An Effective Knowledge Mining Method for Compressor Fault Text Data Based on Large Language Model

Xiaodong Qin
National Engineering Laboratory for Pipeline Safety
MOE Key Laboratory of Petroleum Engineering
Beijing Key Laboratory of Urban Oil and Gas Distribution Technology
China University of Petroleum-Beijing
Beijing, 102249, China

Yuxuan He
National Engineering Laboratory for Pipeline Safety
MOE Key Laboratory of Petroleum Engineering
Beijing Key Laboratory of Urban Oil and Gas Distribution Technology
China University of Petroleum-Beijing
Beijing, 102249, China

Jie Ma
National Engineering Laboratory for Pipeline Safety
MOE Key Laboratory of Petroleum Engineering
Beijing Key Laboratory of Urban Oil and Gas Distribution Technology
China University of Petroleum-Beijing
Beijing, 102249, China

Weiyuan Peng
National Engineering Laboratory for Pipeline Safety
MOE Key Laboratory of Petroleum Engineering
Beijing Key Laboratory of Urban Oil and Gas Distribution Technology
China University of Petroleum-Beijing
Beijing, 102249, China

Enrico Zio
Dipartimento di Energia
Politecnico di Milano
La Masa 34, 20156, Milano, Italy
MINES ParisTech PSL Research University
CRC Sophia Antipolis, France

Huai Su*
National Engineering Laboratory for Pipeline Safety
MOE Key Laboratory of Petroleum Engineering
Beijing Key Laboratory of Urban Oil and Gas Distribution Technology
China University of Petroleum-Beijing
Beijing, 102249, China
*Corresponding Author Email: suhuai@cup.edu.cn

***Abstract*—The fault diagnosis method of compressors determines the reliability of the gas transmission pipeline station. Existing compressor fault diagnosis methods mostly relies on data-driven, which leads to a high application threshold from the mechanism. To address this issue, this paper introduces the knowledge graph into the compressor fault diagnosis for the first time and proposes a compressor fault text data knowledge mining method based on large language model. Firstly, the characteristics and principles of compressor faults are analyzed. Then, a text data knowledge mining model called CFRTE for compressors is constructed. Experimental results show that the *F*1 score of the CFRTE model can reach 0.98, meeting the requirements of compressor fault knowledge mining. Finally, combined with the results of knowledge mining and the graph database, a new system for the storage and indexing of the compressor fault knowledge graph is proposed. To further verify the role of the large language model in compressor fault knowledge mining, this paper conducts a comparative experiment of CFRTE models based on RNN encoder and BERT encoder. Experimental results show that compared with GRU, BiGRU, LSTM, and BiLSTM as the encoder layer, the *F*1 score of the CFRTE model with BERT as the encoder layer has increased by 26.78%, 6.18%, 21.89%, and 5.49% respectively. This work provides a systematic feasible scheme for introducing knowledge graphs into compressor fault diagnosis, which can be used for reference in the fault diagnosis of related equipment.**

Keywords—compressor station; compressor; large language model; knowledge mining; knowledge graph

I. Introduction

Compressors are key equipment to ensure the delivery of natural gas, and guaranteeing their safe and stable operation is of paramount importance. During the long-term operation of compressors, various failures are bound to occur. Therefore, accurate diagnosis of compressor failures is of great significance in ensuring the safe operation of the gas pipeline system. As the industrial sector continues to progress, the complexity of compressors is increasing along with the volume of monitoring data, propelling the advancement of intelligent and sophisticated fault diagnostic approaches, which have progressed into a data-driven fault diagnosis phase. For example, Milad Golmoradi[1]et al. proposed a compressor fault diagnosis method based on support vector machines and genetic algorithms. Experimental results showed that this method has a high safety margin when used for multi-stage centrifugal compressors. Van Gompel Jonas[2]et al. proposed a fault diagnosis model for photovoltaic systems based on Graph Neural Network (GNN). The research results show that GNN can accurately diagnose faults whether or not there is weather data.

979-8-3503-7143-7/23/$31.00 ©2023 IEEE
DOI 10.1109/CSAT61646.2023.00024

However, purely data-driven compressor fault diagnosis methods have shortcomings, namely they highly depend on the quality of fault data, but the cost of obtaining this data in actual projects is often very high, especially for data with small sample characteristics. Human experience, as a highly condensed form of data, becomes a feasible angle to address the above-mentioned shortcoming. The text data, such as maintenance records produced by the compressor in the design, production, and operation processes, contains human experience in compressor fault diagnosis, that is, "knowledge" in the field of compressor faults. Using this text data as a data source can compensate for the high threshold of obtaining quality fault data, but there are the following difficulties in extracting and utilizing compressor fault knowledge from unstructured text data: (1) Data diversity: Unstructured text data may come from various sources, and the formats, languages, style differences are substantial, and there may also be some errors, making data integration and normalization complicated; (2) Key information extraction: The vocabulary and sentences in text data may have different meanings, requiring the development of natural language processing (NLP) algorithms to accurately extract key compressor fault information from large amounts of unstructured text data.

Compressor fault knowledge is often represented in the form of triples, that is, in the form of (head entity, relationship, tail entity). The target of compressor fault knowledge mining is to extract structured compressor fault knowledge triples from unstructured text data. In traditional methods, rule-based methods are often used, matching compressor fault knowledge from unstructured text data by artificially constructing some rules, such as matching through regular expressions. However, these methods require a lot of manual work to define rules and have limited performance when dealing with complex natural language expressions. In recent years, there have been significant advances in the field of natural language processing, the most famous of which is large language model, especially ChatGPT. It can accurately understand the semantic information contained in the text and interact with humans in conversations. The powerful language understanding ability of large language models can significantly improve the performance of downstream natural language processing tasks such as entity recognition and relation extraction, providing a new idea for the compressor fault knowledge mining research in this paper. Therefore, this paper constructs a compressor fault knowledge mining model based on the large language model, which can mine structured compressor fault knowledge triples from unstructured compressor fault text data quickly and accurately for the compressor fault knowledge graph.

At present, in the oil and gas field, the research on introducing knowledge graphs into compressor fault diagnosis needs to be further deepened. In response to the above issues, this paper proposes a compressor fault knowledge mining method based on the large language model. The main innovations include: (1) For non-structured fault text data of compressor faults, a professional knowledge mining method for compressor fault diagnosis based on a large language model (CFRTE) is proposed for the first time, which broadens the data source of compressor fault diagnosis; (2) A knowledge mining method for compressor fault text data coupled with a double-verification mechanism is proposed for the first time, knowledge graph construction accuracy can reach 98.9%.

II. METHODOLOGY

This paper proposes a compressor fault knowledge mining method based on a large language model. The overall framework is mainly divided into two parts: data preprocessing and knowledge mining, as shown in Fig 1.

A. *Construction of Compressor Fault Knowledge Ontology*

Before performing compressor fault knowledge mining, the knowledge mining norms need to be clarified, that is, to construct the concept ontology of compressor fault knowledge. The concept ontology of compressor fault knowledge is a standardized description of the concepts and relationships between concepts in the field of compressor faults. It determines the conceptual nodes and associated relationships in the knowledge graph, and is the basis and template for constructing the knowledge graph. This paper analyzes the compressor fault text data, summarizes the principles and characteristics of seven compressor fault types, and combines relevant expert experience to model the concept ontology of compressor fault knowledge, as shown in Fig 2.

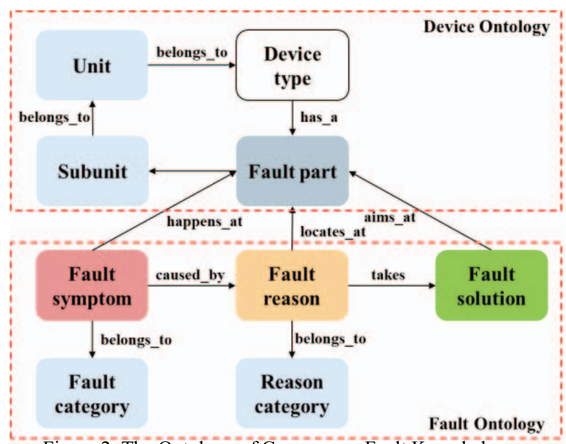

Figure 2. The Ontology of Compressor Fault Knowledge.

B. *Compressor Fault Knowledge Mining Based on Large Language Model*

In 2017, the introduction of the Transformers[3] model marked the beginning of large language models. The encoder layer of this model, known as BERT[4], has achieved significant effectiveness in various natural language processing tasks. BERT is mainly composed of multiple stacked Encoder modules of Transformers, with the self-attention mechanism as its core, enabling the model to understand the contextual meaning of text and automatically learn language representations. The formula for the attention mechanism is shown in Equation (1).

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

In the equation, X represents the input matrix, and Q, K, and V are obtained by multiplying the input matrix by the corresponding weights. dk is the dimension of K, and Attention represents the attention coefficient. The BERT model is based on the self-attention mechanism and learns the vector representation of words in an unsupervised manner. The above equation only shows the core idea of the BERT model. For more information about the BERT model, please refer to references [3] and [4].

In recent years, many researchers have proposed their own relationship extraction models based on the BERT model, such as CasRel[5] and GRTE[6] and so on. Among them, the BiRTE[7] model, with its characteristic of bidirectional relationship extraction, has the potential to recognize relationships in the text more comprehensively. In this paper, based on the pre-training of the large language model BERT, we fine-tuned BERT to construct a model called CFRTE (Compressor Fault Relation Triple Extraction) with BiRTE structure for mining compressor fault knowledge, with the compressor fault knowledge ontology as the standard. The structure of CFRTE model is shown in Fig 3.

(1) Encoder Layer: The encoder layer utilizes the pre-trained BERT model to transform the text into vector representations. The encoded text is then passed through three fully connected layers to obtain text features for extracting the head entity, relation, and tail entity.

(2) Entity Pair Recognition Layer: The entity pair recognition layer identifies entities based on the text features obtained from the encoder layer from two directions. The subject recognizer and the object recognizer both adopts a binary classification approach to predict the probability of whether a certain word in the text is the starting or ending position of the head and tail entity, that is:

$$\begin{aligned} p_s^{i,start} &= Sigmoid(W_s^{start} h_s^i + b_s^{start}),\ p_s^{i,end} = Sigmoid(W_s^{end} h_s^i + b_s^{end}) \\ p_o^{i,start} &= Sigmoid(W_o^{start} (h_o^i \odot v_s^{s\_k}) + b_o^{start}),\ p_o^{i,end} = Sigmoid(W_o^{end} (h_o^i \odot v_s^{s\_k}) + b_o^{end}) \end{aligned} \quad (2)$$

In the equation, $p_s^{i,start}$, $p_s^{i,end}$, $p_o^{i,start}$ and $p_o^{i,end}$ respectively represent the probability that the i-th word in the text is the start and end position of the head entity or tail entity. h_s^i and h_o^i represents the text features through the encoder layer. $v_s^{s\_k}$ represents the features of the head entity extracted through max pooling. W_s^{start}, W_s^{end}, W_o^{start} and W_o^{end} are trainable weights; b_s^{start}, b_s^{end}, b_o^{start} and b_o^{end} are trainable biases; $\odot$ represents the element-wise product of two vectors. If the probability predicted by the model is greater than the set threshold, then it is confirmed that the word is the start or end position of the head or tail entity.

(3) Relationship classification layer: The relationship classification layer classifies the relationships of the entity pairs obtained by the entity extraction layer. Combining the features of the entity pair and the text features, the relationship classification layer uses a double affine transformation method to predict the probability that the two entities have a certain relationship, that is:

$$p_r^i = Sigmoid(\begin{bmatrix} v_r^{s\_k} \\ 1 \end{bmatrix}^T W_r^i \begin{bmatrix} v_r^{o\_j} \\ 1 \end{bmatrix}) \quad (3)$$

In the equation, p_r^i represents the probability that the k-th head entity and the j-th tail entity have the i-th relation. $v_r^{s\_k}$ and $v_r^{o\_j}$ respectively represent the features of the head entity and the tail entity obtained through max pooling. W_r^i is a trainable weight. If the predicted probability by the model is greater than the set threshold, it indicates that the k-th entity and the j-th entity have the i-th relation, resulting in the formation of a knowledge triple.

III. CASE STUDY

A. Data Foundation

The data used in this paper consists of maintenance records of gas pipeline compressor failures, totaling 4,860 maintenance record entries. With techniques such as text segmentation and data augmentation, a total of 83,000 compressor fault knowledge text data were obtained, which were then divided into training, validation, and testing sets in an 8:1:1 ratio.

B. Model Selection for Knowledge Mining

In order to select the optimal model for compressor fault knowledge extraction task, this paper conducts comparative experiments with five models, namely SpERT[8], Casrel[5], OneRel[9], PRGC[10] and GRTE[6], as well as the CFRTE model. The evaluation metrics used for comparison are commonly used in the field of deep learning, including precision, recall, and $F1$ score. The CFRTE model is constructed based on the PyTorch deep learning framework. After the model training is completed, the models are tested on the test set, and the precision, recall, and $F1$ score of each model on the test set are calculated, as shown in Table 1.

TABLE I. THE PRECISION, RECALL, AND $F1$ SCORE OF THE TRAINED MODEL ON THE TEST SET.

Model	Precision	Recall	$F1$ score
SpERT	0.786	0.793	0.789
CasRel	0.878	0.896	0.887
OneRel	0.928	0.893	0.910
PRGC	0.937	0.924	0.930
GRTE	0.957	0.932	0.944
CFRTE	0.989	0.971	0.980

According to the table, it can be seen that compared to other models, the CFRTE model has the highest precision, recall, and $F1$ score, reaching an $F1$ score of 0.980 on the test set. This demonstrates that the model possesses high accuracy and adaptability, making it suitable for compressor fault knowledge extraction tasks.

C. Comparison Experiment for Encoder Layer

To verify the performance improvement of large language models in compressing compressor fault knowledge extraction models, this study selected the two most common RNN variants: GRU and LSTM. Comparative experiments were between the CFRTE model based on BERT encoder and the CFRTE model based on GRU and LSTM encoder. In the comparative experiments, only the encoder layer of the CFRTE model was replaced with GRU and LSTM, while keeping the model's other structures, training data, and hyperparameters unchanged. The results on the test set after model training are shown in Table 2.

TABLE II. THE COMPARISON OF CFRTEBERT AND CFRTERNN ON THE TEST SET.

Model	Precision	Recall	$F1$ score
$\text{CFRTE}_{\text{GRU}}$	0.784	0.763	0.773
$\text{CFRTE}_{\text{CFGRU}}$	0.956	0.892	0.923
$\text{CFRTE}_{\text{LSTM}}$	0.791	0.817	0.804
$\text{CFRTE}_{\text{CFLSTM}}$	0.976	0.887	0.929
$\text{CFRTE}_{\text{BERT}}$	0.989	0.971	0.980

According to Table 2, compared to the CFRTE model with RNN as the encoder layer, the CFRTE model with BERT encoder has higher precision, recall, and $F1$ score, proving that the large language model BERT can better learn the features of natural language and improve the performance of relationship extraction models. In addition, the performance of the CFRTE model with bidirectional RNN encoder is significantly better than that of the unidirectional RNN, indicating that in the task of relationship extraction, the feature information of each word in the text depends not only on the preceding text but also on the following text.

IV. CONCLUSION

(1) Regarding the difficulty of extracting compressor fault knowledge from unstructured compressor fault text data, a method of compressor fault knowledge mining based on the large language model is proposed. Firstly, the concept ontology of compressor fault knowledge is built according to the characteristics and principles of compressor faults, and then the CFRTE model of compressor fault knowledge mining based on the large language model is established according to this standard. This method can also be extended to the fault diagnosis of other equipment in the gas transmission pipeline station.

(2) Experimental verification shows that the CFRTE model proposed in this paper has the highest $F1$ score compared with five models - SPERT, CasRel, OneRel, PRGC, GRTE, which reached 0.98, meeting the requirements of compressor fault knowledge mining.

(3) Compared with CFRTE model using four conventional RNN as encoder layer - GRU, CFGRU, LSTM, CFLSTM, the $F1$ score of CFRTE model based on BERT has increased by 26.78%, 6.18%, 21.89% and 5.49% respectively, proving the improvement of the performance of the compressor fault knowledge mining model by the pre-trained large language model BERT.

REFERENCES

[1] G. Milad, E. Ebrahim and J. Mohamad, "Compressor fault diagnosis based on SVM and GA," Vibroengineering PROCEDIA, vol. 12, 2017-6-30 2017.

[2] J. Van Gompel, D. Spina and C. Develder, "Cost-effective fault diagnosis of nearby photovoltaic systems using graph neural networks," Energy, vol. 266, 2023-3-1 2023.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," 2017-6-12 2017.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019 CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES (NAACL HLT 2019), VOL. 1, pp. 4171-4186, 2019-1-1 2019.

[5] Z. Wei, J. Su, Y. Wang, Y. Tian, and Y. Chang, "A Novel Cascade Binary Tagging Framework for Relational Triple Extraction," 58TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL 2020), pp. 1476-1488, 2020-1-1 2020.

[6] F. Ren, L. Zhang, S. Yin, X. Zhao, S. Liu, B. Li, and Y. Liu, "A Novel Global Feature-Oriented Relational Triple Extraction Model based on Table Filling," 2021 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP 2021), pp. 2646-2656, 2021-1-1 2021.

[7] F. Ren, L. Zhang, X. Zhao, S. Yin, S. Liu, and B. Li, "A Simple but Effective Bidirectional Framework for Relational Triple Extraction," 2021-12-9 2021.

[8] M. Eberts and A. Ulges, "Span-Based Joint Entity and Relation Extraction with Transformer Pre-Training," ECAI 2020: 24TH EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE, vol. 325, pp. 2006-2013, 2020-1-1 2020.

[9] Y. Shang, H. Huang and X. Mao, "OneRel : Joint Entity and Relation Extraction with One Module in One Step," THIRTY-SIXTH AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE / THIRTY-FOURTH CONFERENCE ON INNOVATIVE APPLICATIONS OF ARTIFICIAL INTELLIGENCE / TWELVETH SYMPOSIUM ON EDUCATIONAL ADVANCES IN ARTIFICIAL INTELLIGENCE, pp. 11285-11293, 2022-1-1 2022.

[10] H. Zheng, R. Wen, X. Chen, Y. Yang, Y. Zhang, Z. Zhang, N. Zhang, B. Qin, M. Xu, and Y. Zheng, "PRGC: Potential Relation and Global Correspondence Based Joint Relational Triple Extraction," 2021-6-18 2021.

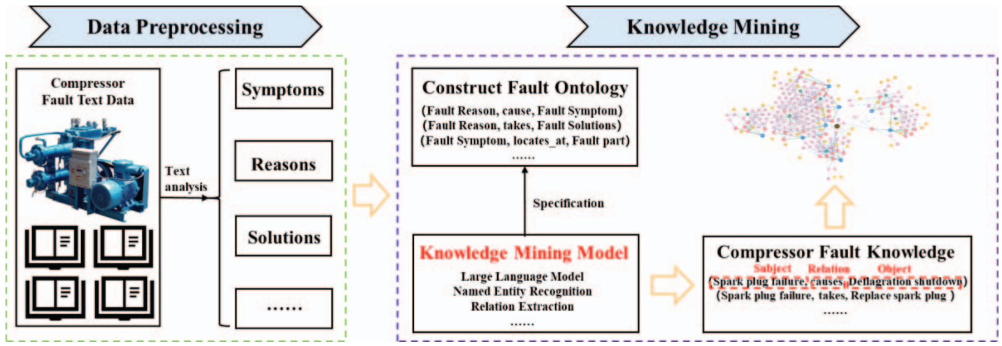

Figure 1. Framework of Compressor Fault Knowledge Mining Process.

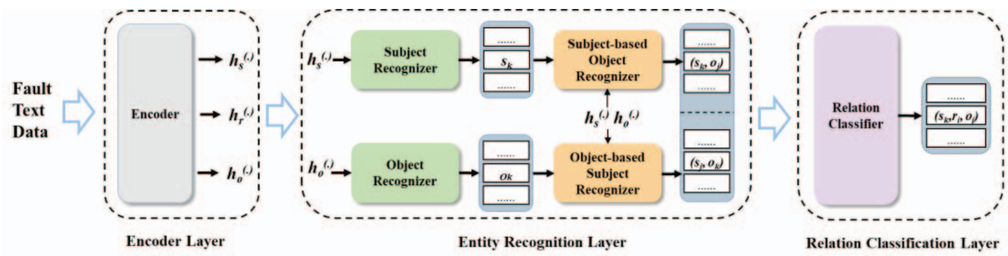

Figure 3. The Structure of CFRTE model.