning rate: In view of the significant difference in the learning rate requirements between BERT and BILSTM-CRF models, we adopt a double-layer classification strategy to meet their respective specific needs, thus improving the training efficiency; Moreover, when the training process does not reach the predetermined number of times and the model F1 score is stagnant, we adopt the strategy of decreasing learning rate in order to enhance the performance of the model. Another technique employed to boost model robustness is the introduction of the Fast Gradient Symbolic Method (FGM) in adversarial training. In the Bert architecture, the gradient and fluctuation of the merged level are calculated in the merged level, and then the fluctuation is injected and replaced. The gradient descent method is performed to modify the original training process, and then the modified training process is given to the model for adversarial training [16]. As shown in Table 1, the detailed model configuration information is presented.

Table 1. Parameters for Model Improvement

Parameter	Value
Lstm Units	128
Bert Learning Rate	5e-5
Attenuation Rate	0.1
Layered Rate	500
Disturbance	0.5
Epoch	15
Batch size	16
Dropout	0.1

4.3 Validation of the Effectiveness of the Improved Model

Evaluation metrics for deep learning models use F1 values

$$P = \frac{A}{B} \times 100\% \tag{14}$$

$$R = \frac{A}{C} \times 100\% \tag{15}$$

$$F_1 = \frac{2PR}{P+R} \quad (16)$$

Where: Ais the number of correctly recognized entities; B is the number of recognized entities; C is the total number of entities; P is the accuracy rate; R is recall rate. The specific model parameters are shown in the following Table 2.

BiLSTM-CRF, LSTM-CRF, BERT-BILSTM-CRF and the improved model are taken to compare the effect of entity extraction in the hydraulic system failure dataset.

Table 2. Comparison of the Effect of Different Models

Model	P	R	F1
LSTM-CRF	0.8157	0.7871	0.8011
BILSTM-CRF	0.8789	0.8495	0.8639
BERT-BiLSTM-CRF	0.9254	0.9022	0.9137
This Model	0.9412	0.9384	0.9393

The accuracy, return rate and F1 value bar effects of the improved model according to the division of the extracted entity classes are shown in Figure 7.

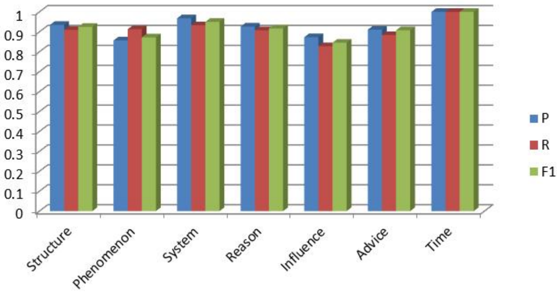

Figure 7. Entity Recognition Effect

Refer to Figure 7. This study successfully improved the performance of the Bert-BILSTM-CRF model. By combining the output of Bert and BILSTM, this improvement strategy fully integrates the semantic features learned by different models [17]. Meanwhile, the hierarchical training strategy optimizes the learning efficiency, ensures that the learning rate of each model is appropriate, and enhances the accuracy of the model by reducing the learning rate.

5. Knowledge Graph System Development

5.1 Knowledge Graph Visualization

Neo4j is a graph database composed of nodes, relationships and attributes. Compared with relational database, graph database can directly express the associated attributes of data, and can store a large number of heterogeneous data from multiple sources without entity and mapping relationship tables, which saves storage space [18].

The entities and associated information are imported into the Neo4j graph database and displayed in the node relationship diagram. Figure 8 shows the visualization of the knowledge graph.

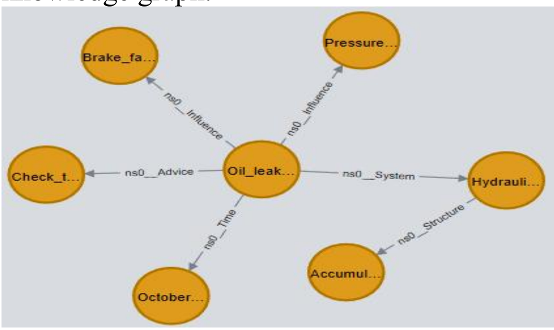

Figure 8. Example of Knowledge Graph Visualization

5.2 Web OS Development and Applications

Knowledge mapping plays an important role in recording detailed structured data of hydraulic system faults, which helps professional technicians to carry out fault analysis and maintenance quickly and effectively, thus greatly improving the efficiency of fault handling. However, since the Neo4j database is operated by the Cypher language, it is not convenient to translate the expert's knowledge directly into a knowledge base. Therefore, we choose to use Python programming language and Django framework, with Neo4j as the background database, to create a Web front-end human-computer interaction system, so that technicians can more easily map the knowledge of operating system failures [19]. Depending on the needs of different operators, we provide different levels of open access. The specific login interface is shown in Figure 9.

Figure 9. Knowledge Mapping System Login

Taking the fault of the hydraulic brake system of the wind turbine as an example, the speed of fault detection is enhanced with the help of relevant fault sources and maintenance strategies, and the diagnostic effect is shown in the Figure 10. Abnormal noise or an increase in system temperature when the brakes are applied may be a sign of a problem. By inputting these signals in the knowledge structure map, a specific fault condition can be determined, as in the image shown Figure 11.

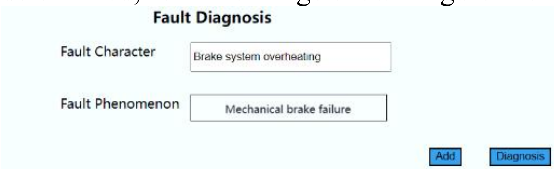

Figure 10. Knowledge Mapping System Troubleshooting

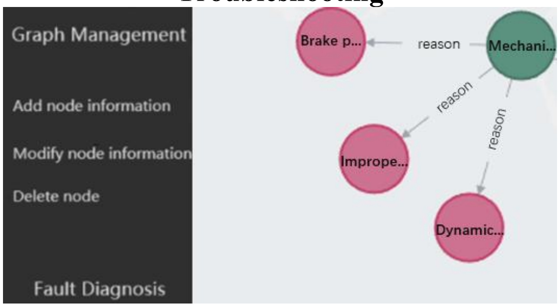

Figure 11. Diagnostic Results of the Knowledge Mapping System

By analyzing the potential risk parameters caused by hydraulic failure, the purpose and steps of troubleshooting are defined, which is helpful for maintenance personnel to make wise choices.

6. Conclusions

In this study, bioinformatics-based preprocessing methods were used to analyze the failure of wind turbine hydraulic system in the past 15 years. The standard Bert-BILSTM-CRF entity detection model is optimized, and the effectiveness of the optimization is verified by P, R and F1 evaluation criteria.

1) First, we use the Bert model to extract the deep semantic features of the text, and then integrate the output vector with the output vector of the BILSTM model, and finally achieve the task of text classification through the CRF model. Adversarial training is introduced in the Bert layer to improve the robustness of the entity recognition model. The improved model accuracy achieved an F1 score of 93. The proportion increased by 2 percentage points to 93%.56% improvement compared to the original Bert-BILSTM-CRF model.

2) The collected data structure is analyzed, and the analysis results are saved in the Neo4j graph database. The knowledge map system of wind turbine hydraulic fault based on Python is constructed, which lays a theoretical foundation for exploring the fault analysis of wind turbine transmission system in the future. The optimized Bert-BILSTM-CRF model can accurately identify the abnormality of the hydraulic system, which provides a new direction for the fault prediction and diagnosis research of the wind turbine transmission system based on the model in the future.

References

[1] Cabrera P, Carta J A, Matos C, et. al. Lessons learned in wind-driven desalination systems in the canary islands: Useful knowledge for other world islands. Desalination, 2024, 583.

[2] Leng D, You S, Wang R, et. al. Fatigue damage mitigation of monopile offshore wind turbines utilizing a rotational inertia double tuned mass damper (ridtmd) under realistic wind-wave conditions. Ocean Engineering, 2024, 305.

[3] Maldonado-Correa J, Torres-Cabrera J, Martín-Martínez S, et. al. Wind turbine fault detection based on the transformer model using scada data. Engineering Failure Analysis, 2024, 162.

[4] Ortega-López V, Faleschini F, Hurtado-Alonso N, et. al. Analysis of raw-crushed wind-turbine blade as an overall concrete addition: Stress–strain and deflection performance effects. Composite Structures, 2024, 340.

[5] Peng F Z. Impedance sources (z sources) with inherent fault protection for resilient and fire-free electricity grids. Scientific Reports, 2024, 14.

[6] Su H, Cao Q, Li Y. Condition-based opportunistic maintenance strategy for multi-component wind turbines by using stochastic differential equations. Scientific Reports, 2024, 14.

[7] Pan S J, Wang J B. Failure analysis and technical transformation of hydraulic transmission system of domestic wind turbine. Inner Mongolia Electric Power Technology, 2011, 29(6):60-62.

[8] Huang W T, Wen C B. MISG algorithm based wind turbine hydraulic pitch system fault diagnosis. Electrical Machines and Control Applications, 2018, 45(11):110-115.

[9] Hao J., Gao H. N., Wang F., et al. A review of hydraulic system fault diagnosis based on big data analysis. Machine Tools and Hydraulics, 2022, 50(10):166-173.
[10] Hu C X, Wen C B. Fault diagnosis of wind turbine hydraulic pitch system based on adaptive observer. Electrical Engineering and Electricity, 2019, 11:5-10.
[11] Li X P, Xu J H, Guo Z M, et al. Construction and application of knowledge graph of dispatching automation system. China Electric Power, 2019, 52(02):70-77+157.
[12] Han T, Huang H S, Yao L G. Entity extraction for aero-engine fault knowledge graph construction. Combined Machine Tools and Automated Machining Technology, 2021(10):69-73+78.
[13] Li X Q, Shi T Y, Li P, et al. Research on text-based knowledge extraction method of high-speed railroad signaling equipment faults. Journal of Railway, 2021, 43(03):92-100.
[14] Li X Y, Zhang H, Zhou X H. Chinese clinical named entity recognition with variant neural structures based on BERT methods. Journal of Biomedical Informatics, 2020, 107(prepublished).
[15] Ye X Z, Shang L, Dong X Z, et al. Research and application of knowledge graph for fault disposal in distribution networks. Grid Technology, 2022, 46(10):3739-3749.
[16] Sun J D, Gu X S, Li Y, et al. Research on Chinese entity relation extraction algorithm based on COAE2016 dataset. Journal of Shandong University (Science Edition), 2017, 52(9):7-12+18.
[17] Liu T, Zhang S R, Wang C, et al. BERT-BiLSTM hybrid model for intelligent analysis of water conservancy construction accident text. Journal of Hydropower Generation, 2022, 41(07):1-12.
[18] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation, 1997, 9(8):1735-1780.
[19] Gong Faming, Li Hastiness. Research on massive petroleum domain ontology data storage based on Neo4j. Computer Science, 2018, 45(S1):549-554.