Research on Knowledge Graph Construction for Intelligent Operation and Maintenance of Electrical Transformers

Bowen Zhang¹*, Fei Gao², Ning Yang, Chengbo Hu² and Ziquan Liu²

¹China Electric Power Research Institute, Beijing, China
²Electric Power Research Institute of State Grid Jiangsu Power Company, Jiangsu, China
*Corresponding author’s e-mail: zhangbowen@epri.sgcc.com.cn

Abstract. In the process of transformer equipment operation and maintenance, power companies encounter some problems, including unstructured text data difficult to use, full caliber data difficult to deeply integrate and shallow equipment knowledge application. Based on artificial intelligence technology such as semantic web, knowledge map, natural language processing, etc., this paper studies the key technologies of equipment intelligent service, and proposes a knowledge intelligent technology framework for transformer. This framework includes three components: unstructured text intelligent recognition and extraction, device centered device knowledge representation and storage, and device knowledge intelligent application.

1. Introduction
One of the goals of artificial intelligence research in the field of power equipment operation and inspection is to realize the empirical knowledge of experts in machine understanding and equipment operation and maintenance[1], including understanding the proprietary language of the power field, and fusion processing equipment operation and inspection related structured and unstructured full-caliber data, allowing the machine to use knowledge to enrich the intelligent means of transportation and inspection. Current research hotspots include equipment text data mining, fault reasoning and auxiliary diagnosis, and equipment defect retrieval. Equipment text data mining refers to the extraction of power domain knowledge from the complex text data of power equipment, such as power equipment defect knowledge. Knowledge mining of data in the field of power equipment is only the first step of knowledge engineering. The ultimate goal is to effectively improve the efficiency of equipment management [3], such as intelligent classification of defect text [4-5], and realization of transformer equipment related Retrieval, reasoning, fault diagnosis [6-7]. At present, such research is mainly based on traditional expert system thinking, and the cost of knowledge construction and update is high, and it is difficult to meet the requirements of large-scale knowledge extraction and low-cost application of power systems. In recent years, artificial intelligence technology has developed vigorously, especially the knowledge graph technology proposed by Google and the Bidirectional Encoder Representations from Transformers (BERT) model based on transformers, etc., which provide new directions for solving the above problems. This paper draws on the key technologies related to the construction and query of knowledge graphs mentioned in the literature, combined with the author's research and practical results in the key technology of knowledge graphs in the field of power equipment, and proposes a device knowledge graph that supports intelligent management of transformers. Technical framework, researched the key technology of intelligent operation and inspection of transformer.
equipment based on knowledge, formed intelligent models such as equipment semantic extraction model, summarized the application results of this technology in scenarios such as automatic extraction of transformer fault reports and flexible query of equipment information, and proposed the next research direction of knowledge-based equipment intelligent operation and inspection technology is provided, which provides the feasibility for the realization of machine understanding of equipment semantics.

1.1. Equipment knowledge graph construction technology
Knowledge graph is a knowledge network that connects and organizes entities and attributes through relationships [7], which is composed of "entity-relation-entity" or "entity-attribute" triples. The knowledge graph is divided into two types: open domain and closed domain [8]. This paper studies the key technology of knowledge graph construction for the field of intelligent operation and inspection of power transformer equipment, which belongs to a closed domain. The process of building a knowledge graph is shown in Figure 1. It consists of three parts: knowledge extraction, knowledge fusion, and knowledge storage.

![Figure 1. Construction process of transformer equipment knowledge graph](image)

1.1.1. Knowledge extraction technology of transformer structured data
The structured data related to transformer equipment includes equipment ledger data, equipment defect data, work order data, etc. These data are stored by RDB, which has obvious structural characteristics. Through the field mapping of the structured data source table structure and the knowledge graph ontology concept, the RDB data is quickly converted into Resource Description Framework (RDF) data.

1.1.2. Knowledge extraction technology of transformer unstructured text data
In this paper, Named Entity Recognition (NER) was used to extract entities and relationships from long texts of transformer equipment (such as fault reports, test data, etc.), we use rule-based template extraction + intelligent model-based approach to achieve NER. 1) Identification based on rule templates. It is suitable for scenarios where knowledge is extracted from the description text with a relatively fixed format or structure, such as "Huangnitou 110kV Substation #2 Main Transformer Trip Reason Analysis Report", etc. The name of the substation and the transformer name show obvious template characteristics. Text-like statements, based on Python regular expressions (Regular expressions, RE) script template extraction, high efficiency, accuracy, and low cost. 2) Recognition based on intelligent model. According to the transformer fault report data of the electric power company over the years, three models of conditional random field (CRF), bi-directional long short-term memory (BiLSTM) and BERT-BiLSTM-CRF are used to extract, Comprehensive comparison of the three model evaluation indicators accuracy rate (Precision, P), recall rate (Recall, R) and F1 value (F1) to optimize the model used. Among them, the BERT-BiLSTM-CRF model structure is shown in Figure 2.
As can be seen in Figure 2, the BERT-BiLSTM-CRF named entity recognition model is added to the BiLSTM model by using the BERT model as a feature representation layer, applying a forward and backward LSTM network to each input sentence sequence, and this The two networks are connected to a common output layer. This structure can provide the output layer with complete context information of the input sentence sequence. Finally, through the CRF model, the label information before and after the sequence is effectively considered.

1.1.3. Transformer knowledge fusion technology:
After extracting transformer knowledge from structured data and unstructured text data sources from equipment management systems and other systems, when storing the knowledge, it is necessary to resolve the conflict and fusion of the two types of knowledge. Using the transformer entity name, numerical attributes, and relationship attributes as feature quantities, the semantic similarity of the two entities is calculated. Before carrying out entity similarity calculation, it is necessary to perform data preprocessing on the attribute value of the entity, and normalize the enumerated attribute value of the transformer-related entity. For example, the voltage level value of the transformer includes 110 kV, 110 kV, 22 Ten thousand volts, etc.; transformer failure phenomena include "serious oil leakage from the main body", "serious oil leakage from the main body", "liquid oil leakage from the main body", etc. The data needs to be normalized and adjusted to "serious oil leakage from the main body". The entity similarity calculation is obtained by weighting the text similarity calculation of the entity feature quantity. The similarity calculation of two entity feature quantities is shown in formula (1):

$$\text{Sim}(A, B) = \alpha \text{Sim}(A_0, B_0) + \beta \sum_{i=1}^{n} (A_i, B_i) + \gamma \sum_{j=1}^{m} (A_j, B_j)$$

Where: $A_0$, $B_0$ refer to the entity names of entity A and entity B; $A_i$, $B_i$ refer to the numerical attribute value of entity A and entity B; $A_j$, $B_j$ refer to the object attribute value of entity A and entity B; $\text{Sim}(A, B)$ refers to the semantic similarity of two attribute values; $\alpha+\beta+\gamma=1$, where $\alpha$, $\beta$, and $\gamma$ respectively represent the similarity of entity name, the similarity of the attribute value of the entity, and the similarity of the attribute value of the entity object. The weight of the degree. Entity attribute values are divided into three types: numeric type, collection type, and text type. For numerical attributes, use the formula (2) to calculate the similarity:
In formula (2): the value range of \( D \) is from 0 to 1. The larger the gap between \( d_i \) and \( d_j \), the larger the value of \( D \), the lower the similarity between the two. For set attributes, the similarity of two sets is judged by calculating Jaccard similarity, as shown in formula (3):

\[
Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|}
\]

In formula (3): for the two sets of set \( A \) and set \( B \), the value range of Jaccard similarity is from 0 to 1. The larger the value, the higher the similarity between the two sets. For text-based attributes, the open source jieba word segmentation tool is used for word segmentation, the word vector calculation tool word2vec is used to train to form a word vector model, the attribute value of the entity is converted into a word vector, and the vector of the same attribute value of different entities is calculated by cosine similarity similarity, as shown in formula (4):

\[
\text{CosSim} = \frac{\sum_{i=1}^{n} (A \times B)}{\sqrt{\sum_{i=1}^{n} (A)^2} \times \sqrt{\sum_{i=1}^{n} (B)^2}}
\]

In formula (4): the range of cosine similarity is from 0 to 1. The larger the value, the higher the similarity of the two entities.

1.1.4. Knowledge storage technology
At present, knowledge graph storage technology mainly includes triple database based on RDF model and graph database based on attribute graph model. In the knowledge-based equipment intelligent application scenario, considering the standard openness, semantic analysis requirements, support component richness, easy scalability and other factors, this paper adopts the semantic web RDF model as the knowledge representation model of the transformer equipment knowledge graph, comprehensive use support The triple database stored in RDF and the open source MongoDB database component realize the storage of transformer knowledge graph data.

2. Technical component framework of equipment knowledge graph
The technical component framework of the equipment knowledge graph can be divided into four layers: data layer, equipment knowledge extraction component, equipment map construction component, and equipment knowledge service component. 1) The data layer covers various data sources of transformers, including RDB data, long text document data (including WPS, PDF, etc.). 2) The equipment knowledge extraction component extracts various knowledge of transformer equipment from the data provided by the data layer, which is divided into structured knowledge extraction and unstructured knowledge extraction. 3) The equipment map building component provides services such as transformer ontology definition, transformer entity knowledge storage, knowledge fusion and reasoning, and path query.

This paper proposes a technical framework for knowledge extraction from transformer long text data. The framework provides functions such as transformer corpus labeling, extraction model training, and knowledge extraction, and supports the automatic extraction of knowledge from transformer long text reports in a process-based manner.

2.1. Data preprocessing
Perform preprocessing operations such as format conversion, removal of pictures, and removal of directories for transformer-related reports.

2.2. Ontology structure design
The research in this paper is aimed at the field of transformer equipment, which requires high accuracy of the ontology structure. It uses top-down and business experts to determine the ontology structure of the transformer, and clarify the attributes and relationships of transformer-related concepts.
2.3. Corpus labeling
Business experts annotate the text on the corpus tagging tool, and use {B, I, O} to tag the corpus (taking the transformer entity as an example, B represents the first word of the transformer entity, I represents the remaining words of the transformer entity, and O represents Words that do not belong to the transformer entity) format representation, and the BERT-BiLSTM-CRF extraction model is generated based on the marked corpus.

2.4. Document batch extraction
Based on the intelligent extraction model, the entities and relationships in long text reports can be automatically identified in batches, and large sections of unstructured text can be converted into structured graphs. Take chapter 3.2-3.3 of "State Grid Corporation's Substation Maintenance Management Regulations" as an example. The original content of the chapter is shown in Figure 3. After identification based on the intelligent extraction model, the "on-load tap-changer maintenance" is identified as the maintenance process, "A) Disconnect before maintenance...f) It is strictly prohibited to step on the explosion-proof membrane of the on-load tap changer" as the attribute value of "Safety Precautions" of the "On-load tap-changer maintenance" entity. The map formed after identification is shown in Figure 4.

Figure 3. Original long text document
In Figure 4, the legend in the upper left corner represents different types of entities, one type of entity has one legend, and the text on the connection represents the attribute name; the lower right legend represents the attribute value and the color of the entity, click on the entity/attribute value legend to control whether the entity and attribute value elements are displayed or not.

3. Transformer entity extraction model
This paper verifies and analyzes the knowledge graph construction and intelligent question answering technology for transformer intelligent operation inspection, including transformer entity extraction model, knowledge fusion and storage model, and transformer entity information intelligent question answering model.

3.1. Model preparation
According to the unstructured text data knowledge extraction technology method described in 2.1.2 of this article, three models of CRF, BiLSTM, and BERT-BiLSTM-CRF were constructed for experiments. The model is built based on the Tensorflow framework and determines the hyperparameters of the model. The Tensorflow framework is an end-to-end open source deep learning framework platform developed by Google. When selecting hyper-parameters, use the default hyper-parameter settings, observe the change of loss function value (loss), preliminarily determine the range of each hyper-parameter, and then adjust the parameters. For each hyperparameter, adjust only one parameter each time you adjust, and then observe the loss value change. Among them, the selection of the training set batch size parameter (batch_size) refers to the GPU card cache size, and the dropout parameter takes the default value. The sequence length parameter (seq_length) is determined according to the sentence length of the corpus, which can cover most sentences. The final experimental parameters are as follows: seq_length is 128, lstm unit number parameter (lstm_size) value is 128, training set batch size The parameter (batch_size) parameter value is 64, the test set batch size parameter (batch_size) is 64, and the learning rate parameter is 10-5. In order to prevent gradient explosion during training, use gradient clipping technology and set the clipping parameter (clip) to 0.5 and the dropout value to 0.5.
3.2. Experimental data and verification indicators
Collect the equipment field and transformer fault reports over the years, and form a training set and a test set after preprocessing the reports. 1) Transformer corpus collection, collecting relevant reports in the field of transformer equipment over the years, including fault reports, equipment status evaluation reports, family defect reports, etc. The corpus data is preprocessed, the text data is retained, the pictures, tables, directories, charts and other elements in the report are removed, and the text is segmented and sentenced to form a typical sentence that can be independently labeled. 2) Divide the corpus, select 298 transformer failure case reports, involving a total of 34 transformer concepts. The core concepts are shown in Table 1. The reports are divided according to a certain proportion, 232 copies are used for training, and 66 copies are used for training test.

| Index | Entity Type | Attribute List | Number |
|-------|-------------|----------------|--------|
| 1     | Transformer| Name, Voltage Level, Model, et al. | 228    |
| 2     | Failure    | Failure time, Operation status, Trip status, et al. | 76     |
| 3     | Manufacture| Manufacture Name | 243    |
| 4     | Equipment parts | Production date, Model, Commissioning date | 35     |
| 5     | Failure cause | Failure cause, Responsibility reason | 237    |
| 6     | Experiment conclusion | qualified or not | 190    |
| 7     | Examination result | Examination content | 319    |
| 8     | DGA        | Test date, Test conclusion | 22     |
| 9     | Substation | Substation name, Voltage level | 217    |
| 10    | Failure phenomenon | Phenomenon tag | 193    |

The corpus sequence labeling is to label the reports in the training set of the corpus. According to the transformer fault concept, the data processed by the corpus is labeled in sentence units. This article uses the ternary mark set \{B, I, O\}. The number of main entities in the test set is shown in Table 2.

| Entity Type | Entity Number | Entity Type | Entity Number |
|-------------|--------------|-------------|--------------|
| Transformer | 62           | Failure cause | 67           |
| Failure     | 15           | Experiment conclusion | 51           |
| Manufacture | 71           | Examination result | 84           |
| Equipment parts | 11           | DGA         | 29           |
| Failure phenomenon | 54           | Substation | 60           |

For various named entities, the values of P, R and F1 are used as model evaluation indicators. It is defined as formula (5), formula (6), and formula (7):

\[
P = \frac{\text{Number of correctly identified named entities}}{\text{Number of identified named entities}} \times 100\% \tag{5}
\]

\[
R = \frac{\text{Number of correctly identified named entities}}{\text{The number of named entities in the standard answer}} \times 100\% \tag{6}
\]

\[
F1 = \frac{2 \times P \times R}{P + R} \times 100\% \tag{7}
\]
3.3. Experimental results
On the data set, CRF, BiLSTM, and BERT-BiLSTM-CRF models are used for analysis. The experimental results are shown in Table 3. It can be seen that the model based on BERT-BiLSTM-CRF is more accurate than the CRF and LSTM models.

| Task               | Recall rate (%) | Precision (%) | F1-Score (%) |
|--------------------|-----------------|---------------|--------------|
| CRF                | 61.3            | 65.13         | 63.16        |
| LSTM               | 63.45           | 66.78         | 65.07        |
| BERT-BiLSTM-CRF    | 72.37           | 76.51         | 74.38        |

For each entity type, the experimental results of the extraction model are shown in Table 4:

| Task                  | Recall rate (%) | Precision (%) | F1-Score (%) |
|-----------------------|-----------------|---------------|--------------|
| Transformer           | 84.50           | 81.37         | 82.90        |
| Failure               | 94.90           | 92.08         | 93.47        |
| Manufacture           | 89.18           | 87.82         | 88.49        |
| Equipment parts       | 69.57           | 66.67         | 68.09        |
| Failure cause         | 72.31           | 71.26         | 71.78        |
| Experiment Conclusion | 10.00           | 9.09          | 9.52         |
| Examination result    | 75.00           | 75.00         | 75.00        |
| DGA                   | 77.78           | 77.78         | 77.78        |
| Substation            | 81.13           | 78.03         | 79.55        |
| Failure phenomenon    | 73.81           | 68.89         | 71.26        |

3.4. Knowledge storage model
Based on the document structure of MongoDB, this paper designs a storage model for storing semantic triples data, covering the SPO triples of RDF. The model adopts the lightweight JavaScript data exchange format (JavaScript Object Notation, JSON) through 5 types of fields. Definition, the definition of each field is shown in Table 5.

| Index | SPO     | Attribute                                 |
|-------|---------|-------------------------------------------|
| 1     | Subject | ID, Task ID, Corresponding concept         |
| 2     | Predicate | ID, Name, Attribute                      |
| 3     | Object  | ID, Task ID, Corresponding concept, Value  |
| 4     | Other Filed | Starting time, End time               |
| 5     | Reserved Field | For subsequent expansion                |

Based on this storage model, it provides a seven-tier composite index for the three elements of SPO, including S, P, O, SP, PO, SO, and SPO, and provides efficient data query through the seven-tier index.

4. Conclusion
This article summarizes the author's research and practice in the field of transformer equipment in the field of knowledge graph technology in recent years, studies the algorithm models of transformer entity...
extraction model, intent recognition and slot extraction model based on BERT-BiLSTM-CRF, and proposes equipment knowledge graph technology Component framework, and the framework was verified in the flexible question and answer of power company’s transformer equipment information and automatic extraction of transformer fault reports. Due to the length of the paper, some models and algorithm technical details were not explained in depth. The research results of this paper are not limited to transformer equipment, but are also applicable to other substation equipment such as circuit breakers, and have strong practical and popularization application value.

Acknowledgments
We thank the State Grid Corporation of Science and Technology Project Research on Text Intelligent Analysis Technology for Quality Information of Power Grid Main Equipment (grant number 5500-202019090A-0-0-00).

References
[1] Bingfeng Luo, Yansong Feng, Zheng Wang, et al. (2018) Mar-rying up regular expressions with neural networks: A case study for spoken language understanding. arXiv preprint arXiv:1805.05588.
[2] Lai, Y., Jia, el al. A Chinese Question Answering System for Single-Relation Factoid Questions. (2017) In: 6th National CCF Conference on Natural Language Processing and Chinese Computing, pp.124–135. Springer, Cham.
[3] Xu K., Lai Y.X., Feng Y.S., and Wang Z.G. (2019). Enhancing Key-Value Memory Neural Networks for Knowledge Based Question Answering. InNAACL-HLT.